Springer Proceedings in Mathematics & Statistics

Christian Klingenberg
Michael Westdickenberg *Editors*

Theory, Numerics and Applications of Hyperbolic Problems II

Aachen, Germany, August 2016

Springer

Springer Proceedings in Mathematics & Statistics

Volume 237

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Christian Klingenberg · Michael Westdickenberg
Editors

Theory, Numerics and Applications of Hyperbolic Problems II

Aachen, Germany, August 2016

Springer

Editors
Christian Klingenberg
Department of Mathematics
Würzburg University
Würzburg
Germany

Michael Westdickenberg
Department of Mathematics
RWTH Aachen University
Aachen
Germany

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-91547-0 ISBN 978-3-319-91548-7 (eBook)
https://doi.org/10.1007/978-3-319-91548-7

Library of Congress Control Number: 2018941540

Mathematics Subject Classification (2010): 35Lxx, 35M10, 35Q30, 35Q35, 35Q60, 35Q72, 35R35, 65Mxx, 65Nxx, 65Txx, 65Yxx, 65Z05, 74B20, 74Jxx, 76L06, 76Rxx, 76Txx, 80A32, 80Mxx, 83C55, 83F05

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Proceedings of the 16th International Conference on Hyperbolic Problems: Theory, Numerics and Application

Organizers: Christian Klingenberg and Michael Westdickenberg

Organizer's Introduction

This series of bi-yearly conferences began in 1986 and celebrated its 30th anniversary with this conference. From its very beginning, the conference set out to bring together researchers studying theoretical issues, numerical methods and applications in hyperbolic partial differential equations. Initially, this pursuit was under the influence of Glimm's major result [1], where the convergence of a useful numerical method (in one space dimensions) gave rise to an existence proof. Even though 30 years later, the areas of theory, numerics and applications are no longer as closely intertwined as they were in the beginning, the organizers feel that having them together in one conference today is much more than historical nostalgia. One area may give impulses to another area. Given that fundamental issues in the field of hyperbolic problems are open (e.g. is there an admissibility condition that gives rise to well-posedness for the Euler equations in multiple space dimensions?), new impulses from different areas are thoroughly needed.

This conference and these proceedings provide a snapshot of the activity in its field at the time of this conference. The field is quite broad, as seen by a partial list of subjects covered:

- hyperbolic conservation laws,
- wave equations,
- partial differential equations of mixed type,
- kinetic equations,
- theoretical questions and numerical schemes for all of the above and
- applications in physics and engineering using all of the above.

This conference over the last 30 years has developed into one of the main conferences in applied mathematics. This is due to the vigour of the field, the enormously challenging questions that still lie ahead and its extreme usefulness in applications. Many researches around the world contribute to this field, so that we expect this series of conferences will be vital for many years to come.

The organizers want to acknowledge the financial support of the German Science Foundation (DFG). The organizers wish to thank all participants, the local staff and organizers, and everybody else who was involved in this event in one way or another, for making the 2016 edition of the "International Conference on Hyperbolic Problems: Theory, Numerics, and Applications" a success.

Reference

1. J. Glimm, Solutions in the large for nonlinear hyperbolic systems of equations. Commun. Pure Appl. Math. **18**(4), 697–715 (1965)

Speakers at the 16th International Conference on Hyperbolic Problems that Contributed to these Proceedings

A Stochastic Galerkin Method for the Fokker–Planck–Landau Equation with Random Uncertainties

Jingwei Hu, Shi Jin and Ruiwen Shu

Abstract We propose a generalized polynomial chaos-based stochastic Galerkin method (gPC-sG) for the Fokker–Planck–Landau (FPL) equation with random uncertainties. The method can handle uncertainties from initial or boundary data and the neutralizing background. By a gPC expansion and the Galerkin projection, we convert the FPL equation with uncertainty into a system of deterministic equations. A consistency result is proven for the approximation of the collision operator. To compute efficiently the collision kernel under the gPC expansion, we use a singular value decomposition (SVD) combined with a fast spectral method for the collision operator. For high-dimensional random inputs, we adopt a sparse basis and use the sparsity of a set of basis-related coefficients and the Lax–Friedrichs splitting to avoid all the SVD involved. Numerical experiments verify the efficiency of the gPC-sG method.

Keywords Fokker-Planck-Landau equation · Uncertainty quantification
Stochastic Galerkin method · Polynomial chaos · Sparse grids

J. Hu
Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA
e-mail: jingweihu@purdue.edu

S. Jin (✉) · R. Shu
Department of Mathematics, University of Wisconsin-Madison,
Madison, WI 53706, USA
e-mail: sjin@wisc.edu

R. Shu
e-mail: rshu2@math.wisc.edu

S. Jin
Department of Mathematics, Institute of Natural Sciences,
MOE-LSEC and SHL-MAC, Shanghai Jiao Tong University, Shanghai 200240, China

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_1

1 Introduction

First derived by Landau [6] as the grazing collision limit of the Boltzmann equation, the Fokker–Planck–Landau (FPL) or Landau equation is a collisional kinetic model that describes the non-equilibrium dynamics of charged particles in a plasma [16].

Let $f = f(t, x, v)$ be the density distribution function of particles, where t is the time, x is the space, and v is the velocity. The FPL equation with the mean-field term (also known as the Vlasov–Poisson–Landau equation) reads

$$\partial_t f + v \cdot \nabla_x f + E(t, x) \cdot \nabla_v f = Q(f, f), \quad t > 0,\ x \in \Omega \subset \mathbb{R}^{d_x},\ v \in \mathbb{R}^{d_v}, \tag{1}$$

where $E(t, x)$ is the electric field given by

$$E(t, x) = -\nabla_x \phi(t, x), \tag{2}$$

and $\phi(t, x)$ is a self-consistent electrostatic potential function satisfying the Poisson equation

$$\Delta_x \phi(t, x) = \mu(x) - \int_{\mathbb{R}^{d_v}} f(t, x, v)\, \mathrm{d}v, \tag{3}$$

where $\mu(x)$ is a neutralizing background satisfying

$$\int_{\mathbb{R}^{d_x}} \mu(x)\, \mathrm{d}x = \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_v}} f(t, x, v)\, \mathrm{d}v\, \mathrm{d}x. \tag{4}$$

$Q(f, f)$ on the right-hand side of (1) is the FPL collision operator that models binary interactions among particles:

$$Q(f, f)(v) = \nabla_v \cdot \int_{\mathbb{R}^{d_v}} A(v - v_*)[f(v_*)\nabla_v f(v) - f(v)\nabla_{v_*} f(v_*)]\, \mathrm{d}v_*. \tag{5}$$

Here $A(w)$ is a semi-positive definite matrix defined by

$$A(w) = \Psi(w)\left(I - \frac{w \otimes w}{|w|^2}\right), \tag{6}$$

where I is the identity matrix. For inverse power law potentials, $\Psi(w) = |w|^{\gamma+2}$ with $-3 \le \gamma \le 1$. The case $\gamma = -3$ corresponds to the Coulomb interaction which is of primary importance in plasma applications.

The collision operator $Q(f, f)$ possesses some important physical properties: it preserves mass, momentum, and energy

$$\int_{\mathbb{R}^{d_v}} Q(f, f)\phi(v)\, \mathrm{d}v = 0, \quad \phi(v) = 1, v, |v|^2; \tag{7}$$

and satisfies the entropy dissipation inequality (the H-theorem)

$$\int_{\mathbb{R}^{d_v}} Q(f, f) \log f \, \mathrm{d}v \leq 0. \tag{8}$$

The equality only holds when f attains the local equilibrium (Maxwellian)

$$f(v) = M(v) = \frac{\rho}{(2\pi T)^{d_v/2}} e^{-\frac{(v-u)^2}{2T}}, \tag{9}$$

where ρ, u, T are the density, bulk velocity, and temperature defined as

$$\rho = \int_{\mathbb{R}^{d_v}} f \, \mathrm{d}v, \quad u = \frac{1}{\rho} \int_{\mathbb{R}^{d_v}} v f \, \mathrm{d}v, \quad T = \frac{1}{d_v \rho} \int_{\mathbb{R}^{d_v}} (v-u)^2 f \, \mathrm{d}v. \tag{10}$$

Equation (1) needs to be supplemented with appropriate initial condition

$$f(0, x, v) = f^0(x, v), \tag{11}$$

where f^0 can be chosen as, for example, the local equilibrium (9). For boundary condition, a commonly used one is the Maxwell boundary condition: For any boundary point $x \in \partial\Omega$, let $n(x)$ be the unit inward normal vector to the boundary, then the inflow boundary condition is specified as

$$f(t, x, v) = g(t, x, v), \quad v \cdot n > 0, \tag{12}$$

$$g(t, x, v) := (1-\alpha) f(t, x, v - 2(v \cdot n)n) + \frac{\alpha}{(2\pi)^{\frac{d_v-1}{2}} T_w^{\frac{d_v+1}{2}}} e^{-\frac{v^2}{2T_w}} \int_{v \cdot n < 0} f(t, x, v) |v \cdot n| \, \mathrm{d}v, \tag{13}$$

where $T_w = T_w(t, x)$ is the temperature of the wall (boundary). The constant α ($0 \leq \alpha \leq 1$) is the accommodation coefficient with $\alpha = 1$ corresponding to the purely diffusive boundary, and $\alpha = 0$ the purely specular reflective boundary.

In the past decades, the FPL equation (1) has been studied extensively both theoretically and numerically. The readers are referred to [16] for a review of the main mathematical aspects, and the recent paper [1] and references therein for relevant numerical methods. In spite of the vast amount of existing research, the study of the FPL equation has mostly remained deterministic and ignored uncertainty. In reality, however, there are many sources of uncertainties that can arise in this equation: imprecise measurements for initial boundary conditions and physical parameters; incomplete knowledge of the fundamental interaction mechanism between particles; and so on. Understanding the impact of these uncertainties is critical to the simulations of complex plasma systems and will allow scientists and engineers to obtain more reliable predictions and perform better risk assessment. The goal of this paper

is to develop an efficient stochastic numerical method for uncertainty quantification (UQ) of the FPL equation (1).

The basic framework of our work is built on a probabilistic approach which models the uncertain parameters as random variables. In the FPL equation (1), this amounts to consider the distribution function

$$f = f(t, x, v, z), \quad z \in I_z \in \mathbb{R}^d, \tag{14}$$

now depending on an extra argument z—a d-dimensional random vector with support I_z collecting all possible uncertainties in the system. For instance, one may consider $z = (z^\mu, z^{\text{ini}}, z^{\text{bdry}})$, where z^μ, z^{ini}, z^{bdry} denote, respectively, the random parameters arising from

- the neutralizing background (3): $\mu = \mu(x, z^\mu)$;
- the initial condition (11): $f(0, x, v, z) = f^0(x, v, z^{\text{ini}})$ for $x \in \Omega$;
- the boundary condition (12): $f(t, x, v, z) = g(t, x, v, z^{\text{bdry}})$ for $x \in \partial\Omega$.

We will further assume the components of z are already mutually independent random variables obtained through some dimension reduction technique, e.g., Karhunen–Loève expansion [9], and do not pursue the issue of random input parameterization in this paper.

To properly model the propagation of uncertainties, we adopt the generalized polynomial chaos-based stochastic Galerkin (gPC-sG) method, which is widely used in the UQ simulations nowadays [2, 3, 10, 13, 17, 18]. Simply speaking, this method seeks to approximate the unknown function f via an orthogonal polynomial series:

$$f(t, x, v, z) \approx \sum_{k=1}^{K} f_k(t, x, v)\Phi_k(z) := f^K(t, x, v, z), \ f_k(t, x, v) = \int_{I_z} f(t, x, v, z)\Phi_k(z)\pi(z)\,\mathrm{d}z, \tag{15}$$

where $\{\Phi_k(z)\}$ is a set of d-variate polynomials of degree up to m which satisfy

$$\int_{I_z} \Phi_j(z)\Phi_k(z)\pi(z)\,\mathrm{d}z = \delta_{jk}, \quad 1 \le j, k \le K,$$

with $\pi(z)$ being the probability distribution of z and δ_{jk} the Kronecker delta function. The number of basis functions is $K = \binom{m+d}{m}$. Equipped with this gPC representation, one then proceeds as follows: (1) Substitute the expansion (15) into the original equation and conduct a Galerkin projection. This usually results in a system of coupled deterministic equations for the gPC coefficients $\{f_k\}_{k=1}^K$ requiring different treatment from the corresponding deterministic equation. (2) Solve the gPC system. (3) Use $\{f_k\}_{k=1}^K$ to reconstruct the solution in I_z via (15), or construct the solution statistics directly, e.g., the mean and standard deviation can be retrieved as

$$\mathbb{E}[f] = f_1, \quad S[f] = \sqrt{\sum_{k=2}^{K} f_k^2}. \tag{16}$$

For the FPL equation, the main difficulty associated with solving the gPC-sG system lies in the evaluation of the nonlinear collision operator. Similar to the work [5], we propose a fast algorithm to efficiently compute the collision operator under the Galerkin projection. The acceleration is achieved by combining a singular value decomposition (SVD) of the collision kernel with the fast spectral method in the deterministic case [11].

In the cases where the random domain I_z is high-dimensional, i.e., d is large, the usual gPC-sG method may fail to be affordable since the number of basis functions $K = \binom{m+d}{m}$ is too large. To circumvent this difficulty, we use the sparse wavelet basis as in our previous work [15]. We use N-level hierarchical piecewise polynomial functions of degree at most m as basis functions in one dimension and use a standard sparse grid construction to obtain basis functions in d-dimensional random spaces. With this basis, one can achieve an accuracy of $O(N^d 2^{-N(m+1)})$ with $K = O((m+1)^d 2^N N^{d-1})$ basis functions. The accuracy is $O(K^{-(m+1)} (\log K)^{(m+2)(d-1)})$ in terms of K. This method is much more efficient than the usual gPC-sG method if d is large.

When using the sparse grid method, K can still be too large to make an SVD of order K affordable. Thus the following two difficulties arise. The first one is that one can no longer afford the SVD approach for the collision operator. To avoid it, we notice the sparsity of a basis-related tensor $S_{b,ijk}$ proved in [15]. As a result, one can compute the collision operator directly with low computational cost. The second difficulty is that a direct computation of the numerical flux for the mean-field term requires a diagonalization of constant flux matrices of order K. To avoid this diagonalization, we utilize the local Lax–Friedrichs splitting [7]. In this way, one can compute the second-order upwind flux without diagonalization of the flux matrices.

The rest of this paper is organized as follows. Section 2 describes in detail the gPC-sG method for the FPL equation with uncertainty. Section 3 discusses the spatial and time discretization. In Sect. 4, we give a consistency analysis of the gPC-sG method for the collision operator. In Sect. 5, we give a sparse wavelet method for problems with high-dimensional random inputs. Extensive numerical results are presented in Sect. 6. Finally, the paper is concluded in Sect. 7.

2 The gPC-sG Method for the FPL Equation with Uncertainties

In this section, we describe the gPC-sG method for the FPL equation with uncertainty. We start by substituting the truncated gPC expansion (15) into Eq. (1). Upon a standard Galerkin projection, this yields a system of equations for the gPC coefficients f_k:

$$\begin{aligned}\partial_t f_k(t,x,v)+v\cdot\nabla_x f_k(t,x,v)+\nabla_v\cdot\int_{I_z}E(t,x,z)f^K(t,x,v,z)\Phi_k(z)\pi(z)\,\mathrm{d}z\\ =Q_k(f^K,f^K)(t,x,v),\quad 1\le k\le K,\end{aligned}\tag{17}$$

where $Q_k(f^K, f^K)$, the kth mode of the collision operator, is defined as

$$Q_k(f^K,f^K):=\int_{I_z}Q(f^K,f^K)(t,x,v,z)\Phi_k(z)\pi(z)\,\mathrm{d}z.\tag{18}$$

To simplify the forcing term, note that

$$E_k(t,x)=-\nabla_x\phi_k(t,x),\quad \Delta_x\phi_k(t,x)=\mu_k(x)-\int_{\mathbb{R}^{d_v}}f_k(t,x,v)\,\mathrm{d}v,\tag{19}$$

where $\mu_k(x)=\int_{I_z}\mu(x,z)\Phi_k(z)\pi(z)\,\mathrm{d}z$ are the gPC coefficients of the neutralizing background μ. Then the integral term in (17) becomes

$$\int_{I_z}\left(\sum_{i=1}^K E_i(t,x)\Phi_i(z)\right)\left(\sum_{j=1}^K f_j(t,x,v)\Phi_j(z)\right)\Phi_k(z)\pi(z)\,\mathrm{d}z=\sum_{j=1}^K A_{kj}(t,x)f_j(t,x,v),\tag{20}$$

with

$$A_{kj}(t,x):=\sum_{i=1}^K S_{ijk}E_i(t,x),\quad S_{ijk}=\int_{I_z}\Phi_i(z)\Phi_j(z)\Phi_k(z)\pi(z)\,\mathrm{d}z.\tag{21}$$

To simplify the collision term, we define the bilinear FPL collision operator as

$$Q(f,g)(v)=\nabla_v\cdot\int_{\mathbb{R}^{d_v}}A(v-v_*)(f(v_*)\nabla_v g(v)-f(v)\nabla_{v_*}g(v_*))\,\mathrm{d}v_*,\tag{22}$$

Then the collision term (18) can be expressed as

$$Q_k(f^K,f^K)=\sum_{i,j=1}^K S_{ijk}Q(f_i,f_j).\tag{23}$$

Due to the double summation in (23), a direct evaluation of the collision operator Q_k would be very expensive. To reduce the computational cost, we follow the approach proposed in [5]. Specifically, we pre-compute the singular value decomposition (SVD) of the matrix $\{S_{ijk}\}_{ij}$ for each k:

$$S_{ijk}=\sum_{r=1}^{R_k}U_{ir}^k V_{rj}^k,\tag{24}$$

where R_k is the numerical rank of the matrix. Plugging (24) into (23) and rearranging terms give

$$Q_k(f^K, f^K) = \sum_{r=1}^{R_k} Q\left(g_r^k, h_r^k\right), \quad g_r^k := \sum_{i=1}^{K} U_{ir}^k f_i, \quad h_r^k := \sum_{j=1}^{K} V_{rj}^k f_j. \tag{25}$$

Therefore, we reduce the original double summation into a single one. To compute the bilinear term $Q\left(g_r^k, h_r^k\right)$, we apply the fast spectral method introduced in [11] for the deterministic FPL collision operator. See Appendix for a brief description of this method. The numerical complexity of such a computation is $O(N_v^{d_v} \log N_v)$ where N_v is the number of mesh points in each velocity direction, and d_v is the dimension of the velocity space. Thus, for each k, the cost of computing Q_k is $O(R_k N_v^{d_v} \log N_v)$ with $R_k \le K$, and $K = \binom{m+d}{m}$ is the dimension of d-variate polynomials of degree up to m (note that the direct evaluation of Q_k based on (23) requires $O(K^2 N_v^{2d_v})$ operations).

The initial data is given by

$$f_k(0, x, v) = f_k^0(x, v) = \int_{I_z} f^0(x, v, z)\Phi_k(z)\pi(z)\,\mathrm{d}z. \tag{26}$$

The Maxwell boundary condition is given by

$$f_k(t, x, v) = g_k(t, x, v), \quad x \in \partial\Omega, \quad v \cdot n > 0, \tag{27}$$

with n the inward normal of $\partial\Omega$, and

$$g_k(t, x, v) := \int_{I_z} g(t, x, v, z)\Phi_k(z)\pi(z)\,\mathrm{d}z. \tag{28}$$

We consider the case where the wall temperature T_w and the accommodation coefficient α may depend on z. We assume that $\alpha(z) = \sum_{k=1}^{K} \alpha_k \Phi_k(z)$. Then

$$\begin{aligned} g(t, x, v, z) :=& (1 - \alpha(z)) f^K(t, x, v - 2(v \cdot n)n, z) \\ &+ \frac{\alpha(z)}{(2\pi)^{\frac{d_v-1}{2}} T_w(x, z)^{\frac{d_v+1}{2}}} e^{-\frac{v^2}{2T_w(x,z)}} \int_{v\cdot n<0} f^K(t, x, v, z)|v \cdot n|\,\mathrm{d}v. \end{aligned} \tag{29}$$

Substitute into (28), one gets

$$\begin{aligned} g_k =& \sum_{j=1}^{K} \left(\int_{I_z} (1 - \alpha(z))\Phi_k(z)\Phi_j(z)\pi(z)\,\mathrm{d}z \right) f_j(t, x, v - 2(v \cdot n)n) \\ &+ \sum_{j=1}^{K} D_{kj}(x, v) \int_{v\cdot n<0} f_j(t, x, v)|v \cdot n|\,\mathrm{d}v, \end{aligned} \tag{30}$$

where

$$D_{kj}(x,v)=\int_{I_z}\frac{\alpha(z)}{(2\pi)^{(d_v-1)/2}T_w(x,z)^{(d_v+1)/2}}e^{-\frac{|v|^2}{2T_w(x,z)}}\Phi_k(z)\Phi_j(z)\pi(z)\,\mathrm{d}z, \tag{31}$$

is a matrix that is time-independent hence can be pre-computed.

3 The Spatial and Time Discretization

In order to solve the Galerkin system (17), we split it into three steps:

$$\begin{cases}\partial_t f_k+v\cdot\nabla_x f_k=0,\\ \partial_t f_k+\sum_{j=1}^{K}A_{kj}(t,x)\cdot\nabla_v f_j=0,\\ \partial_t f_k=Q_k(f^K,f^K).\end{cases} \tag{32}$$

Note that each A_{kj} is a vector of length d_v. To achieve second-order accuracy in time, we use the Strang splitting and the second-order Runge–Kutta method for each step. For the transport step, we employ a second-order MUSCL scheme with the minmod slope limiter [7]. For the forcing step, we discuss the case $d_v=1$ for simplicity. The general case follows by computing the fluxes dimension by dimension. In the case $d_v=1$, for each fixed x, since (A_{kj}) is a symmetric matrix depending on x but not on v, the equation becomes a system of linear hyperbolic equations in v with constant characteristic speeds which can be solved by upwind schemes. Thus we can diagonalize the matrix A, find the Riemann invariants, and use the MUSCL scheme on each Riemann invariant. To be precise, suppose A is written as

$$A=P^{-1}DP,$$

where $P=(P_{kj})$ is an invertible matrix, and D is a diagonal matrix. Then the forcing step equations can be written as

$$\partial_t\bar{f}_k+D_{kk}\partial_v\bar{f}_k=0,$$

where $\bar{f}_k=\sum_{j=1}^{K}P_{kj}f_j$. These equations in $\bar{f}_k$ are hyperbolic with constant characteristic speeds and therefore can be solved by the MUSCL scheme. And then f_k is computed by

$$f_k=\sum_{j=1}^{K}(P^{-1})_{kj}\bar{f}_j.$$

For the collision step, we use the fast algorithm mentioned above to compute Q_k.

To choose the time step Δt, we notice first that it has to satisfy the CFL condition from the transport step, which is $\Delta t \leq \frac{\Delta x}{R_v}$, where R_v is the largest possible characteristic speed. In addition, it has to satisfy the CFL condition from the forcing step, which is $\Delta t \leq \frac{\Delta v}{C_1}$, where the constant $C_1 = \max_{x,z} |E(t, x, z)|$ is the maximum of the electric field. Furthermore, due to the parabolic nature of the FPL collision operator, one has the following constraint for the collision step $\Delta t \leq \frac{\Delta v^2}{C_2}$, where the constant $C_2 \sim \max_{x,z} \int_{\mathbb{R}^{d_v}} A(v - v_*) f(t, x, v_*, z)\, \mathrm{d}v_*$ is the maximum of the strength of diffusion of the collision operator. Thus one should choose Δt to satisfy the three restrictions.

4 Consistency Analysis of the gPC-sG Method for the Collision Operator

Here we give a consistency analysis of the gPC-sG method for the FPL collision operator. For simplicity, the random variable z is assumed to be one-dimensional in this section.

Suppose the exact solution to the spatial homogeneous FPL equation

$$\partial_t f = Q(f, f), \tag{33}$$

is

$$f(t, v, z) = \sum_{k=1}^{\infty} f_k(t, v)\Phi_k(z), \quad f_k(t, v) = \int_{I_z} f(t, v, z)\Phi_k(z)\pi(z)\, \mathrm{d}z. \tag{34}$$

Given the gPC approximation of f:

$$f \approx f^K(t, v, z) = \sum_{k=1}^{K} f_k(t, v)\Phi_k(z), \tag{35}$$

To analyze the consistency of the gPC-sG method, one substitutes the exact solution f into the scheme

$$\partial_t f_k \approx Q_k(f^K, f^K), \tag{36}$$

and estimate the difference of the LHS and the RHS. Since f solves Eq. (33), one has

$$\partial_t f_k = Q_k(f, f). \tag{37}$$

Thus it suffices to analyze $Q_k(f, f) - Q_k(f^K, f^K)$, the numerical truncation error of the collision operator. We will use the following lemma proved by Pareschi et al. [12]:

Lemma 1. *Let $g, h \in L_v^2$, then*

$$\|Q(g,h)\|_{L_v^2} \le C\|h\|_{L_v^1}\|g\|_{H_v^2}. \tag{38}$$

We estimate the error of collision operator as follows:

$$\begin{aligned}
&|Q_k(f,f) - Q_k(f^K,f^K)|^2 \\
&= \left| \int_{I_z} [Q(f,f) - Q(f^K,f^K)]\Phi_k(z)\pi(z)\,\mathrm{d}z \right|^2 \\
&\le \int_{I_z} |Q(f,f) - Q(f^K,f^K)|^2\pi(z)\,\mathrm{d}z \int_{I_z} |\Phi_k(z)|^2\pi(z)\,\mathrm{d}z.
\end{aligned}$$

Notice

$$\begin{aligned}
|Q(f,f) - Q(f^K,f^K)|^2 &= |Q(f,f-f^K) - Q(f^K-f,f^K)|^2 \\
&\le 2[|Q(f,f-f^K)|^2 + |Q(f^K-f,f^K)|^2],
\end{aligned}$$

Then one gets

$$|Q_k(f,f) - Q_k(f^K,f^K)|^2 \le 2\int_{I_z} \left[|\bar{Q}(f,f-f^K)|^2 + |\bar{Q}(f^K-f,f^K)|^2 \right] \pi(z)\,\mathrm{d}z. \tag{39}$$

Integrating in v and using the lemma, we get

$$\begin{aligned}
&\|Q_k(f,f) - Q_k(f^K,f^K)\|_{L_v^2}^2 \\
&\le C\int_{I_z} (\|f^K\|_{L_v^1}^2\|f-f^K\|_{H_v^2}^2 + \|f-f^K\|_{L_v^1}^2\|f\|_{H_v^2}^2)\pi(z)\,\mathrm{d}z \\
&\le C\int_{I_z} (\|f-f^K\|_{H_v^2}^2 + \|f-f^K\|_{L_v^1}^2)\pi(z)\,\mathrm{d}z,
\end{aligned}$$

where C will be a generic positive constant in the sequel. The second inequality above is obtained by assuming that the L_v^1 and H_v^2 norms of f are bounded, and those norms of f^K are uniformly bounded in K. Also, notice that

$$\|f - f^K\| \le C_N K^{-N}, \quad \forall N \ge 1, \tag{40}$$

in which the norms are L_v^1 or H_v^2. The term $C_N K^{-N}$ comes from the spectral accuracy of the projection operator, assuming that $f \in H_v^{N+2}$. Plug into (40), we end up with the estimate

$$\|Q_k(f,f) - Q_k(f^K,f^K)\|_{L_v^2}^2 \le C_N K^{-2N}. \tag{41}$$

which shows the spectral consistency of the gPC-sG method for the collision operator.

Remark 1. In the proof, we assume that the L^1_v and H^2_v norms of f are bounded, and those norms of f^K are uniformly bounded in K. The regularity of f for the FPL equation without the forcing term is proved by Guo [4] assuming that the initial data is close enough to the global Maxwellian in a suitable Sobolev space. The result was extended to the equation with external force by Li and Yu [8]. No result is known for the equation we consider, where the force is self-consistent. Furthermore, the regularity of f^K is completely open. However, numerically we always observe the boundedness of these norms. Therefore, these assumptions are reasonable.

5 A Remark on High-Dimensional Random Spaces

If the random space is high-dimensional, the usual gPC expansion, which requires $K = \binom{m+d}{d}$ basis functions where m is the maximal degree of polynomials and d is the dimension of the random space, can be prohibitively expensive. To handle this difficulty, we adopt the sparse technique we proposed in [15]. Using locally supported piecewise polynomials and a hierarchical construction, this technique gives a basis with $K = O((m+1)^d 2^N N^{d-1})$ basis functions, where N is the number of hierarchical levels, and m is the maximal degree of polynomials. The accuracy is $O(N^d 2^{-N(m+1)})$, which is $O(K^{-(m+1)}(\log K)^{(m+2)(d-1)})$ in terms of K.

With this sparse basis, the number of basis can still be moderately large so that the SVD method for the collision operator as well as the diagonalization of the forcing term matrix A in (32) are no longer affordable. To avoid the SVD for the collision operator computation, we follow [15] and compute $Q_k = \sum_{i,j=1}^K S_{ijk} Q(f_i, f_j)$ directly. The following sparsity result was proven: The number of pairs (i, j) for which there is at least one k with $S_{ijk} \neq 0$ is no more than $O((m+1)^{2d} 2^{2N} N^{d+1})$, compared to the total number of pairs $O((m+1)^{2d} 2^{2N} N^{2d-2})$. Only for such pairs it is required to compute $Q(f_i, f_j)$, and thus the computational cost for Q_k is still greatly reduced if N and d are large.

To avoid the diagonalization of the forcing term matrix A, we discuss the case $d_v = 1$ for simplicity. The cases with larger d_v can be treated dimension by dimension. In the case of $d_v = 1$, we use the local Lax–Friedrichs splitting for the second equation of (32) as follows:

$$\partial_t \mathbf{f} + \frac{1}{2}(A(x) - \beta(x)\mathscr{I})\partial_v \mathbf{f} + \frac{1}{2}(A(x) + \beta(x)\mathscr{I})\partial_v \mathbf{f} = 0, \tag{42}$$

where $\mathbf{f} = (f_1, \ldots, f_K)$, $\mathscr{I}$ is the identity matrix of order K, and $\beta(x)$ is a local (in each cell) upper bound of the absolute values of the eigenvalues of the symmetric matrix $A(x)$. The eigenvalues of the first flux matrix $(A(x) - \beta(x)\mathscr{I})$ are all negative, while those of the second one are all positive. Thus one can use a second-order upwind scheme with the minmod slope limiter on each flux terms without diagonalizing the matrices.

6 Numerical Results

In all the numerical examples here, except for the Landau damping, we take the physical domain to be the one-dimensional ($d_x = 1$) interval $[0, 1]$ and the velocity domain to be two-dimensional ($d_v = 2$). In all the examples, except for the six-dimensional random domain example, we take the periodic boundary condition. We discretize the physical domain into N_x grid points uniformly:

$$x_j = \left(j - \frac{1}{2}\right)\Delta x, \quad \Delta x = \frac{1}{N_x}, \quad j = 1, \ldots, N_x.$$

The velocity domain is truncated into $[-R_v, R_v]^2$ and discretized into N_v points in each dimension:

$$v_{j_1,j_2} = \left(-R_v + (j_1 - \frac{1}{2}\Delta v), -R_v + (j_2 - \frac{1}{2}\Delta v)\right), \quad \Delta v = \frac{2R_v}{N_v}, \quad j_1, j_2 = 1, \ldots, N_v.$$

R_v is big enough so $[-R_v, R_v]^2$ contains the support of the solution.

We assume the random variable z obeys uniform distribution on $[-1, 1]^d$. In the first three examples, we take $d = 1$. In the fourth example, we take $d = 2$. These examples are computed by the gPC-sG method with the gPC basis being the normalized Legendre polynomials. For the last example, $d = 6$, and we use the sparse method given in the previous section.

6.1 Random Initial Data: A Shock Tube Problem

We take the random initial data to be the equilibrium with macroscopic quantities

$$\begin{cases} \rho_l = 1 + 0.2\left(\frac{z+1}{2}\right), \quad u_l = 0, \quad T_l = 1, \quad x \le 0.5, \\ \rho_r = 0.125, \quad u_r = 0, \quad T_r = 0.25, \quad x > 0.5. \end{cases}$$

We take

$$N_x = 100, \quad N_v = 32, \quad R_v = 6, \quad K = 7, \quad \Delta t = 0.001,$$

and compute the solution at $t = 0.1$ by the sG method. The result is compared with the solution by the stochastic collocation (sC) method with the same parameters and $N_z = 10$ Gauss–Legendre quadrature points; see Fig. 1. To implement the sC method, we take N_z Gauss–Legendre quadrature points $\{z_j\}_{j=1}^{N_z}$ in the random domain and then solve the (deterministic) FPL equation at each z_j. Finally, the mean and standard deviations of any quantity f are computed by

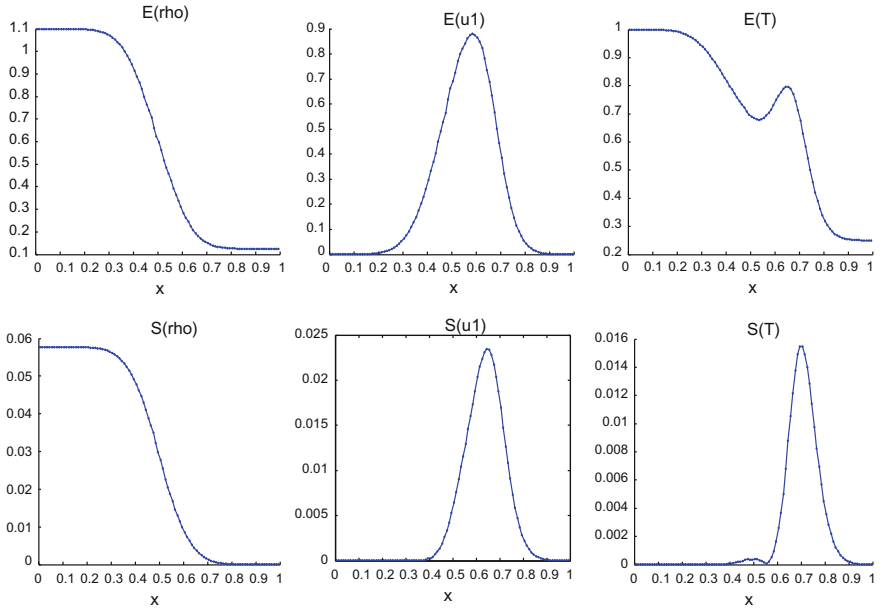


Fig. 1 Random initial data: expectation and standard deviation of macroscopic quantities. Solid line: sC, $N_x = 100$, $N_v = 32$, $R_v = 6$, $N_z = 10$, $\Delta t = 0.001$. Dots: sG, $N_x = 100$, $N_v = 32$, $R_v = 6$, $K = 7$, $\Delta t = 0.001$

$$\mathbb{E}[f] = \sum_{j=1}^{N_z} f(z_j) w_j, \quad S[f] = \sqrt{\sum_{j=1}^{N_z} f(z_j)^2 w_j - (\mathbb{E}[f])^2}, \tag{43}$$

where w_j is the quadrature weight of the point z_j. For the sC method, we verified that the solution with $N_z = 20$ quadrature points is indistinguishable with the solution with $N_z = 10$ quadrature points. Therefore, the $N_z = 10$ solution is good enough as a reference solution. This is also true for other numerical examples except the last one.

One can see from Fig. 1 that the results of two methods agree well. This shows that the sG method has good accuracy.

6.2 The Landau Damping

We use the Landau damping to test our sG method for the forcing term. For simplicity, we omit the collision term. The physical space is the interval $[0, 4\pi]$ with periodic boundary condition, and the velocity domain is one-dimensional. The random initial condition is

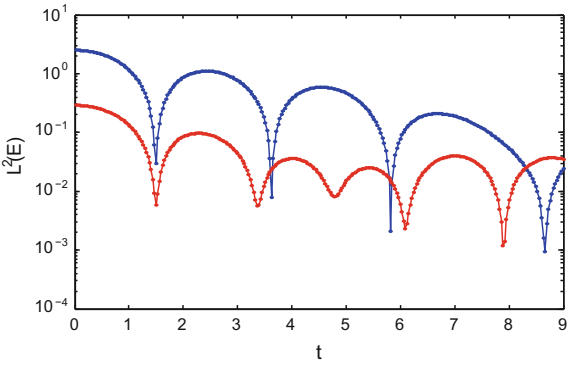


Fig. 2 Landau damping: expectation and standard deviation. Solid line: sC, $N_x = 64,\ N_v = 128,$ $R_v = 6,\ N_z = 10,$ $\Delta t = 0.03$. Dots: sG, $N_x = 64,\ N_v = 128,$ $R_v = 6,\ K = 7,\ \Delta t = 0.03.$ Upper curve: expectation. Lower curve: standard deviation

$$f^0(x, v) = \frac{1}{\sqrt{2\pi}}(1 + (0.5 + 0.1z)\cos(0.5x))e^{-\frac{|v|^2}{2}}.$$

We take

$$N_x = 64, \quad N_v = 128, \quad R_v = 6, \quad K = 7, \quad \Delta t = 0.03,$$

for the sG method and compare with the sC method with the same parameters and $N_z = 10$. We compare the expectation and standard deviation of the magnitude of the electric field for t from 0 to 9.

It can be seen from Fig. 2 that the results from two methods agree well, which shows the accuracy of the sG method. Since the uncertainty is small, the expectation is similar to the result of [14]. The standard deviation in both examples also shows oscillation in time, and this needs further theoretical explanations.

6.3 A Random Neutralizing Background

We take the deterministic initial data as the equilibrium with macroscopic quantities

$$\rho = (2 + \sin(2\pi x))/3, \quad u = (0.2, 0), \quad T = (3 + \cos(2\pi x))/4, \tag{44}$$

and the random background as

$$\mu(x, z) = \frac{2}{3}(1 + 0.2z\sin(4\pi x)). \tag{45}$$

We take

$$N_x = 100, \quad N_v = 32, \quad R_v = 8, \quad K = 7, \quad \Delta t = 0.001, \tag{46}$$

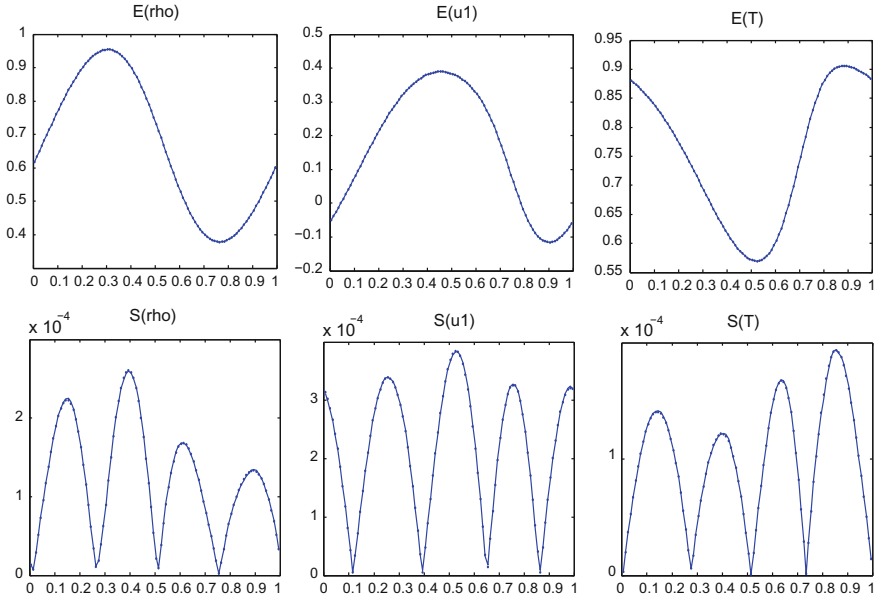


Fig. 3 Random background: expectation and standard deviation of macroscopic quantities. Solid line: sC, $N_x = 100$, $N_v = 32$, $R_v = 8$, $N_z = 10$, $\Delta t = 0.001$. Dots: sG, $N_x = 100$, $N_v = 32$, $R_v = 8$, $K = 7$, $\Delta t = 0.001$

and compare the solution by the sC method with the same parameters and $N_z = 10$ Gauss–Legendre quadrature points at $t = 0.1$. One can see from Fig. 3 that the results of two methods agree well, even for the standard deviations whose magnitude is small. This shows that the sG method can efficiently handle the uncertainties from the neutralizing background.

6.4 An Example with a Two-Dimensional Random Variable

To demonstrate that our sG method is efficient for more than one random dimension, we give a test of our method on an example with two-dimensional random domain $I_{z_1,z_2} = [-1, 1]^2$. The gPC basis is taken to be $\{\Phi_{k_1}(z_1)\Phi_{k_2}(z_2)\}$ where $\Phi_k(z)$ is the normalized Legendre polynomial of degree k, and $k_1 + k_2 \leq m$. The initial data is given by

$$f^0(x, v) = \frac{\rho^0(x)}{4\pi T^0(x)} \left(e^{-\frac{|v-u^0(x)|^2}{2T^0(x)}} + e^{-\frac{|v+u^0(x)|^2}{2T^0(x)}} \right),$$

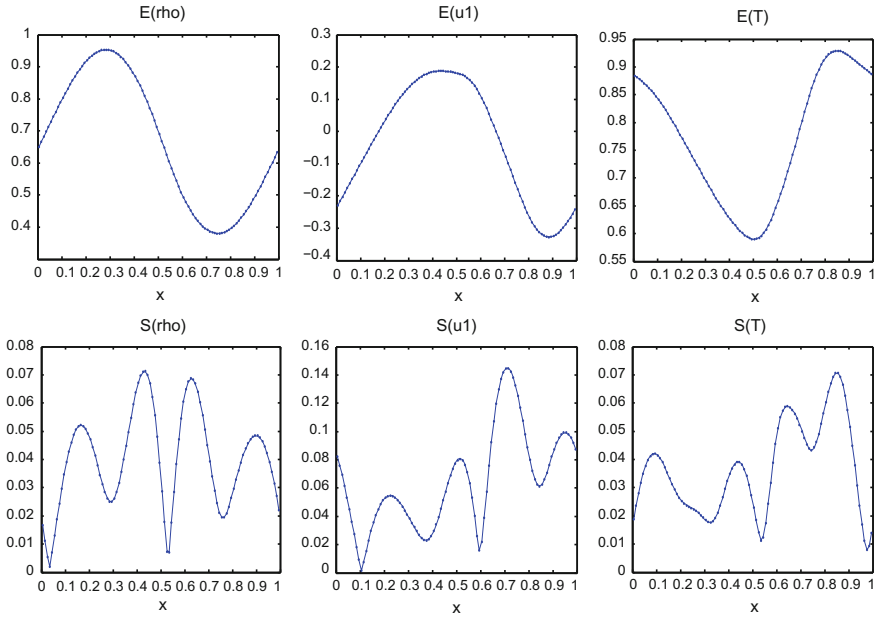


Fig. 4 Two-dimensional test with random initial data: expectation and standard deviation of macroscopic quantities. Solid line: sC, $N_x = 100,\ N_v = 32,\ R_v = 6,\ N_z = 10,\ \Delta t = 0.001$. Dots: sG, $N_x = 100,\ N_v = 32,\ R_v = 6,\ m = 5,\ \Delta t = 0.001$

where

$$\begin{cases} \rho^0(x,z) = \dfrac{1}{3}\left(2 + \sin(2\pi x) + \dfrac{1}{2}\sin(4\pi x)z_1 + \dfrac{1}{3}\sin(6\pi x)z_2\right), \\ u^0 = (0.2, 0), \\ T^0(x,z) = \dfrac{1}{4}\left(3 + \cos(2\pi x) + \dfrac{1}{2}\cos(4\pi x)z_1 + \dfrac{1}{3}\cos(6\pi x)z_2\right). \end{cases}$$

The numerical parameters are

$$N_x = 100, \quad N_v = 32, \quad R_v = 6, \quad m = 5, \quad \Delta t = 0.001,$$

and the result is compared at $t = 0.1$ with the sC method with the same parameters and $N_z = 10$ collocation points in each dimension. The result is shown in Fig. 4. It can be seen that the results of the two methods agree well, which shows the accuracy of the sG method for two-dimensional random domains.

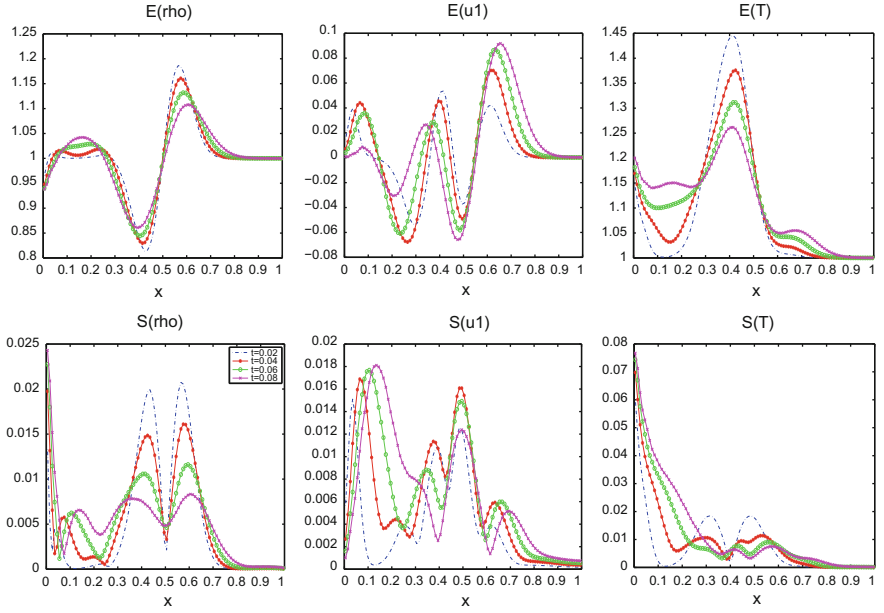


Fig. 5 Six-dimensional random domain example, using sparse sG: expectation and standard deviation of macroscopic quantities. $N_x = 100,\ N_v = 32,\ R_v = 8,\ \Delta t = 0.001,\ m = 0,\ N = 3$

6.5 An Example with a Six-Dimensional Random Domain

We finally give an example with a six-dimensional random domain. To deal with the high-dimensionality, we use the sparse sG method mentioned in Sect. 5.

We take the initial data as the equilibrium with

$$\begin{aligned} \rho(x, z) &= 1 + \exp(-100(x - 0.5)^2)\sin(10(x - 0.5))(0.5 + 0.1z_2), \\ u(x, z) &= 0, \quad T = 1 + 0.5\exp(-100(x - 0.4 - 0.01z_1)^2), \end{aligned} \tag{47}$$

and boundary data as the Maxwell boundary with

$$T_w = 1 + 0.2z_3, \quad \alpha = 0.5 + 0.3z_4. \tag{48}$$

The random background is given by

$$\mu(x, z) = 1 + 0.1z_5\sin(2\pi x) + 0.2z_6\cos(2\pi x). \tag{49}$$

We choose numerical parameters as

$$N_x = 100, \quad N_v = 32, \quad R_v = 8, \quad \Delta t = 0.001.$$

We use the sparse basis with $m = 0,\ N = 3$ to solve the equation, and the result is shown in Fig. 5. One can clearly see that near the center of the domain, the mean and standard deviation of the density and the temperature are diffused due to the kinetic transport term, and those of the velocity exhibit more complicated behavior due to the forcing term. The most interesting phenomena is that near the left boundary, the effect of the boundary condition is not influential on the mean, but is dominating the standard deviation. In fact, for all the three standard deviations, one can see that the uncertainty comes from boundary and propagates into the domain. Note that for this example with six random dimensions, with the sparse approach, only 138 basis functions are needed.

7 Conclusion

In this paper, we propose a gPC-based stochastic Galerkin method for the Fokker–Planck–Landau equation with random uncertainties. By a gPC expansion and Galerkin projection, we convert the FPL equation with uncertainty into a system of deterministic equations. We prove the consistency of the gPC-sG method for the collision operator as well as accelerate the computation of the collision kernel by a singular value decomposition combined with a fast spectral method. We adopt the sparse method from [15] to handle high-dimensional random inputs. To avoid the expensive SVD operations, we take advantage of the sparsity of the tensor $S_{b,ijk}$ for the computation of the collision operator and use a flux splitting for the mean-field term. Numerical results show the efficiency of the stochastic Galerkin method.

Acknowledgements This research was supported by NSF grants DMS-1522184 and DMS-1107291: RNMS KI-Net, by NSFC grant No. 91330203, and by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation. The first author's research was partially supported by NSF grant DMS-1620250 and a startup grant from Purdue University.

References

1. G. Dimarco, Q. Li, L. Pareschi, B. Yan, Numerical methods for plasma physics in collisional regimes. J. Plasma Phys. **81**, 1–31 (2015)
2. R.G. Ghanem, P.D. Spanos, *Stochastic Finite Elements: A Spectral Approach* (Springer, New York, 1991)
3. M.D. Gunzburger, C.G. Webster, G. Zhang, Stochastic finite element methods for partial differential equations with random input data. Acta Numer. **23**, 521–650 (2014)
4. Y. Guo, The Landau equation in a periodic box. Commun. Math. Phys. **231**(3), 391–434 (2002)
5. J. Hu, S. Jin, A stochastic Galerkin method for the Boltzmann equation with uncertainty. J. Comput. Phys. **315**, 150–168 (2016)
6. L. Landau, The transport equation in the case of the Coulomb interaction, *Collected Papers of L.D. Landau* (Pergamon Press, Oxford, 1981), pp. 163–170

7. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems* (Cambridge University Press, Cambridge, 2002)
8. F. Li, H. Yu, Decay rate of global classical solutions to the Landau equation with external force. Nonlinearity **21**, 1813–1830 (2008)
9. M. Loève, *Probability Theory*, 4th edn. (Springer, New York, 1977)
10. O.P.L. Maitre, O.M. Knio, *Spectral Methods for Uncertainty Quantification: with Applications to Computational Fluid Dynamics* (Springer, Berlin, 2010)
11. L. Pareschi, G. Russo, G. Toscani, Fast spectral methods for the Fokker-Planck-Landau collision operator. J. Comput. Phys. **165**, 216–236 (2000)
12. L. Pareschi, G. Toscani, C. Villani, Spectral methods for the non cut-off Boltzmann equation and numerical grazing collision limit. Numer. Math. **93**, 527–548 (2003)
13. M.P. Pettersson, G. Iaccarino, J. Nordstrom, *Polynomial Chaos Methods for Hyperbolic Partial Differential Equations* (Springer, Berlin, 2015)
14. J.A. Rossmanith, D.C. Seal, A positivity-preserving high-order semi-Lagrangian discontinuous Galerkin scheme for the Vlasov-Poisson equations. J. Comput. Phys. **230**, 6203–6232 (2011)
15. R. Shu, J. Hu, S. Jin, A stochastic Galerkin method for the Boltzmann equation with multi-dimensional random inputs using sparse wavelet bases. Numer. Math. Theor. Methods Appl. **10**, 465–488 (2017)
16. C. Villani, A review of mathematical topics in collisional kinetic theory, *Handbook of Mathematical Fluid Dynamics*, vol. 1 (North-Holland, Amsterdam, 2002), pp. 71–305
17. D. Xiu, *Numerical Methods for Stochastic Computations* (Princeton University Press, New Jersey, 2010)
18. D. Xiu, G.E. Karniadakis, The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002)

On Robust and Adaptive Finite Volume Methods for Steady Euler Equations

Guanghui Hu, Xucheng Meng and Tao Tang

Abstract In this paper, a robust and adaptive framework of finite volume solutions for steady Euler equations is introduced. On a given mesh, the numerical solutions evolve following the standard Godunov process and the algorithm consists of a Newton method for the linearization of the governing equations and a geometrical multigrid method for solving the derived linear system. To improve the simulations, an h-adaptive method is proposed for more efficient discretization by means of local refinement and coarsening of the mesh grids. Several numerical issues such as the regularization of the system, selection of the reconstruction patch, treatment of the curved boundary, as well as the design of the error indicator will be discussed in detail. The effectiveness of the proposed method is successfully examined on a variety of benchmark tests, and it is found that all simulations can be implemented well with one set of parameters, which shows the robustness of the method.

Keywords Steady Euler equations · Finite volume method · Adaptive method
A posteriori error estimation · Newton iteration

1 Introduction

In the study of the compressible flow, Euler equations are one fundamental governing equations and have been playing an important role in a variety of practical applica-

G. Hu
UM Zhuhai Research Institute, Zhuhai, Guangdong Province, China
e-mail: garyhu@umac.mo

G. Hu · X. Meng
University of Macau, Macao S.A.R., Macau, China
e-mail: mxc201409@gmail.com

T. Tang (✉)
Southern University of Science and Technology, Shenzhen, Guangdong Province, China
e-mail: tangt@sustc.edu.cn

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_2

tions such as optimal design of the vehicle shape [15], physical-based simulations in animation [31].

Steady-state flow is a typical phenomenon in the fluid dynamics in which the distributions of the physical quantities will not change with the time evolution. Such phenomena exist in several realistic fluid dynamics applications. For example, when an aeroplane is in its cruise state in the stratosphere, the fluid dynamics around the aeroplane can be described reasonably by the steady state. The theoretical and numerical studies on the steady-state flow have great importance on the applications such as the optimal design on the vehicle shape. In a classical optimization framework for the optimal design, the objective functional is optimized subject to several shape parameters. In the whole simulation, dozens of, or maybe hundreds of, steady-state flows need to be determined with different configurations. Hence, efficiency of the steady-state solver becomes crucial in the practical simulations.

Although there have been lots of work available in the market for solving steady Euler equations by using finite difference methods [54], finite element methods [16], spectral methods [28], the existence of the discontinuous solutions such as shock and contact discontinuity makes the use of the finite volume methods [29, 33], discontinuous Galerkin methods [10], spectral volume methods [51] more competitive. Besides the ability on representing discontinuous solutions, these methods also introduce the flux to preserve the conservation property of the simulation, which makes these methods more attractive towards delivering physical simulations. It is worth mentioning that, recently, the fast sweeping method [12, 13] was proposed to solve steady Euler equations, and excellent numerical results were obtained. In our previous works [21–26], an adaptive framework of finite volume solutions has been developed for solving steady Euler equations.

There are several challenges on developing quality high-order finite volume methods for solving Euler equations. One of the most important challenges is the solution reconstruction. In the original Godunov scheme, the cell average is used directly to evaluate the numerical flux. The advantage of Godunov is very attractive, i.e. the maximum principle can be preserved naturally. However, the piecewise constant approximation makes the scheme too dissipative to generate high-resolution solution; hence, the solution variation needs to be recovered to deliver high-order approximation for the exact solution. In the solution reconstruction, a nontrivial issue is to develop quality limiter functions to restrain the possible nonphysical oscillation, which is listed in [52] as one challenge for developing high-order numerical methods for computational fluid dynamics. Another challenge is efficiency of the algorithm. By propagating the time-dependent system for sufficiently long time is obviously not a good idea for obtaining the steady state of the system since the low efficiency. To effectively accelerate the simulation, several classical techniques such as local time-stepping, enthalpy damping, residual smoothing, multigrid methods and preconditioning techniques [6] have been developed and applied. Towards the efficient discretization of the governing equations, adaptive methods such as r-adaptive methods [37, 38, 46], h-adaptive methods [5, 18, 39, 43], and hp-adaptive methods [19, 50] have been developed and still attract more and more research attention. Nowadays, with the dramatic development of the computer hardware, the capacity of the

high-performance computing cluster is also improved significantly. Hence, parallel algorithms based on OpenMP [1], OpenMPI [2] as well as GPU [53] become more and more popular in the community of computational fluid dynamics [34].

In this paper, we introduce an adaptive framework of finite volume solutions for the steady Euler equations. On a given mesh, the solver consists of a Newton iteration for the linearization of the governing equations and a geometrical multigrid method for solving the linear system. To resolve the issue on the quality high-order solution reconstruction, the non-oscillatory k-exact reconstruction is proposed which provides a unified strategy for high-order reconstruction. To handle the efficiency issue, h-adaptive method is introduced in our method and an adjoint-based a posteriori error estimation method is developed to generate quality error indicator. Some numerical issues such as regularization of the linearized system are also introduced. Numerical tests successfully show the robustness and effectiveness of the proposed method.

The rest of the paper is organized as follows. In Sect. 2, the steady Euler equations and finite volume discretization are introduced. In Sect. 3, the solution reconstruction will be introduced and the non-oscillatory k-exact reconstruction method will be described in detail. In Sect. 4, our methods on partially resolving the efficiency issue of the simulations are summarized and the adjoint weighted residual indicator as well as implementation are introduced in detail. Three numerical tests are delivered in Sect. 5 in which the robustness and effectiveness of the proposed framework are successfully demonstrated. Finally, the conclusion is given.

2 Finite Volume Framework for Steady Euler Equations

2.1 Governing Equations

The inviscid two-dimensional steady Euler equations are given as

$$\nabla \cdot F(U) = 0, \tag{1}$$

where U and $F(U)$ denote the conservative variables and flux given by

$$U = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix}, \quad \text{and } F(U) = \begin{bmatrix} \rho u & \rho v \\ \rho u^2 + p & \rho uv \\ \rho uv & \rho v^2 + p \\ u(E+p) & v(E+p) \end{bmatrix}, \tag{2}$$

respectively. Here $(u, v)^T$, ρ, p, and E denote the velocity, density, pressure, and total energy, respectively. To close the system, we use the following equation of state in this paper,

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho(u^2 + v^2), \tag{3}$$

where $\gamma = 1.4$ is the ratio of the specific heats of the perfect gas.

Before we get involved in the numerical methods for solving (1), let us introduce the notations as follows to facilitate the description. The computational domain is denoted by Ω, and $\mathscr{T} = \{\mathscr{K}_i\}, i = 1, 2, \ldots, N_{tri}$ is its associated triangulation in which $\mathscr{K}_i$ is the ith triangle in the mesh, and N_{tri} is the total number of the triangle elements in the mesh.

2.2 Newton Linearization

Certain linearization is needed since the nonlinearity of the governing Eq. (1), and Newton iteration is employed in our work. Below we would briefly summarize the implementation of the Newton iteration on our problem. People may refer to [21, 23, 24, 26, 39] for the details.

The governing Eq. (1) is discretized as follows. First of all, the integral form of (1) on Ω is given by

$$\int_{\Omega} \nabla \cdot F(U) dxdy = \sum_{\mathscr{K}_i} \int_{\mathscr{K}_i} \nabla \cdot F(U) dxdy = 0. \tag{4}$$

Then Green's theorem gives the following equation,

$$\sum_{\mathscr{K}_i} \sum_{e_{i,j} \in \partial \mathscr{K}_i} \int_{e_{i,j}} F(U) \cdot n_{i,j} ds = 0, \tag{5}$$

where $e_{i,j}$ means the common edge of the element $\mathscr{K}_i$ and its neighbour element $\mathscr{K}_j$, and $n_{i,j}$ means the unit out normal vector of $e_{i,j}$ with respect to the element $\mathscr{K}_i$. In the simulation, numerical flux $\bar{F}(U_i, U_j)$ is used to replace the unknown flux $F(U)$. Hence, the above equations are approximated by the following ones

$$\sum_{\mathscr{K}_i} \sum_{e_{i,j} \in \partial \mathscr{K}_i} \int_{e_{i,j}} \bar{F}(U_i, U_j) \cdot n_{i,j} ds = 0. \tag{6}$$

To resolve the nonlinearity of (6), Newton method is employed here. We assume that the approximation of the solution at the kth step, $U^{(k)}$, is known, and then the approximation of the solution at the $(k+1)$th step, $U^{(k+1)} = U^{(k)} + \Delta U^{(k)}$, can be found by solving

$$\sum_{\mathscr{K}_i}\sum_{e_{i,j}\in\partial\mathscr{K}_i}\int_{e_{i,j}}\bar{F}(U_i^{(k+1)},U_j^{(k+1)})\cdot n_{i,j}ds = \sum_{\mathscr{K}_i}\sum_{e_{i,j}\in\partial\mathscr{K}_i}\int_{e_{i,j}}\bar{F}(U_i^{(k)}+\Delta U_i^{(k)},U_j^{(k)}+\Delta U_j^{(k)})\cdot n_{i,j}ds = 0, \tag{7}$$

for $\Delta U_i^{(k)}$ which is increment of the conserved quantity on the element $\mathscr{K}_i$ to the kth approximation of the solutions. By Taylor theorem and only keeping the linear part, the linear system for ΔU can be written as

$$\begin{aligned}\sum_{\mathscr{K}_i}\sum_{e_{i,j}\in\partial\mathscr{K}_i}\int_{e_{i,j}}\frac{\partial\bar{F}}{\partial U_i}\cdot n_{i,j}ds\Delta U_i^{(k)}+\sum_{\mathscr{K}_i}\sum_{e_{i,j}\in\partial\mathscr{K}_i}\int_{e_{i,j}}\frac{\partial\bar{F}}{\partial U_j}\cdot n_{i,j}ds\Delta U_j^{(k)}\\ =-\sum_{\mathscr{K}_i}\sum_{e_{i,j}\in\partial\mathscr{K}_i}\int_{e_{i,j}}\bar{F}(U_i^{(k)},U_j^{(k)})\cdot n_{i,j}ds.\end{aligned} \tag{8}$$

Regularization is necessary to solve the linear system (8). The issue is resolved by introducing the local residual $LR_i=\sum_{e_{i,j}\in\partial\mathscr{K}_i}\int_{e_{i,j}}\bar{F}(U_i^{(k)},U_j^{(k)})\cdot n_{i,j}ds$, i.e. the regularized system is written as

$$\begin{aligned}\alpha\sum_{\mathscr{K}_i}||LR_i||_1\Delta U_i^{(k)}+\sum_{\mathscr{K}_i}\sum_{e_{i,j}\in\partial\mathscr{K}_i}\int_{e_{i,j}}\frac{\partial\bar{F}}{\partial U_i}\cdot n_{i,j}ds\Delta U_i^{(k)}\\ +\sum_{\mathscr{K}_i}\sum_{e_{i,j}\in\partial\mathscr{K}_i}\int_{e_{i,j}}\frac{\partial\bar{F}}{\partial U_j}\cdot n_{i,j}ds\Delta U_j^{(k)}=-\sum_i LR_i,\end{aligned} \tag{9}$$

where $||\cdot||_1$ is the l_1 norm, and $\alpha>0$ is a parameter to weight the regularization.

So far, the only unknown quantity in (9) is the numerical flux $\bar{F}$. In the simulation, this quantity is obtained by solving a local Riemann problem in which the left and right states are determined by the solutions in the element $\mathscr{K}_i$ and its neighbour $\mathscr{K}_j$. There are several Riemann solvers available in the market, and HLLC [48] is used in our simulations.

A natural choice for the left and right states for Riemann problem is the cell average of each conserved quantity. In this case, a piecewise constant approximation of the conserved quantity is supposed, and only first-order numerical accuracy can be expected. To improve the numerical accuracy, more accurate left and right states in Riemann problem are desired and this can be achieved by high-order solution reconstruction.

3 Solution Reconstruction

With the assumption of sufficient regularity, Taylor theorem gives the following substitution for the unknown function $U(x, y)$ in the element $\mathscr{K}$

$$\begin{aligned} U(x,y) = U(x_{\mathscr{K}}, y_{\mathscr{K}}) &+ \frac{\partial U}{\partial x}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(x - x_{\mathscr{K}}) + \frac{\partial U}{\partial y}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(y - y_{\mathscr{K}}) \\ &+ \frac{1}{2}\frac{\partial^2 U}{\partial x^2}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(x - x_{\mathscr{K}})^2 + \frac{\partial^2 U}{\partial x \partial y}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(x - x_{\mathscr{K}})(y - y_{\mathscr{K}}) \\ &+ \frac{1}{2}\frac{\partial^2 U}{\partial y^2}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(y - y_{\mathscr{K}})^2 \\ &+ \cdots, \end{aligned} \tag{10}$$

where $(x_{\mathscr{K}_i}, y_{\mathscr{K}_i})$ is the barycentre of the element $\mathscr{K}_i$. The task of the reconstruction is to recover those coefficients $\partial^\alpha U/(\partial x^{\alpha_1}\partial y^{\alpha_2})$, $\alpha = \alpha_1 + \alpha_2$, with the cell average $\bar{U}_i = 1/|\mathscr{K}_i| \int_{\mathscr{K}_i} U(x, y)dxdy$ of the conserved quantity $U(x, y)$ in the element $\mathscr{K}_i$, where $|\mathscr{K}_i|$ is the area of the element $\mathscr{K}_i$.

The most popular reconstruction in the market is the linear reconstruction, i.e.

$$U(x,y) \approx U(x_{\mathscr{K}}, y_{\mathscr{K}}) + \frac{\partial U}{\partial x}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(x - x_{\mathscr{K}}) + \frac{\partial U}{\partial y}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(y - y_{\mathscr{K}}) := P^1(x,y). \tag{11}$$

It is noted that with the assumption of the linear distribution of $U(x, y)$ in $\mathscr{K}_i$, the constant term in (11) is the cell average, i.e. $U(x_{\mathscr{K}_i}, y_{\mathscr{K}_i}) = \bar{U}_i$. Hence, the linear reconstruction is to recover the gradient of $U(x, y)$ in $\mathscr{K}_i$. There are two ways to evaluate the gradient $\nabla U = (\partial U/\partial x, \partial U/\partial y)^T$. One is the following Green–Gauss theorem [6],

$$\int_{\mathscr{K}_i} \nabla U dxdy = \int_{\partial \mathscr{K}_i} Unds. \tag{12}$$

Since the linearity of U, ∇U is a constant. Hence,

$$\nabla U|_{\mathscr{K}_i} = \frac{1}{|\mathscr{K}_i|}\int_{\partial \mathscr{K}_i} Unds. \tag{13}$$

Replacing U on the edge $e_{i,j}$ by using the average $(\bar{U}_i + \bar{U}_j)/2$, the above integral can be approximated by

$$\nabla U|_{\mathscr{K}_i} \approx \frac{1}{|\mathscr{K}_i|}\sum_{e_{i,j}} \frac{1}{2}(\bar{U}_i + \bar{U}_j)n_{i,j}|e_{i,j}|. \tag{14}$$

The implementation of Green–Gauss approach is quite simple. However, the numerical accuracy of such approximation heavily depends on the regularity of the mesh grids. Also, it is not trivial to extend the method to the high-order cases. People may refer to [11] for the quadratic reconstruction with Green–Gauss method.

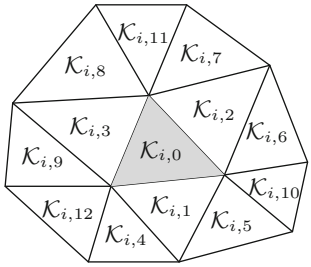


Fig. 1 Reconstruction patch for the element $\mathscr{K}_{i,0}$

To overcome the above issues, the least square method becomes a very competitive candidate on solution reconstruction since its ability on delivering accurate solution even on skewed unstructured grids and on natural extension to high-order cases. To implement the least square reconstruction on the element $\mathscr{K}_i$, a reconstruction patch $\mathscr{P}_i$ is needed first. In the case of the linear reconstruction, a natural choice for $\mathscr{P}_i$ is $\mathscr{K}_i$ itself as well as its three Neumann neighbours. For example, for the element $\mathscr{K}_{i,0} = \mathscr{K}_i$ in Fig. 1, the patch of the linear reconstruction for it can be chosen as $\mathscr{P}_i = \{\mathscr{K}_{i,0}, \mathscr{K}_{i,1}, \mathscr{K}_{i,2}, \mathscr{K}_{i,3}\}$.

With $\mathscr{P}_i$ for $\mathscr{K}_i$, the gradient $\nabla U|_{\mathscr{K}_{i,0}} = (\partial U/\partial x|_{\mathscr{K}_{i,0}}, \partial U/\partial y|_{\mathscr{K}_{i,0}})^T$ can be solved from the following minimization problem,

$$\underset{\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}}{\operatorname{argmin}} \sum_{\mathscr{K}_j \in \mathscr{P}_i, \mathscr{K}_j \neq \mathscr{K}_{i,0}} ||P_i^1(x_{\mathscr{K}_j}, y_{\mathscr{K}_j}) - \bar{U}|_{\mathscr{K}_j}||_2^2. \tag{15}$$

The extension to the high-order reconstruction is straightforward for the least square approach. In the case of quadratic reconstruction, a larger patch containing at least 6 elements is needed since there are more unknowns included in (10). A method to enlarge $\mathscr{P}_i$ is to introduce Neumann neighbours of the Neumann neighbours of $\mathscr{K}_i$. However, it is found that generating $\mathscr{P}_i$ by selecting $\mathscr{K}_i$ and its Moore neighbours is a better choice, especially when the adaptive strategy is used in the simulation, based on our numerical experience. In this case, the patch $\mathscr{P}_i$ becomes

$$\mathscr{P}_i = \{\mathscr{K}_{i,0}, \mathscr{K}_{i,1}, \mathscr{K}_{i,2}, \mathscr{K}_{i,3}, \mathscr{K}_{i,4}, \mathscr{K}_{i,5}, \mathscr{K}_{i,6}, \mathscr{K}_{i,7}, \mathscr{K}_{i,8}, \mathscr{K}_{i,9}, \mathscr{K}_{i,10}, \mathscr{K}_{i,11}, \mathscr{K}_{i,12}\}.$$

Now the unknown quantity $U(x, y)$ is approximated by

$$\begin{aligned} U(x, y) \approx\ & P^1(x, y) + \frac{1}{2}\frac{\partial^2 U}{\partial x^2}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(x - x_{\mathscr{K}})^2 + \frac{\partial^2 U}{\partial x \partial y}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(x - x_{\mathscr{K}})(y - y_{\mathscr{K}}) \\ & + \frac{1}{2}\frac{\partial^2 U}{\partial y^2}|_{x_{\mathscr{K}}, y_{\mathscr{K}}}(y - y_{\mathscr{K}})^2 \\ :=\ & P^2(x, y) \end{aligned} \tag{16}$$

To preserve the conservative property of the reconstructed polynomial, the minimization problem we need to solve becomes

$$\underset{\frac{\partial U}{\partial x},\frac{\partial U}{\partial y},\frac{\partial^2 U}{\partial x^2},\frac{\partial^2 U}{\partial y},\frac{\partial^2 U}{\partial y^2}}{\operatorname{argmin}} \sum_{\mathscr{K}_j \in \mathscr{P}_i} \left\| \frac{1}{|\mathscr{K}_j|} \int_{\mathscr{K}_j} P_i^2(x,y)dxdy - \bar{U}|_{\mathscr{K}_j} \right\|_2^2. \tag{17}$$

Remark 1 The above method is k-exact reconstruction [3]. To solve (17) directly, a large amount of integrals need to be evaluated during the reconstruction. In [42], a numerical trick is introduced to effectively save the computational resource. In the trick, several integrals are calculated beforehand, and then the linear system consists of those integrals by algebraic operations. Recently, the parallel k-exact reconstruction is developed [17], which significantly improves the efficiency of the reconstruction.

Remark 2 The conservative of U in $\mathscr{K}_i$ cannot be guaranteed strictly by solving (17) in the least square sense. To preserve the conservative property rigorously, the constant term in $P_i^2(x,y)$ is adjusted to make $\frac{1}{|\mathscr{K}_{i,0}|}\int_{\mathscr{K}_{i,0}} P_i^2(x,y)dxdy = \bar{U}_i$.

For all high-order reconstructions ($\geq$ linear reconstructions), limiting process is necessary to restrain the nonphysical oscillation, especially when there is shock in the solution. For linear reconstruction, there are several mature limiters available for the unstructured meshes such as the limiter of Barth and Jespersen [4], and the limiter of Venkatakrishnan [49]. Compared with the limiter of Barth and Jespersen, the limiter of Venkatakrishnan has better property towards the differentiability; hence, it has better performance on the steady-state convergence. Although these limiters work well for the linear reconstruction, it is nontrivial for the higher-order extension. People may refer to [41] for the contribution towards this direction. It is worth mentioning that quality limiter for high-order methods was listed as one of the challenges in developing high-order numerical methods for computational fluid dynamics in [52].

Weighted essentially non-oscillatory (WENO) scheme is well known for its ability on delivering high-order and non-oscillatory numerical solutions [30, 55]. For WENO implementation on unstructured meshes, people may refer to [30] for details. Besides the solution reconstruction, WENO has been also used as a limiter in the discontinuous Galerkin framework [40, 44, 45, 56]. In our works [21–26], WENO reconstruction is introduced for the solution reconstruction. Below is a brief summarization for the WENO reconstruction with the assumption of the locally linear distribution of the solutions.

In WENO reconstruction, besides the reconstruction patch $\mathscr{P}_{i,0} = \mathscr{P}_i$ for $\mathscr{K}_{i,0}$ in Fig. 1, we also solve the optimization problem (15) on patches $\mathscr{P}_{i,1} = \{\mathscr{K}_{i,0}, \mathscr{K}_{i,1}, \mathscr{K}_{i,4}, \mathscr{K}_{i,5}\}$, $\mathscr{P}_{i,2} = \{\mathscr{K}_{i,0}, \mathscr{K}_{i,2}, \mathscr{K}_{i,6}, \mathscr{K}_{i,7}\}$, and $\mathscr{P}_{i,3} = \{\mathscr{K}_{i,0}, \mathscr{K}_{i,3}, \mathscr{K}_{i,8}, \mathscr{K}_{i,9}\}$. Correspondingly, besides the polynomial $P_{i,0}^1 = P_i$ from $\mathscr{P}_{i,0}$, we also have the candidate polynomials $P_{i,1}^1$, $P_{i,2}^1$, $P_{i,3}^1$ from $\mathscr{P}_{i,1}$, $\mathscr{P}_{i,2}$ and $\mathscr{P}_{i,3}$, respectively. For each candidate $P_{i,j}^1$, $j = 0, 1, 2, 3$, a smoothness indicator is defined by

$$S_j = \left((\frac{\partial U}{\partial x}|_j)^2 + (\frac{\partial U}{\partial y}|_j)^2 \right) |\mathscr{K}_{j,0}|. \tag{18}$$

Then the weight for each polynomial is calculated by

$$\omega_j = \frac{\tilde{\omega}_j}{\sum_k \tilde{\omega}_k}, \quad \tilde{\omega}_j = \frac{1}{(\epsilon + S_j)^2}, \tag{19}$$

and the final polynomial for the element $\mathscr{K}_i$ is given by

$$P_i^1 = \sum_j \omega_j P_{i,j}^1. \tag{20}$$

Remark 3 In the definition of $\tilde{\omega}_j$ in (19), a parameter γ_j [20, 30] is used as the numerator. γ_j there is designed for preserving the higher order accuracy of P_i^1, i.e. $P_i^1(x_{GQ}, y_{GQ}) = P_i^2(x_{GQ}, y_{GQ})$ where $P_i^2(x, y)$ is a quadratic polynomial obtained by solving (16). With γ_j and the nonlinear weight ω_j, the reconstructed polynomial P_i can preserve the third-order numerical accuracy and restrain the nonphysical oscillation effectively in the meantime [20, 30]. However, an extra quadratic reconstruction problem (16) as well as the parameters γ_j need to be calculated, which would slow down the simulation efficiency. In our algorithm, the numerator 1 is used instead of γ_j to avoid the extra calculations and the h-adaptive method is introduced to remedy the accuracy issue.

The WENO reconstruction can be extended to higher order directly. People may refer to [25, 26] for our works on non-oscillatory k-exact reconstruction.

In the traditional reconstructions, the polynomial is obtained by certain method first, and then the limiter is introduced to remove or restrain the possible oscillation. Recently, Chen and Li developed an integrated linear reconstruction (ILR) method [8] in which an optimization method is proposed and solved locally for each element to construct the polynomial. The advantages of ILR include (i) the reconstruction can be finished by solving a single problem, i.e. the reconstructing and limiting processes are combined together, (ii) the local maximum principle is preserved theoretically by ILR, and (iii) no parameter is used in the reconstruction. An improved ILR method can be found in the paper [7].

4 Towards Efficiency

Efficiency is crucial for an algorithm in its practical applications. Since the Newton iteration is used for the linearization, a series of linearized system need to be solved in solving a steady Euler system, which means that the efficiency of the linear solver is important for an efficient simulation. Furthermore, in one of the most important applications for steady Euler solver, i.e. the optimal design of the vehicle shape, a series of steady Euler systems with different configurations need to be solved in a single design process. Hence, how to improve the efficiency of the steady Euler solver is also worth studying in detail.

For the first issue, a geometrical multigrid solver is developed to solve the linearized system in our algorithm [21, 23–26, 39]. In this geometrical multigrid solver,

the coarse meshes are generated by the volume agglomeration method [6, 32]. Then the error on the coarse meshes is smoothed by blocked lower-upper Gauss–Seidel method proposed in [9]. People may refer to our works for the details of the implementation and performance of the solver.

To resolve the second issue mentioned above, the algorithm can be improved from the following aspects. First of all, it is the acceleration of the convergence of the Newton iteration. In (9), the local residual of the system is used to regularize the system. It is noted that this is a similar acceleration technique to the local time-stepping method [6]. In both methods, local information is used to improve the simulation. In local time-stepping method, the time-dependent Euler equations are solved and the Courant–Friedrichs–Lewy (CFL) number is chosen locally depending on the characteristic speed in the current control volume; hence, the evolution of the system is not uniform in the whole flow field. In the region with low characteristic speed, a larger CFL number can be chosen to speedup the convergence to the steady state. In our method, there is no temporal term in the equations and we use local residual to regularize the system. If the system is far from the steady state locally, the local residual is a large quantity, which corresponds to effect in solving time-dependent problem with a small CFL number. On the other hand, local residual would be a small quantity when the system is close to the steady state locally which corresponds to the large CFL number case. Based on our numerical experience, the local residual regularization works very well in all cases and the simulations are not sensitive to the selection of the parameter α in (9).

The second way to improve the implementation efficiency is to develop efficient discretization. In the case that there is large variation of the solution in the domain, especially there is shock in the solution, numerical discretization on a uniform mesh is obviously not a good idea since too many mesh grids are wasted in the region with gentle solution. In the market, adaptive mesh methods are popular towards the efficient and nonuniform discretization of the governing equations. For example, r-adaptive methods have been successfully used in solving Euler equations [27, 36–38, 46, 47]. In our algorithm, h-adaptive methods are introduced towards the efficient numerical discretization [21, 22, 25, 26, 39]. To handle the local refinement or coarsening of the mesh grids efficiently, an hierarchy geometry tree (HGT) is developed. People may refer to [35] for HGT details. It is worth mentioning that CPU time on local refinement or coarsening is nothing compared with the whole CPU time in the simulation with HGT.

Another important component in adaptive method is the error indicator. The quality of the error indicator determines the quality of the nonuniform discretization. There are basically two types error indicators in the market. One is feature-based error indicators which depend on the numerical solution, and the other one is error indicators based on the a posteriori error estimation. In our works, several feature-based error indicators are tested in the h-adaptive framework such as the gradient of the pressure [21, 26, 39] and entropy [21, 26]. Recently, an adjoint-based a posteriori error estimation method is developed towards minimizing the numerical error of a quantity of interest [25]. Adjoint-based analysis is a very useful tool in the applications of optimal design of vehicle shape [15] and the error estimation [14]. Below is

a brief summary of our adjoint-based error indicator, and people may refer to [25] for the details.

Suppose that U^H is the solution on the mesh $\mathscr{T}^H$, and $J(U^H)$ is the quantity of interest. In the practical applications, the quantity of interest $J(U^H)$ could be the drag or lift in the simulations of flow through an airfoil, or other application-related quantities. Now, we are interested in the error of $J(U^H)$, i.e. $J(U) - J(U^H)$ where $J(U)$ is the exact evaluation of the quantity of interest depending on the exact solution U. In most cases, $J(U)$ is nonlinear. Then the linearization of the difference gives

$$J(U) - J(U^H) \approx \frac{\partial J}{\partial U}(U - U^H). \tag{21}$$

By defining the residual $R(U) := \nabla \cdot F(U)$, the linearization of the difference between the exact residual and approximate residual gives

$$R(U) - R(U^H) \approx \frac{\partial R}{\partial U}(U - U^H), \tag{22}$$

which follows

$$U - U^H \approx \left(\frac{\partial R}{\partial U}\right)^{-1} (R(U) - R(U^H)). \tag{23}$$

By plugging the above expression into (21), we get

$$J(U) - J(U^H) \approx \frac{\partial J}{\partial U}\left(\frac{\partial R}{\partial U}\right)^{-1} (R(U) - R(U^H)) := \psi^T (R(U) - R(U^H)), \tag{24}$$

where the adjoint ψ^T can be obtained by solving

$$\left(\frac{\partial R}{\partial U}\right)^T \psi = \frac{\partial J}{\partial U}. \tag{25}$$

The implementation in [25] is as follows. First, the mesh $\mathscr{T}^H$ is uniformly refined one time to get the new mesh $\mathscr{T}^h$. Then the solution U^H on $\mathscr{T}^H$ is interpolated onto $\mathscr{T}^h$ to get an approximation U_h^H which is used in (24) to replace U. Since we assume that the system is solved completely on $\mathscr{T}^H$, the quantity $R(U^H)$ can be reasonably ignored in (24). There are two ways mentioned in [25] to solve the adjoint problem (25). One is to evaluate two Jacobian matrices in (25) on $\mathscr{T}^h$ first, and then the equation is solved on $\mathscr{T}^h$. The other one is to do the same thing on $\mathscr{T}^H$. Compared with the former one, the advantage of the latter strategy is that the size of the system is much smaller, i.e. the size is only 25% of the one in former case. Furthermore, since U^H is a quality approximation to U on $\mathscr{T}^H$, the linear problem (25) can be solved smoothly. It is noted that based on our numerical experience, direct evaluation of $\partial J/\partial U$ and $\partial R/\partial U$ on $\mathscr{T}^h$ with the interpolation approximation U_h^H would bring difficulty on solving (25) and several Newton iterations for (9) with U_h^H as the initial

guess are necessary for the improvement. On the other hand, the disadvantage of the latter strategy is that the convergence order of the numerical method will be sacrificed a little bit. This is understandable since the information from the finer mesh would generate more accurate error estimation.

The third strategy to improve the efficiency of our steady Euler solver is to resort to the parallel computing. Since the operations on solution reconstruction, evaluation of the numerical flux, and the cell average update are local, OpenMP [1] has been introduced to realize the parallel computing on these operations in [22] in which a reactive Euler system is solved to simulate detonation. To handle large-scale simulations, the parallelization based on MPI becomes necessary. We are working on the parallelization of our algorithm based on domain decomposition method and OpenMPI [2], and the results will be reported in the forthcoming paper.

5 Numerical Tests

In this section, the following three numerical tests will be implemented to demonstrate the effectiveness of our method,

- Subsonic flow around a circular cylinder,
- Inviscid flow through a channel with a smooth bump,
- Transonic flow around a NACA 0012 airfoil.

All simulations in this paper are supported by AFVM4CFD [21–26, 39] which is a C++ library developed and maintained by the authors and collaborators.

5.1 Subsonic Flow Through a Circular Cylinder

In this section, the subsonic flow passing a circular cylinder is simulated. The computational domain is a ring, and the radii for the inner and outer circles are 0.5 and 20, respectively. The configuration of the flow in the far field is as follows. The density is 1, the Mach number is 0.38, the velocity vector is $(\cos\theta, \sin\theta)^T$ where θ is attack angle and $\theta = 0^\circ$ in this case. The configuration for far field flow is also used as the initial condition for our Newton iteration.

The method with non-oscillatory 2-exact reconstruction is implemented on five meshes with 240, 504, 800, 1776, and 3008 grid points, respectively. Since the flow in the domain is subsonic, inviscid, and vortex free, the entropy of the flow should be a constant same to that in the far field. Hence, we use the L_2 error of the entropy production to evaluate the convergence of the method which shown

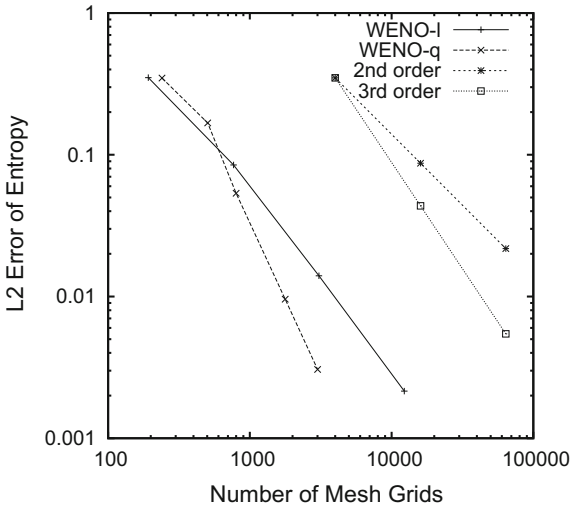


Fig. 2 Convergence curves for the inviscid flow through the circle

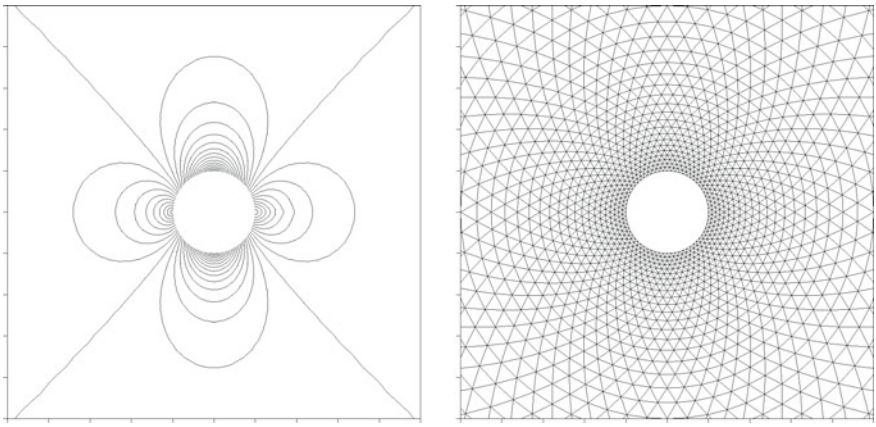


Fig. 3 Left: The Mach number isolines generated with WENO 2-exact reconstruction. Right: The corresponding mesh

in Fig. 2. As a comparison, the results obtained with linear reconstruction in [24] are also demonstrated here. It can be observed from the figure that both linear and quadratic methods successfully generate theoretical convergence curves. The mesh grids around the inner circle as well as the isolines of the Mach number can be observed from Fig. 3.

5.2 *Inviscid Flow Through a Channel with a Smooth Bump*

In this subsection, the inviscid flow through a channel with a smooth bump is simulated by adaptive method with non-oscillatory 2-exact reconstruction. This test is a benchmark test listed in [52] in which the detailed setup for the simulation can be found.

In Fig. 4, the following three results are shown. The first result is the convergence curve generated on four successively and uniformly refined meshes. It can be observed obviously the theoretical curve is recovered very well. The second result is the convergence curve generated by adaptive method with error indicator obtained only by the local residual. It is observed that the adaptive method generates much better convergence curve, compared with the one generated by uniformly refining the mesh. The nonuniform distribution of the mesh grids with 5940 points as well as the corresponding isolines of the Mach number can be observed from Fig. 5 (bottom). The third result is the convergence curve generated by adaptive method with error indicator obtained by adjoint weighted residual. In the simulation, the following functional is used as the quantity of interest,

$$\mathscr{J}(U) = \frac{1}{|\Omega|}\int_{\Omega} \frac{|s_\infty - s|}{s_\infty} dxdy, \tag{26}$$

where $s_\infty = p_\infty/\rho_\infty^\gamma$ is the far field entropy, and p_∞ and ρ_∞ are the far field pressure and density, respectively. From Fig. 4, it can be observed that adjoint weighted residual gives the best convergence result among three results. In Fig. 5 (top), the distribution of the mesh grids with 3387 points and the isolines of Mach number are shown with adjoint weighted residual method. It can be seen that the adjoint method

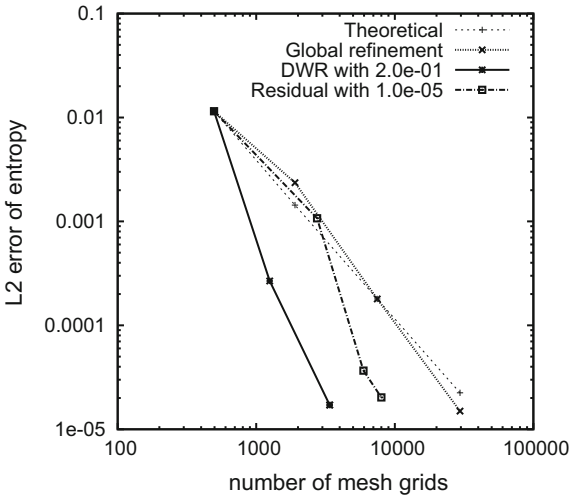


Fig. 4 Convergence curves of the entropy production with different methods

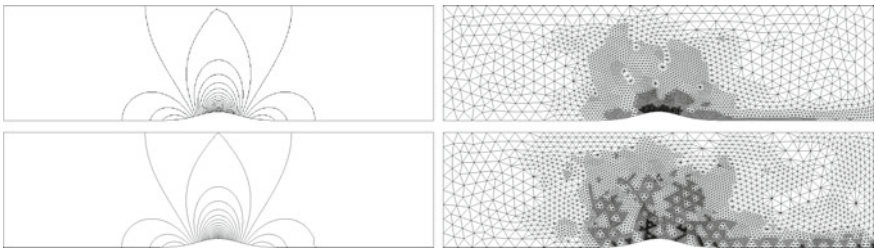


Fig. 5 Mesh profiles and Mach isolines obtained from adjoint-based mesh adaptation with 3387 mesh grids (upper row) and residual-based mesh adaptation with 5940 mesh grids (lower row)

helps to assign more mesh grids in the region in which the entropy is more sensitive to the local residual. Hence, this explains that with adjoint weighted residual, better result can be generated with less mesh grids, compared with the second result in which only local residual is used.

5.3 Transonic Flow Around a NACA 0012 Airfoil

The last numerical test is for the transonic flow through a NACA0012 airfoil. The purpose is to show the advantage of adjoint weighted method on accurately calculating the quantity of interest in the practical applications such as drag coefficient in this test, i.e.

$$\mathscr{J}(U) = \int_{\partial\Omega_a} p\beta \cdot n ds, \tag{27}$$

where $\partial\Omega_a$ is the surface of the airfoil, and n is the unit outer normal vector with respect to $\partial\Omega_a$. The parameter β in the above formula is given as

$$\beta = \begin{cases} (\cos\alpha, \sin\alpha)^T / C_\infty, & \text{for drag calculation,} \\ \\ (-\sin\alpha, \cos\alpha)^T / C_\infty, & \text{for lift calculation,} \end{cases}$$

where $C_\infty = 0.5\gamma p_\infty Ma_\infty^2 l$, and Ma_∞ and l are the far field Mach number of the flow and the chord length of the airfoil, respectively.

The far field flow is set up with the following configuration. The density is 1, the Mach number is 0.8, and the velocity vector is $(\cos\theta, \sin\theta)^T$ with the attack angle $\theta = 1.25°$. The far field flow state is again used as the initial guess for the Newton iteration.

In Fig. 6 (left), the convergence history of Newton iteration on 11 successively and adaptively refined meshes is shown and it can be observed that the residual can be reduced towards the machine epsilon efficiently in all meshes which demonstrates the effectiveness of the algorithm. In Fig. 6 (right), the advantage on using adaptive

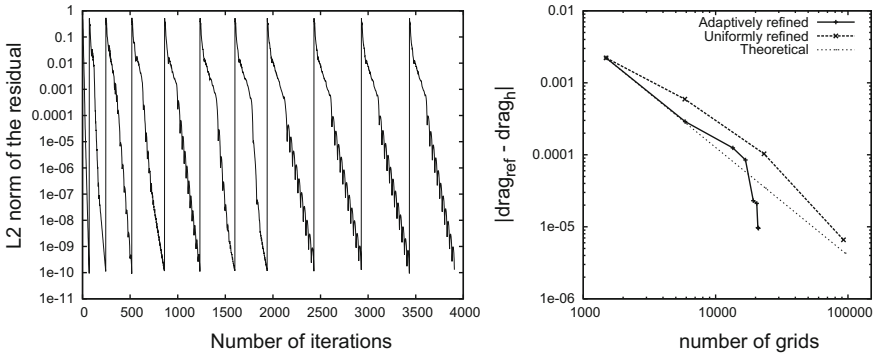


Fig. 6 Left: Residual convergence history with adaptively and successively mesh refinements for NACA0012 airfoil with 0.8 Mach number and 1.25° attack angle; Right: the corresponding convergence history of the drag coefficient (solid line), while the dashed line shows the results given by the uniformly refining mesh

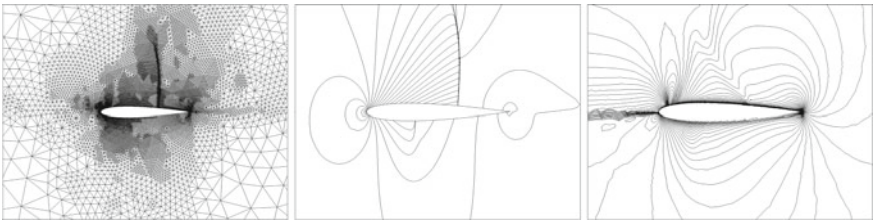


Fig. 7 Left: The mesh profile after 5 adaptive refinement. Middle: The corresponding isolines of the Mach number. Right: The isolines of the x-momentum from the adjoint problem

method with error indicator generated by adjoint weighted residual is demonstrated obviously, i.e. the convergence curve of the drag coefficient generated by the adaptive method is much superior to that generated by uniformly refining the mesh and to reach almost the same numerical accuracy (around $1.0e-05$), only over 10% mesh grids are needed by the adaptive mesh method, compared with the uniform refinement strategy. Figure 7 shows the mesh grids around the airfoil (left), the isolines of the Mach number (middle), and the isolines of x-momentum from the adjoint problem (right). It can be seen that with the adjoint weighted residual, the upper and lower shocks as well as leading edge and tail region are successfully resolved, which guarantees the accurate calculation of drag coefficient.

Remark 4 It is worth mentioning that in all simulations in this paper and our previous works [21, 23–26], the convergence of Newton iteration is smooth and efficient. Furthermore, the convergence is not sensitive to the selection of the parameters, which shows the robustness of our method.

Remark 5 In simulations with curved boundary, the direction of the out normal vector on the Gauss quadrature point is adjusted according to the exact curve. With this

correction, the performance of the method with high-order solution reconstruction can be significantly improved and people may refer to the simulation on Ringleb problem in [26] for details. However, there are still errors on the other quadrature information such as the position and weight of the quadrature point. Moreover, to develop a framework for the optimal design of the vehicle, a flexible and powerful tool to handle the curved boundary approximation is desirable. In our forthcoming paper, the nonuniform rational B-splines (NURBS) will be introduced in our method to handle the curved boundary issue and preliminary results show the excellent performance of the new method.

6 Conclusion

In this paper, an efficient and robust framework of adaptive finite volume solutions on steady Euler equations is introduced. The governing equations are discretized with finite volume method, and the framework consists of the Newton iteration for the linearization of the Euler system and a geometrical multigrid method for solving the linearized system. A non-oscillatory k-exact reconstruction is developed to deliver quality solution reconstruction to linear and higher-order cases. To improve the solver efficiency, the h-adaptive method is introduced in the method and an adjoint-based a posteriori error estimation method is developed to generate quality error indicator for the adaptive method. Numerical results successfully show the desired convergence behaviour of the method, and quality nonuniform meshes generated by the adaptive method.

Acknowledgements The research of Guanghui Hu was partially supported by 050/2014/A1 from FDCT of the Macao S. A. R., MYRG2014-00109-FST and MRG/016/HGH/2013/FST from University of Macau and National Natural Science Foundation of China (Grant No. 11401608). The research of Tao Tang was partially supported by the Special Project on High-Performance Computing of the National Key R&D Program under No. 2016YFB0200604, the National Natural Science Foundation of China under No. 11731006, and the Science Challenge Project under No. TZ2018001.

References

1. OpenMP. http://www.openmp.org
2. OpenMPI. http://www.open-mpi.org
3. T.J. Barth, Recent developments in high order k-exact reconstruction on unstructured meshes, in *31st Aerospace Sciences Meeting American Institute of Aeronautics and Astronautics*, pp. 1–15 (1993)
4. T.J. Barth, D.C. Jespersen, The design and application of upwind schemes on unstructured meshes, in *AIAA Paper*, pp. 89–0366 (1989)
5. M.J. Berger, A. Jameson, Automatic adaptive grid refinement for the Euler equations. AIAA J. **23**(4), 561–568 (1085)
6. J. Blazek, *Computational Fluid Dynamics: Principles and Applications*. (Elsevier Science, 2005)

7. L. Chen, G. Hu, R. Li, Integrated linear reconstruction for finite volume scheme on arbitrary unstructured grids, in *ArXiv e-prints*, March 2017. Commum. Comput. Phys. **24**(2), 454–480 (2018)
8. L. Chen, R. Li, An integrated linear reconstruction for finite volume scheme on unstructured grids. J. Sci. Comput. **68**(3), 1172–1197 (2016)
9. R.F. Chen, Z.J. Wang, Fast, block lower-upper symmetric Gauss-Seidel scheme for arbitrary grids. AIAA J. **38**(12), 2238–2245 (2000)
10. B. Cockburn, G.E. Kamiadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods: Theory, Computation and Applications*, Lecture Notes in Computational Science and Engineering, vol. 11 (Springer-Verlag, Heidelberg, 2000)
11. M. Delanaye, Polynomial reconstruction finite volume schemes for the compressible Euler and Navier-Stokes equations on unstructured adaptive grids. Ph.D. thesis, The University of Liege, Belgium, 1996
12. B. Engquist, B.D. Froese, Y.-H.R. Tsai, Fast sweeping methods for hyperbolic systems of conservation laws at steady state. J. Comput. Phys. **255**, 316–338 (2013)
13. B. Engquist, B.D. Froese, Y.-H.R. Tsai, Fast sweeping methods for hyperbolic systems of conservation laws at steady state II. J. Comput. Phys. **286**, 70–86 (2015)
14. D.J. Fidkowski, D.L. Darmofal, Review of output-based error estimation and mesh adaptation in computational fluid dynamics. AIAA J. **49**(4), 673–694 (2011)
15. M.B. Giles, N.A. Pierce, An introduction to the adjoint approach to design. Flow Turbul. Combust. **65**(3), 393–415 (2000)
16. M. Gurris, D. Kuzmin, S. Turek, Implicit finite element schemes for the stationary compressible Euler equations. Int. J. Numer. Meth. Fluids **69**(1), 1–28 (2012)
17. F. Haider, P. Brenner, B. Courbet, J.-P. Croisille, *Parallel Implementation of k-Exact Finite Volume Reconstruction on Unstructured Grids* (Springer International Publishing, Cham, 2014), pp. 59–75
18. R.E. Harris, Z.J. Wang, High-order adaptive quadrature-free spectral volume method on unstructured grids. Comput. Fluids **38**(10), 2006–2025 (2009)
19. P. Houston, E. Süli, hp-adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems. SIAM J. Sci. Comput. **23**(4), 1226–1252 (2001)
20. C.Q. Hu, C.-W. Shu, Weighted essentially non-oscillatory schemes on triangular meshes. J. Comput. Phys. **150**, 97–127 (1999)
21. G.H. Hu, An adaptive finite volume method for 2D steady Euler equations with WENO reconstruction. J. Comput. Phys. **252**, 591–605 (2013)
22. G.H. Hu, A numerical study of 2D detonation waves with adaptive finite volume methods on unstructured grids. J. Comput. Phys. **331**, 297–311 (2017)
23. G.H. Hu, R. Li, T. Tang, A robust high-order residual distribution type scheme for steady Euler equations on unstructured grids. J. Comput. Phys. **229**(5), 1681–1697 (2010)
24. G.H. Hu, R. Li, T. Tang, A robust WENO type finite volume solver for steady Euler equations on unstructured grids. Commun. Comput. Phys. **9**(3):627–648, 003 (2011)
25. G.H. Hu, X.C. Meng, N.Y. Yi, Adjoint-based an adaptive finite volume method for steady Euler equations with non-oscillatory k-exact reconstruction. Comput. Fluids **139**:174–183 (2016). *13th USNCCM International Symposium of High-Order Methods for Computational Fluid Dynamics–A special issue dedicated to the 60th birthday of Professor David Kopriva*
26. G.H. Hu, N.Y. Yi, An adaptive finite volume solver for steady Euler equations with non-oscillatory k-exact reconstruction. J. Comput. Phys. **312**, 235–251 (2016)
27. W.Z. Huang, R.D. Russell, *Adaptive Moving Mesh Methods*, vol. 174 of *Applied Mathematical Sciences* (Springer-Verlag New York, 2011)
28. M.Y. Hussaini, D.A. Kopriva, M.D. Salas, T.A. Zang, Spectral methods for the Euler equations. I – Fourier methods and shock capturing. AIAA J. **23**(1), 64–70 (1985)
29. A. Jameson, W. Schmidt, E. Turkel, Numerical solution of the Euler equations by finite volume methods using Runge-Kutta time stepping schemes, in *14th Fluid and Plasma Dynamics Conference* (1981)

30. G.-S. Jiang, C.-W. Shu, Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**(1), 202–228 (1996)
31. N. Kwatra, J.T. Grétarsson, R. Fedkiw, Practical animation of compressible flow for shock waves and related phenomena, in *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '10, pp. 207–215, Aire-la-Ville, Switzerland, Switzerland (2010). Eurographics Association
32. M.H. Lallemand, H. Steve, A. Dervieux, Unstructured multigridding by volume agglomeration: current status. Comput. Fluids **21**(3), 397–433 (1992)
33. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics (Cambridge University Press, 2002)
34. K. Li, Z. Xiao, Y. Wang, J.Y. Du, K.Q. Li (eds.), *Parallel Computational Fluid Dynamics*, vol. 405, Communications in Computer and Information Science (Springer, Berlin Heidelberg, 2013)
35. R. Li, On multi-mesh h-adaptive methods. J. Sci. Comput. **24**(3), 321–341 (2005)
36. R. Li, T. Tang, Moving mesh discontinuous Galerkin method for hyperbolic conservation laws. J. Sci. Comput. **27**(1), 347–363 (2006)
37. R. Li, T. Tang, P.W. Zhang, Moving mesh methods in multiple dimensions based on harmonic maps. J. Comput. Phys. **170**(2), 562–588 (2001)
38. R. Li, T. Tang, P.W. Zhang, A moving mesh finite element algorithm for singular problems in two and three space dimensions. J. Comput. Phys. **177**(2), 365–393 (2002)
39. R. Li, X. Wang, W.B. Zhao, A multigrid block LU-SGS algorithm for Euler equations on unstructured grids. Numer. Math. Theor. Methods Appl. **1**, 92–112 (2008)
40. H. Luo, J.D. Baum, R. Lhner, A Hermite WENO-based limiter for discontinuous Galerkin method on unstructured grids. J. Comput. Phys. **225**(1), 686–713 (2007)
41. K. Michalak, C. Ollivier-Gooch, Limiters for unstructured higher-order accurate solutions of the Euler equations, in *46th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada* (2008)
42. C. Ollivier-Gooch, A. Nejat, K. Michalak, Obtaining and verifying high-order unstructured finite volume solutions to the Euler equations. AIAA J. **47**(9), 2105–2120 (2009)
43. S. Popinet, Gerris: a tree-based adaptive solver for the incompressible Euler equations in complex geometries. J. Comput. Phys. **190**(2), 572–600 (2003)
44. J.X. Qiu, C.-W. Shu, Hermite WENO schemes and their application as limiters for Runge-Kutta discontinuous Galerkin method: one-dimensional case. J. Comput. Phys. **193**(1), 115–135 (2004)
45. J.X. Qiu, J. Zhu, *RKDG with WENO Type Limiters* (Springer, Heidelberg, 2010), pp. 67–80
46. H.Z. Tang, T. Tang, Adaptive mesh methods for one- and two-dimensional hyperbolic conservation laws. SIAM J. Numer. Anal. **41**(2), 487–515 (2003)
47. T. Tang, Moving mesh methods for computational fluid dynamics. Contemporary Math. **383**, 141–174 (2005)
48. E.F. Toro, M. Spruce, W. Speares, Restoration of the contact surface in the HLL-Riemann solver. Shock Waves **4**(1), 25–34 (1994)
49. V. Venkatakrishnan, Convergence to steady-state solutions of the Euler equations on unstructured grids with limiters. J. Comput. Phys. **118**, 120–130 (1995)
50. L. Wang, D.J. Mavriplis, Adjoint-based h-p adaptive discontinuous Galerkin methods for the compressible Euler equations, in *47th AIAA Aerospace Sciences Meeting including The New Horizons Forum and Aerospace Exposition, Orlando, Florida* (2009)
51. Z.J. Wang, Evaluation of high-order spectral volume method for benchmark computational aeroacoustic problems. AIAA J. **43**(2), 337–348 (2005)
52. Z.J. Wang, K. Fidkowski, R. Abgrall, F. Bassi, D. Caraeni, A. Cary, H. Deconinck, R. Hartmann, K. Hillewaert, H.T. Huynh, N. Kroll, G. May, P.-O. Persson, B. van Leer, M. Visbal, High-order CFD methods: current status and perspective. Int. J. Numer. Meth. Fluids **72**(8), 811–845 (2013)
53. Wikipedia. Geforce–Wikipedia, the free encyclopedia (2016). Accessed 16 Dec 2016

54. X.X. Zhang, C.-W. Shu, Positivity-preserving high order finite difference WENO schemes for compressible Euler equations. J. Comput. Phys. **231**(5), 2245–2258 (2012)
55. Y.-T. Zhang, C.-W. Shu, Chapter 5–ENO and WENO schemes, in *Handbook of Numerical Methods for Hyperbolic Problems Basic and Fundamental Issues, Handbook of Numerical Analysis*, ed. by R. Abgrall, C.-W. Shu, vol. 17 (Elsevier, 2016), pp. 103–122
56. J. Zhu, X.H. Zhong, C.-W. Shu, J.X. Qiu, Runge-Kutta discontinuous Galerkin method with a simple and compact Hermite WENO limiter. Commun. Comput. Phys. **19**(4):944–969, 004 (2016)

The Burgers–Hilbert Equation

John K. Hunter

Abstract The Burgers–Hilbert equation consists of an inviscid Burgers equation with a linear Hilbert-transform source term. We explain how the equation arises as a model for waves on a vorticity discontinuity and surface waves with constant frequency. We survey various results about the Burger–Hilbert equation, including ones on singularity formation, shock structure, weak solutions, and the enhanced life span of small, smooth solutions.

Keywords Conservation laws · Hilbert transform · Shock formation
Normal form transformations

1 The Burgers–Hilbert Equation

The Burgers–Hilbert (BH) equation consists of an inviscid Burgers equation for $u(x,t)$ with a linear source term given by the Hilbert transform of u with respect to x:

$$u_t + \left(\frac{1}{2}u^2\right)_x = \mathbf{H}[u]. \tag{1}$$

The Hilbert transform $\mathbf{H}$ is defined for $f : \mathbb{R} \to \mathbb{R}$ by

$$\mathbf{H}[f](x) = \text{p.v.}\frac{1}{\pi}\int_{\mathbb{R}} \frac{f(y)}{x-y}\,dy, \qquad \widehat{\mathbf{H}[f]}(k) = (-i\ \text{sgn}\ k)\,\hat{f}(k),$$

where $\text{sgn}\, k$ is the sign function. Analogous expressions apply to spatially periodic functions $f : \mathbb{T} \to \mathbb{T}$.

The Hilbert transform is a skew-adjoint singular integral operator of order zero, so the source term $\mathbf{H}[u]$ in (1) is conservative but non-smoothing. The BH equation

J. K. Hunter
University of California at Davis, Davis, CA 95616, USA
e-mail: jkhunter@ucdavis.edu

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_3

is Hamiltonian with respect to the Hamiltonian operator ∂_x,

$$u_t + \partial_x \left(\frac{\delta \mathcal{H}}{\delta u} \right) = 0, \qquad \mathcal{H}(u) = \int \left(\frac{1}{6} u^3 + \frac{1}{2} u \, |\partial_x|^{-1} u \right) dx,$$

where $|\partial_x| = \mathbf{H}\partial_x$ has symbol $|k|$.

Since $\mathbf{H}^2 = -\mathbf{I}$, all solutions of the linearized equation $u_t = \mathbf{H}[u]$ oscillate with frequency 1. Thus, the BH equation is a model equation for nonlinear waves with constant, nonzero linearized frequency. From a mathematical perspective, the equation combines aspects of conservation laws and singular integral operators.

In this paper, we survey a number of results about the BH equation. We explain how the BH equation is related to the motion of vorticity discontinuities in fluids and nonlinear waves with constant linearized frequency. We then describe results on the existence and breakdown of smooth solutions, the structure of shocks, and the global existence of weak solutions. Finally, we discuss the behavior of small-amplitude solutions, which are the ones that model the motion of vorticity discontinuities. In particular, we state a result on the enhanced life span of small, smooth solutions, which is obtained by the use of normal form methods, and give an asymptotic equation that describes the behavior of small solutions over cubically nonlinear timescales.

2 Vorticity Discontinuities

A planar discontinuity in vorticity in an inviscid, two-dimensional, incompressible, fluid flow with velocity $\mathbf{u}(x, y, t) = (u(x, y, t), v(x, y, t))$ corresponds to a shear flow $\mathbf{u} = (u, 0)$ with

$$u(y) = \begin{cases} \alpha_+ y & \text{if } y > 0, \\ \alpha_- y & \text{if } y < 0, \end{cases}$$

where $\alpha_+ \neq \alpha_-$ (see Fig. 1). The vorticity $v_x - u_y$ of this flow is $-\alpha_+$ in $y > 0$ and $-\alpha_-$ in $y < 0$ and jumps across $y = 0$. Since vorticity is advected by the fluid flow, it is consistent to consider the motion of a vorticity discontinuity located at $y = \eta(x, t)$ with constant vorticities $-\alpha_+$ in $y > \eta(x, t)$ and $-\alpha_-$ in $y < \eta(x, t)$.

A planar vorticity discontinuity may be regarded as an approximation of the flow in a wide, two-dimensional channel (see Fig. 2) or as a local approximation to a vortex patch (see Fig. 3). The boundary of a vortex patch remains smooth globally in time (Chemin [7]) but forms thin, high-curvature filaments (see e.g., Dritschel [9]).

Rayleigh [18] showed that a vorticity discontinuity is linearly stable, in sharp contrast to the catastrophic Kelvin–Helmholtz instability of a vortex sheet with a jump in the tangential velocity. Rayleigh also computed the response of a vorticity discontinuity to a sinusoidal perturbation up to fifth order in the amplitude of the perturbation and found no sign of nonlinear instability [19]. However, as we explain

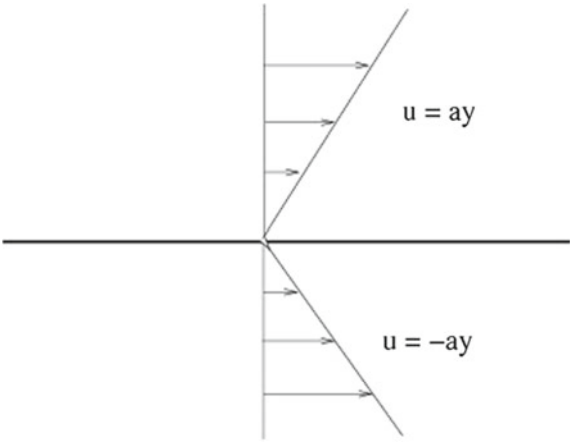


Fig. 1 A vorticity discontinuity with shear rates $\alpha_+ = a$ and $\alpha_- = -a$

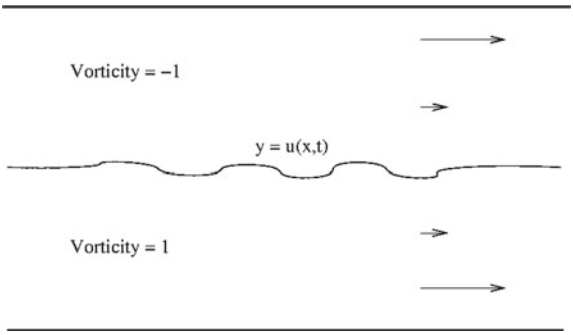


Fig. 2 A vorticity discontinuity in channel flow with $\alpha_\pm = \pm 1$. Figure 1 corresponds to the limit of infinite channel width

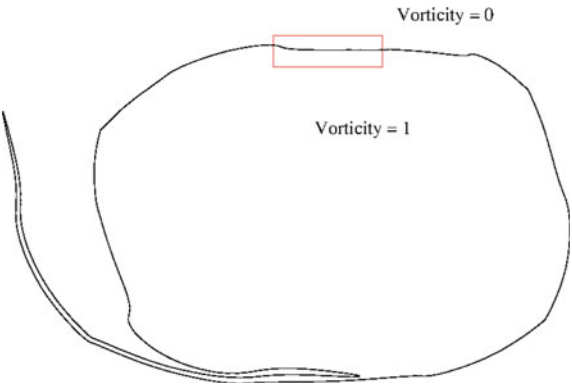


Fig. 3 A vortex patch with $\alpha_+ = 0$, $\alpha_- = 1$. The vorticity discontinuity in Fig. 1 corresponds to a local approximation of the boundary of the patch

in Sect. 7, the restriction of the perturbation to a single Fourier mode leads to a traveling wave and eliminates the most interesting weakly nonlinear dynamics of the discontinuity.

The motion of the fluid is described by the two-dimensional incompressible Euler equations for the velocity $\mathbf{u}(\mathbf{x}, t)$ and the pressure $p(\mathbf{x}, t)$:

$$\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p = 0, \qquad \nabla \cdot \mathbf{u} = 0.$$

The only parameters in an unperturbed planar vorticity discontinuity are the shear rates $\alpha_\pm$, which have the dimension of frequency; in particular, the planar geometry does not define any length scales. As a result, the Euler equations and the unperturbed flow are invariant under spatial dilations

$$\mathbf{x} \mapsto \lambda \mathbf{x}, \qquad t \mapsto t, \qquad \mathbf{u} \mapsto \lambda \mathbf{u}, \qquad p \mapsto \lambda^2 p$$

for $\lambda > 0$. The unperturbed flow is not invariant under spatial reflections $x \mapsto -x$, $t \mapsto t$. Dimensional arguments then imply that the frequency of a small-amplitude wave on a vorticity discontinuity is independent of its wavenumber; in addition, waves propagate along the interface in only one direction (with the larger value of α_+, α_- on their left). This spatial scaling invariance may be compared with the space-time invariance $\mathbf{x} \mapsto \lambda \mathbf{x}$, $t \mapsto \lambda t$ of hyperbolic conservation laws, where the velocity of a small-amplitude wave is independent of its wavenumber.

Biello and Hunter [2] showed that the following BH equation provides an effective equation for small-amplitude motions of a vorticity discontinuity between shear rates α_+, α_- located at $y = \eta(x, t)$:

$$\eta_t + \left(\frac{1}{2}\beta_0 \eta^2\right)_x = \omega_0 \mathbf{H}[\eta], \qquad \beta_0^2 = \frac{\alpha_+^2 + \alpha_-^2}{2}, \quad \omega_0 = \frac{\alpha_+ - \alpha_-}{2}. \tag{2}$$

Here, ω_0 is the linearized frequency of oscillations in the vorticity discontinuity.

As explained in Sect. 7, the remarkable fact is that when the quadratically nonlinear coefficient in the BH equation is renormalized to β_0, then (at least formally) small-amplitude solutions of the BH equation have the same asymptotic behavior on cubically nonlinear timescales as solutions of the incompressible Euler equations for the vorticity discontinuity.

From the point of view of normal forms, when a suitable near-identity transformation is used to eliminate the quadratic term from (2), the resulting cubic term is exactly the one that describes the motion of a vorticity discontinuity. As shown in [2], dimensional analysis provides a partial explanation of why this occurs by showing that the BH equation is an appropriate equation to describe Hamiltonian surface waves with constant frequency.

A BH equation was also obtained by Marsden and Weinstein [16] as a quadratic truncation of the equation for vortex patches. In the case of a vorticity discontinuity, their equation is

$$\eta_t + \left(\frac{1}{2}\gamma_0 \eta^2\right)_x = \omega_0 \mathbf{H}[\eta], \qquad \gamma_0 = \frac{\alpha_+ + \alpha_-}{2}.$$

This equation has a different nonlinear coefficient from (2), which may vanish, and it does not provide a valid approximation for the motion of a vorticity discontinuity on cubic timescales.

The linearized dispersion relation for waves on the vorticity discontinuity in the channel flow shown in Fig. 2 is $\omega = \omega_0 \tanh hk$, where $2h$ is the width of the channel [18]. Thus, a heuristic model equation in that case is

$$\eta_t + \left(\frac{1}{2}\beta_0 \eta^2\right)_x = \omega_0 \mathbf{T}[\eta],$$

where the linear operator $\mathbf{T}$ has symbol $-i \tanh(hk)$. Equation (2) is the short-wave limit of this equation.

3 Smooth Solutions and Singularity Formation

We consider the initial value problem for the BH equation

$$u_t + \left(\frac{1}{2}u^2\right)_x = \mathbf{H}[u], \qquad u(x,0) = u_0(x). \tag{3}$$

This problem is well posed in the L^2-Sobolev space H^s for short times when $s > 3/2$, with the same proof as for the inviscid Burgers equation.

In Fig. 4, we show a numerical solution of (3) with sinusoidal initial data $u_0(x) = \sin x$, computed using a fourth-order WENO scheme to capture shocks. We see the typical Burgers steepening and the formation of a logarithmically cusped shock, whose structure is described in Sect. 4.

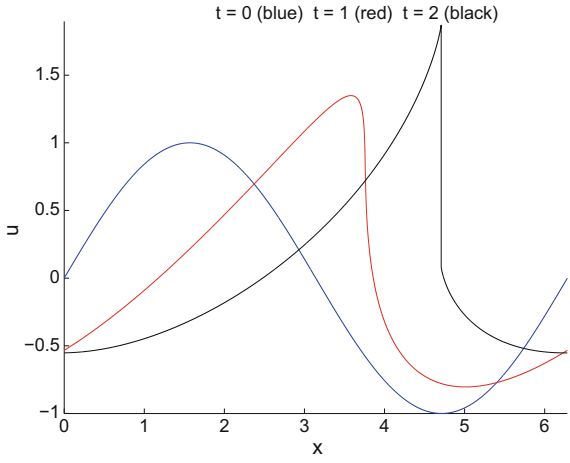


Fig. 4 Solution of (1) with initial data $u(x,0) = \sin x$ at $t = 0$ (blue), $t = 1$ (red), $t = 2$ (black)

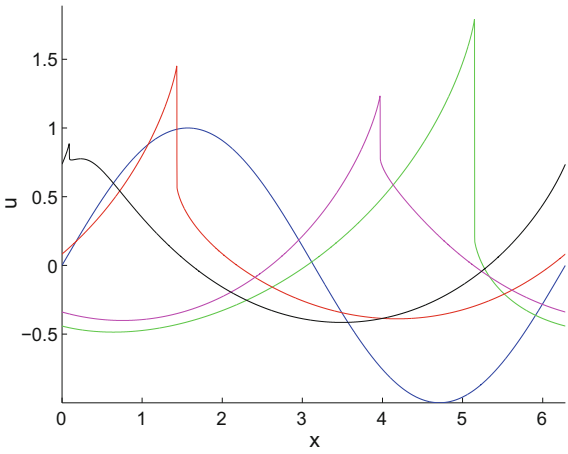


Fig. 5 Solution of (1) with initial data $u(x, 0) = \sin x$ at $t = 0$ (blue), $t = 2.5$ (green), $t = 5$ (red), $t = 7.5$ (magenta), $t = 10$ (black)

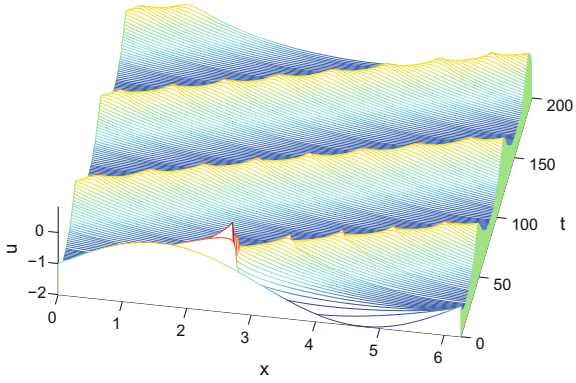


Fig. 6 Solution of (1) for $0 \le t \le 200$ with initial data $u(x, 0) = -1 + \sin x$. The Galilean transformation $u \mapsto -1 + u$, $x \mapsto x - t$ goes into a reference frame moving with the linearized wave speed

At longer times, the shock weakens due the effect of the lower-order oscillatory Hilbert-transform term, not viscous decay, and the solution appears to approach a smooth, temporally quasi-periodic traveling wave as $t \to \infty$ (Figs. 5 and 6). This behavior is not unique to the BH equation; it also occurs for Whitham equations of the form

$$u_t + \left(\frac{1}{2}u^2\right)_x = k * u \tag{4}$$

where $k \in L^1$ is an odd function. Shefter and Rosales [21] carried out a numerical study of (4) with the kernel $k(x) = \sin x$. Their results support the existence of a two-dimensional invariant torus of quasi-periodic solutions that is an attracting manifold for general solutions as a result of shock formation and decay.

Numerical simulations show that smooth initial data typically forms shocks. However, at least some small amplitude, spatially periodic initial data have global smooth solutions.

Theorem 1 *The BH Eq.* (1) *has an analytic branch*

$$a \mapsto (f(x,a), c(a))$$

in $H^k(\mathbb{T})$, $k \geq 1$, *of even, zero-mean,* 2π*-periodic, smooth traveling wave solutions* $u(x,t;a) = f(x - c(a)t, a)$. *Furthermore, as* $a \to 0$,

$$f(x,a) = a\cos x + \frac{1}{2}a^2 \cos 2x + \frac{3}{8}a^3 \cos 3x + O(a^4),$$
$$c(a) = 1 + \frac{1}{4}a^2 + O(a^4).$$

The proof is a standard application of the Crandall–Rabinowitz theorem on bifurcation from simple eigenvalue [22]. A similar result would apply to traveling waves on a vorticity discontinuity, and the expansion in Theorem 1 agrees with Rayleigh's expansion for traveling waves on vorticity discontinuities [19]. For the BH equation, this branch of traveling waves presumably ends at a steepest wave which is not smooth at its crest, as typically occurs for traveling wave solutions of Whitham-type Eq. (4); see e.g., [11, 17].

The existence of solitary wave solutions of the BH equation on $\mathbb{R}$ is unclear. The BH equation is invariant under the scaling $x \mapsto \lambda^{-1}x$, $u \mapsto \lambda u$, so in the long-wave limit $\lambda \to 0$, the profile of the spatially periodic traveling waves remains the same and their amplitude goes to infinity. Any solitary wave would need to have zero integral, otherwise $\mathbf{H}[u](x) \sim \frac{1}{\pi x}\int_{\mathbb{R}} u(y)\,dy$ as $|x| \to \infty$ would be non-integrable.

In the opposite direction, we have the following result of Castro, Córdoba, and Gancedo [6] on the breakdown of smooth solutions.

Theorem 2 *Let* $u_0 \in L^2(\mathbb{R}) \cap C^{1,\delta}(\mathbb{R})$ *with* $0 < \delta < 1$ *and suppose that there exists* $x_0 \in \mathbb{R}$ *such that*

$$\mathbf{H}[u_0](x_0) > 0, \qquad u_0(x_0) \geq \left(32\pi \|u_0\|_{L^2}^2\right)^{1/3}.$$

Then there is a finite time $T > 0$ *such that the solution* $u : \mathbb{R} \times [0,T)$ *of* (3) *satisfies*

$$\|u(\cdot,t)\|_{C^{1,\delta}} \to \infty \quad \text{as } t \to T.$$

The idea of the proof is to introduce characteristic coordinates $x = X(\xi,t)$, $u = U(\xi,t)$, where $X_t(\xi,t) = U(\xi,t)$ with $X(\xi,0) = \xi$ and $U(\xi,0) = u_0(\xi)$, in which case

$$U_{tt}(\xi,t) + U(\xi,t) = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{[u(X(\xi,t),t) - u(y,t)]^2}{[X(\xi,t) - y]^2}\,dy.$$

Under the stated assumptions, we have the pointwise bound

$$\frac{1}{2\pi}\int_{\mathbb{R}}\frac{[u(x,t)-u(y,t)]^2}{(x-y)^2}\,dy \geq Cu^4(x), \qquad C=\frac{1}{32\pi\|u\|_{L^2}^2}.$$

Hence, if $J(t)=u(X(\xi_0,t),t)$, where $X(\xi_0,0)=x_0$, then $J_{tt}+J\geq CJ^4$, so $J(t)$ blows up (and smooth solutions must break down at that time or earlier) if $J_t(0)>0$ and $0<J(0)\leq CJ^4(0)$.

Although this result shows that smooth solutions of the BH equation break down, it does not show that u_x blows up, as observed in the numerical solutions, nor does give a sharp estimate for the smooth life span.[1] However, it is unclear how to use a standard argument for the blow up of $v=u_x$, based on the Ricatti equation $dv/dt+v^2=\mathbf{H}[v]$, since one lacks uniform pointwise bounds for the Hilbert transform.

4 Shocks

A significant difficulty in understanding weak solutions of the BH equation with shocks is that the Hilbert transform is not bounded on L^∞ or L^1, and it generates a logarithmic singularity on a jump discontinuity:

$$\mathbf{H}[\operatorname{sgn} x]=\frac{2}{\pi}\log|x|.$$

Thus, the source term $\mathbf{H}[u]$ in the BH equation is singular along the trajectory of a shock. The speed of a shock differs from the characteristic speeds on either side, and Bressan and Zhang [4] showed that the singularity generated by the Hilbert transform can be balanced by the nonlinear advection term.

Let

$$\phi(x)=\frac{2|x|\log|x|}{\pi}\eta(x), \qquad \eta(x)=\begin{cases}1 & \text{if } |x|\leq 1,\\ 0 & \text{if } |x|\geq 2,\end{cases}$$

where $\eta\in C_c^\infty(\mathbb{R})$ is a suitable smooth cutoff function. Define spaces

$$\mathcal{D}=\left\{v\in H^2(\mathbb{R}\setminus\{x_0\}) : \text{for some } x_0\in\mathbb{R} \text{ with } v(x_0^-)>v(x_0^+)\right\},$$
$$\mathcal{E}=\{w : w(x)=\phi(x-x_0)+v(x) \text{ for } v\in\mathcal{D} \text{ with jump at } x_0\}.$$

Functions in $\mathcal{D}$ are piecewise smooth with a compressive jump discontinuity at x_0; functions in $\mathcal{E}$ are continuous with a logarithmic cusp at x_0. Then, we have the following result [4].

[1] See [1] for a similar example involving an application of functional methods to the inviscid Burger's equation.

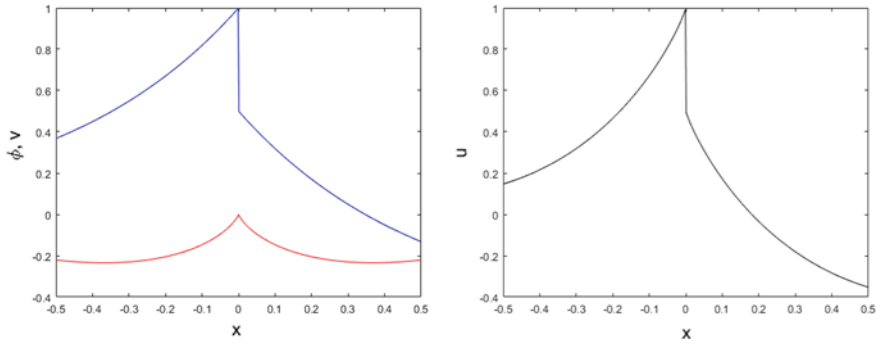


Fig. 7 Shock structure for the BH equation. The solution is the sum of a piecewise smooth function with a jump discontinuity (in blue) and a continuous function with a logarithmic cusp (in red)

Theorem 3 *For every $u_0 \in \mathcal{D}$, there exists $T > 0$ and a unique weak entropy solution $u(x,t)$ of (3) with $u(\cdot,t) \in \mathcal{E}$ for $0 < t \le T$.*

The structure of shocks in $\mathcal{E}$ is illustrated in Fig. 7. A surprising feature of these solutions is that the strength of the logarithmic cusp is independent of the shock strength.

Theorem 3 only holds for short times, since further shocks may form. Bressan and Nguyen [5] proved the following global existence result for general weak solutions.

Theorem 4 *If $u_0 \in L^2(\mathbb{R})$, then there is a weak entropy solution $u(x,t)$ of (3) that is defined for all $(x,t) \in \mathbb{R} \times [0,\infty)$. For this solution, the function $t \mapsto \|u(\cdot,t)\|_{L^2(\mathbb{R})}$ is non-increasing and $u(\cdot,t) \in L^\infty(\mathbb{R})$ for all $t > 0$.*

The proof uses a fractional step method between the flows defined by

$$u_t + \left(\frac{1}{2}u^2\right)_x = 0, \qquad u_t = \mathbf{H}[u].$$

The difficulty is that Burgers equation defines a contractive nonlinear semigroup on L^1 but not on L^2, whereas the Hilbert transform is bounded on L^2 (and L^p for $1 < p < \infty$) but not on L^1. The regularizing properties of the Burgers flow, and the Oleinik inequality, are essential to prove compactness of the fractional step approximations. The proof in [5] does not show that the solution belongs to BV, and the uniqueness of weak entropy solutions is largely open.

5 Small-Amplitude BH Dynamics

If $u_x \ll 1$, then the nonlinear Burgers term in (1) is small compared with the linear Hilbert transform, and (1) can be regarded as a perturbation of the linearized BH

equation

$$u_t = \mathbf{H}[u]. \tag{5}$$

The dynamics of the BH equation in this small-amplitude regime differs from the usual Burgers dynamics.

The solution of (5) is given by

$$u(x,t) = u_0(x)\cos t + h_0(x)\sin t, \qquad h_0 = \mathbf{H}[u_0].$$

Thus, u oscillates in time between u_0, h_0, $-u_0$, and $-h_0$, where u_0 is an arbitrary spatial profile and h_0 is its Hilbert transform. The solution acts like a collection of simple harmonic oscillators at different locations, each of which has the same frequency; the oscillators are coupled only by the condition that their velocity is the spatial Hilbert transform of their displacement.

When a slow nonlinear Burgers dynamics is added to these rapid linear oscillations, the spatial waveform of the solution alternately compresses and expands in each oscillation period, and the quadratic Burgers nonlinearity averages over time to a cubic nonlinearity. The transition from quadratic to cubic nonlinearity is illustrated in Fig. 8, which shows the numerically computed singularity formation time T_s, at which it appears that $|u_x| \to \infty$, as a function of amplitude ϵ for (3) with the initial data

$$u_0(x) = \epsilon\left[2\cos x + \cos 2(x + 2\pi^2)\right]. \tag{6}$$

A numerical solution for $\epsilon = 0.025$ is shown in Fig. 9, and a detail of the shock formation is shown in Fig. 10. The solution steepens and expands during each oscillation period. Eventually, after 104 oscillations, a very weak shock forms for the first time; this shock then becomes too weak to detect; in the next period, a slightly stronger shock forms; and so on. For scalar conservation laws, shocks do not disappear once formed [8], and it seems plausible that a similar result holds for the

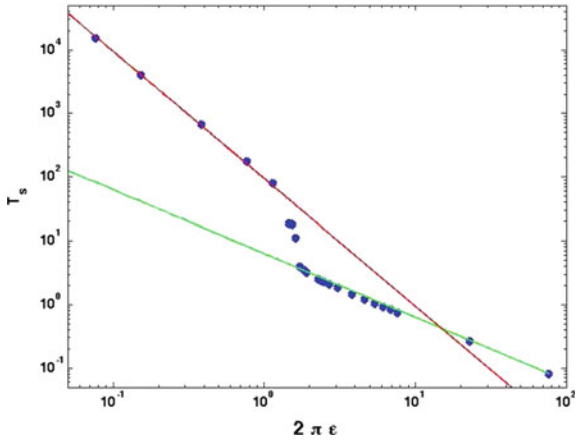


Fig. 8 Singularity time T_s versus amplitude ϵ for the initial data (6). Green line = quadratic Burgers equation asymptotics; Red line = cubic asymptotics; Blue dots = numerical solution. (From [2])

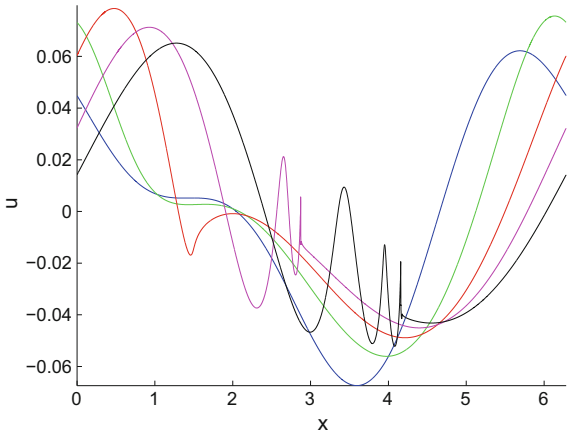


Fig. 9 Solution of (3) with initial data (6) where $\epsilon = 0.025$. The solution is plotted for $t = 2\pi N$ with $N = 0$ (blue), $N = 50$ (green), $N = 100$ (red), $N = 150$ (magenta), $N = 200$ (black). See Fig. 12 for a surface plot

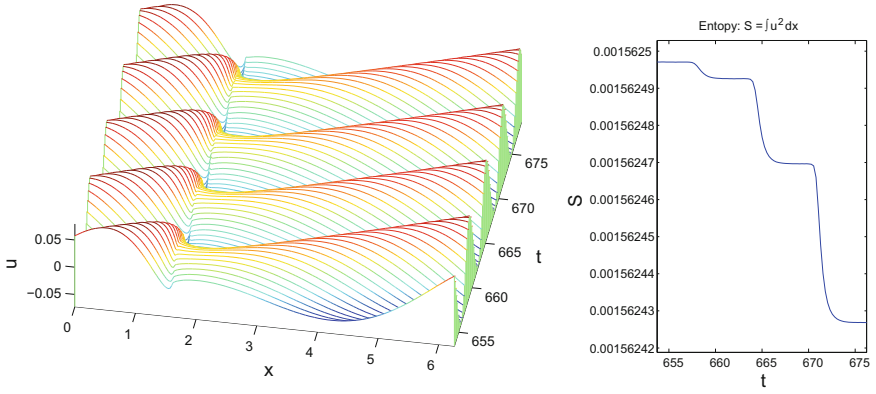


Fig. 10 Detail of shock formation. Left: Previous solution for $104 \cdot 2\pi \leq t \leq 108 \cdot 2\pi$ Right: the L^2-entropy drops by approximately 3×10^{-3} percent in first time period after shock forms

nonlocal BH equation. This argument suggests that, after the smooth solution breaks down, tiny new shocks form in each oscillation period, then become extremely weak as they absorb a large expansion wave generated by the background oscillation.

In the vorticity discontinuity problem, contour dynamics computations show that singularity formation in the BH equation is associated with wave-breaking and filamentatation of the vorticity discontinuity [3]. The discontinuity throws off new, extremely thin filaments in each oscillation period, and it is conceivable that solutions of the BH equation with very weak shocks provide a reasonable approximation to filamented vorticity discontinuities (as in numerical contour surgery methods that snip off the filaments [10]).

6 Enhanced Life Span of Smooth Solutions

Small, smooth solutions of the BH equation, with amplitude of the order ϵ, have an enhanced cubic life span of the order $1/\epsilon^2$, rather than the quadratic life span of the order $1/\epsilon$ for the inviscid Burgers equation. A proof of this result may be based on normal form methods.

Normal form methods for PDEs were introduced by Shatah [20], but standard methods fail for the BH equation. Thus, the BH equation provides a useful test problem for the development of normal form methods for quasi-linear PDEs; for example, normal form methods applicable to two-dimensional water waves in holomorphic coordinates bear a remarkable similarity to the ones for the BH equation [13].

It is convenient to transfer the small parameter ϵ from the initial data to the nonlinear term in the BH equation, and consider the initial value problem

$$u_t + \epsilon \left(\frac{1}{2}u^2\right)_x = \mathbf{H}[u], \qquad u(x, 0; \epsilon) = u_0(x). \tag{7}$$

A standard near-identity transformation $u \mapsto v$ that reduces the BH equation to a cubic normal form is given by

$$v = u + \frac{1}{2}\epsilon \left|\partial_x\right| \left(h^2\right), \qquad h = \mathbf{H}[u]. \tag{8}$$

However, since we use a lower-order linear term $\mathbf{H}[u]$ to remove a higher-order quadratic term $(u^2/2)_x$, there is a loss of derivatives in this transformation, and we cannot carry out cubic energy estimates for v. The following result is proved in [12, 14] by two related, but different, methods that overcome the loss of derivatives in (8).

Theorem 5 *Suppose that $u_0 \in H^2(\mathbb{R})$. There are constants $k > 0$ and $\epsilon_0 > 0$, depending only on $\|u_0\|_{H^2}$, such that for every ϵ with $|\epsilon| \leq \epsilon_0$ there exists a solution $u : I^\epsilon \to H^2(\mathbb{R})$ of (7) defined on the time-interval $I^\epsilon = \left[-k/\epsilon^2, k/\epsilon^2\right]$.*

In the first method, we regard the unbounded near-identity transformation as the forward Euler approximation of a bounded and invertible near-identity flow, which does not have a derivative loss. The BH equation is particularly simple because we can solve explicitly for the near-identity flow.

When written in terms of the Hilbert transforms $h = \mathbf{H}[u]$, $g = \mathbf{H}[v]$, the transformation $h \mapsto g$ in (8) is local,

$$g = h - \frac{1}{2}\epsilon \left(h^2\right)_x.$$

This transformation agrees to $O(\epsilon)$ with the solution of a Burgers equation for $U(x, t, \tau)$, in which t occurs as a parameter and we omit the dependence on ϵ to simplify the notation,

$$U_\tau + \left(\frac{1}{2}U^2\right)_x = 0, \qquad U(x,t,0) = h(x,t), \quad U(x,t,\epsilon) = g(x,t). \tag{9}$$

Solving this equation by the method of characteristics, we get the transformation $h \mapsto g$ where

$$g(x,t) = h(\xi,t), \qquad x = \xi + \epsilon h(\xi,t) \tag{10}$$

If u satisfies the BH equation in (7), then one finds that the function g given in (10) satisfies the following cubically nonlinear integro-differential evolution equation

$$g_t(x,t) - \frac{\epsilon^2}{\pi}\partial_x \int (x-y) g_y(y,t)\phi\left(\frac{g(x,t)-g(y,t)}{x-y}\right) dy = \mathbf{H}[g](x,t),$$
$$\phi(c) = \log(1-c) + c.$$

This equation has closed H^2-energy estimates, and Theorem 5 follows [12].

In the second method, we use a modified energy suggested by the near-identity transformation (8). The L^2-energy associated with $\partial_x^k v$ is

$$\|\partial_x^k v\|^2 = \|\partial_x^k u\|^2 + 2\epsilon\langle \partial_x^k u, \partial_x^k \mathbf{H}[hh_x]\rangle + \epsilon^2 \|\partial_x^k \mathbf{H}[hh_x]\|^2, \tag{11}$$

where $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ are the L^2-norm and inner product. The quartic term of $O(\epsilon^2)$ in (11) involves higher-order derivatives and is comparable with $\|\partial_x^{k+1}u\|^2$, but it is irrelevant for cubically nonlinear energy estimates. The cubic term of $O(\epsilon)$ in (11) is comparable with $\|\partial_x^k u\|^2$ because the higher-order derivatives integrate out:

$$\langle \partial_x^k u, \partial_x^k \mathbf{H}[hh_x]\rangle = \left(k + \frac{1}{2}\right)\langle h_x, (\partial_x^k h)^2\rangle + \text{l.o.t.}$$

The skew-adjointness behind this integration by parts is related to the well-posedness of the normal form flow in (9).

This observation suggests that we drop the higher-derivative quartic term from (11) and define a modified energy by

$$E_k(u) = \frac{1}{2}\|\partial_x^k u\|^2 + \langle \partial_x^k u, \partial_x^k \mathbf{H}[hh_x]\rangle.$$

For small energies and $k \geq 2$, the modified energy $E_k(u)$ is equivalent to $\|\partial_x^k u\|^2$, and one computes directly from (7) that it has cubically nonlinear estimates, so Theorem 5 follows [14].

7 Asymptotic Equation for the BH Equation

The linearized dispersion relation of the BH equation for non-constant harmonic solutions $u(x,t) = e^{ikx-i\omega t}$ is $\omega = \operatorname{sgn} k$. Weakly nonlinear solutions are not subject to quadratic, three-wave resonances since there are no solutions of $\omega_1 = \omega_2 + \omega_3$; this non-resonance condition is what allows us to remove the quadratic term from the BH equation by a near-identity transformation. There are, however, many cubic, four-wave resonances of the form

$$k_1 + k_2 = k_3 + k_4, \qquad k_j > 0.$$

As a result, the nonlinear dynamics of small-amplitude, constant-frequency waves is qualitatively different from that of dispersive waves, which have few four-wave resonances, and two initial spatial harmonics are sufficient to start an energy cascade e.g., from $k = 1, 2$ one gets wavenumbers $3 = 2 + 2 - 1, 4 = 2 + 3 - 1 = 3 + 3 - 2$, and so on.

Thus, small-amplitude, constant-frequency waves have an arbitrary spatial profile that oscillates at the linearized frequency and deforms slowly due to the effects of nonlinearity. Their nonlinearity is cubic, like dispersive waves and the cubic NLS, but four-wave interactions generate an energy cascade to higher spatial wavenumbers, like the three-wave energy cascade for non-dispersive hyperbolic waves and the inviscid Burgers equation. No such four-wave cascade occurs if there is only one initial harmonic, which explains why it did not arise in Rayleigh's analysis of vorticity discontinuities [19].

The BH Eq. (1) has weakly nonlinear, asymptotic solutions of the form [2]

$$u(x,t;\epsilon) = \epsilon v(x,\epsilon^2 t)\cos t + \epsilon \mathbf{H}[v](x,\epsilon^2 t)\sin t + O(\epsilon^2) \qquad \text{as } \epsilon \to 0,$$

where $v(x,\tau)$ satisfies the cubically nonlinear, nonlocal asymptotic equation

$$v_\tau + \frac{1}{2}\partial_x\left\{v^2\,|\partial_x|\,v - v\,|\partial_x|\,v^2 + \frac{1}{3}\,|\partial_x|\,v^3\right\} = 0, \qquad |\partial_x| = \mathbf{H}\partial_x. \tag{12}$$

Exactly the same asymptotic equation, with a suitable nonlinear coefficient, arises from the incompressible Euler equations for a vorticity discontinuity [2]. Consequently, the BH equation with the renormalized coefficient in (2) provides an effective equation for small-slope motions of a vorticity discontinuity on cubically nonlinear timescales.

The following local existence and uniqueness result for spatially periodic solutions of (12) is proved in [15].

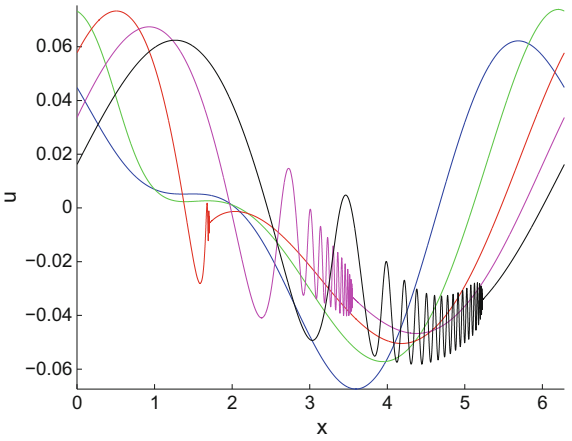


Fig. 11 Solution of (13) with initial data (6) where $\epsilon = 0.025$. The solution is plotted for $t = 2\pi N$ with $N = 0$ (blue), $N = 50$ (green), $N = 100$ (red), $N = 150$ (magenta), $N = 200$ (black)

Theorem 6 *Suppose that $u_0 \in H^2(\mathbb{T})$. Then there exists $T(\|u_0\|_{H^2}) > 0$ such that the initial value problem*

$$\begin{aligned} &u_t + \frac{1}{2}\partial_x \left\{ u^2 \left|\partial_x\right| u - u \left|\partial_x\right| u^2 + \frac{1}{3} \left|\partial_x\right| u^3 \right\} = 0, \\ &u(x,0) = u_0(x), \end{aligned} \tag{13}$$

has a unique solution $u \in C(-T, T; H^2) \cap C^1(-T, T; H^1)$.

Although the nonlinear term in (13) involves two spatial derivatives, it is effectively first order on smooth solutions because of the commutator identity

$$u^2 \left|\partial_x\right| u - u \left|\partial_x\right| u^2 + \frac{1}{3} \left|\partial_x\right| u^3 = 2u \left[u, \mathbf{H}\right] u_x - \left[u^2, \mathbf{H}\right] u_x.$$

This cancelation of derivatives (or its spectral equivalent) enables one to close the energy estimates.

Equation (13) inherits a Hamiltonian structure from the BH equation

$$u_t + \partial_x \left(\frac{\delta \mathcal{H}}{\delta u} \right) = 0, \qquad \mathcal{H}(u) = \int \left(\frac{1}{6} u \left|\partial_x\right| u^3 - \frac{1}{4} u^2 \left|\partial_x\right| u^2 \right) dx.$$

A numerical solution of (13) with initial data (6) is shown in Fig. 11. There is numerical evidence of singularity formation and the existence of a weak solution, but no proof. The solution of the asymptotic equation appears to remain continuous, but forms a propagating, oscillatory singularity. This result is consistent with the

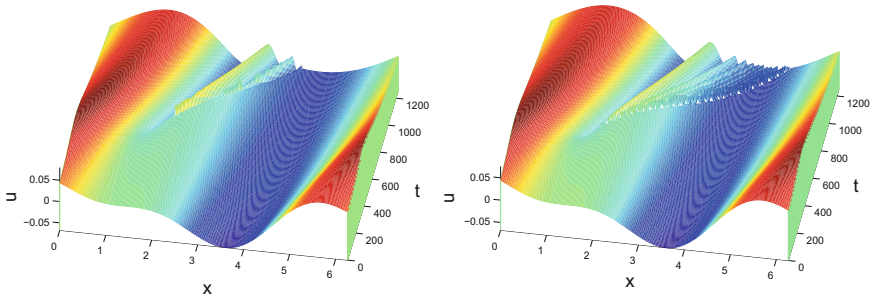


Fig. 12 Left: Solution of BH equation (1) with initial data (6) where $\epsilon = 0.025$. $\epsilon = 0.025$ plotted at 2π-time intervals. Right: Corresponding solution obtained from the asymptotic equation

formation of multiple shocks (or filaments) whose strength (or thickness) approaches zero in the weakly nonlinear limit.

Finally, in Fig. 12, we compare the BH and asymptotic solutions. The asymptotic equation provides a good approximation to the BH equation.

Acknowledgements Supported by the NSF under grant number 1616988.

References

1. S. Alinhac, *Blowup for Nonlinear Hyperbolic Equations* (Birkhäuser, Boston, 1995)
2. J. Biello, Hunter J.K: Nonlinear Hamiltonian waves with constant frequency and surface waves on vorticity discontinuities. Comm. Pure Appl. Math. **63**, 303–336 (2009)
3. J. Biello, J.K. Hunter, unpublished
4. A. Bressan, T. Zhang, Piecewise smooth solutions to the Burgers-Hilbert equations. Comm. Math. Sci. **15**, 165–184 (2017)
5. A. Bressan, K.T. Nguyen, Global existence of weak solutions for the Burgers-Hilbert equation. SIAM J. Math. Anal. **46**, 2884–2904 (2014)
6. A. Castro, D. Córdoba, F. Gancedo, Singularity formation for a surface wave model. Nonlinearity **23**, 2835–2847 (2010)
7. J.Y. Chemin, Persistance de structures gomtriques dans les fluides incompressibles bidimensionnels. Ann. Sci. École Norm. Sup. **26**, 517–542 (1993)
8. C.M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, 4th edn. (Springer-Verlag, Heidelberg, 2016)
9. D.G. Dritschel, The repeated filamentation of two-dimensional vorticity interfaces. J. Fluid Mech. **194**, 511–547 (1988)
10. D.G. Dritschel, Contour dynamics and contour surgery. Comput. Phys. Rep. **10**, 77–146 (1989)
11. J.K. Hunter, Numerical solution of some nonlinear dispersive wave equations. Lect. Appl. Math. **26**, 301–316 (1990)
12. M. Ifrim, J.K. Hunter, Enhanced life span of smooth solutions of a Burgers-Hilbert equation. SIAM J. Math. Anal. **44**, 2039–2052 (2012)
13. J.K. Hunter, M. Ifrim, D. Tataru, Two dimensional water waves in holomorphic coordinates. Comm. Math. Phys. **346**, 483–552 (2016)

14. J.K. Hunter, M. Ifrim, D. Tataru, T.K. Wong, Long time solutions for a Burgers-Hilbert equation via a modified energy method. Proc. Am. Math. Soc. **143**, 3407–3412 (2015)
15. M. Ifrim, Normal form transformations for Quasilinear wave equations. Thesis (Ph.D.), University of California, Davis, 2012
16. J. Marsden, A. Weinstein, Coadjoint orbits, vortices, and Clebsch variables for incompressible fluids. Phys. D **7**, 305–323 (1983)
17. R.L. Pego, Some explicit resonating waves in weakly nonlinear gas dynamics. Stud. Appl. Math. **79**, 263–270 (1988)
18. L. Rayleigh, On the stability or instability of certain fluid motions. Proc. Lond. Math. Soc. **11**, 57 (1880)
19. L. Rayleigh, On the propagation of waves upon the plane surface separating two portions of fluid of different vorticities. Proc. Lond. Math. Soc. **27**, 13–18 (1895)
20. J. Shatah, Normal forms and quadratic nonlinear Klein-Gordon equations. Comm. Pure Appl. Math. **38**, 685–696 (1985)
21. M. Shefter, R.R. Rosales, Quasiperiodic solutions in weakly nonlinear gas dynamics. I. Numerical results in the inviscid case. Stud. Appl. Math. **103**, 279–337 (1999)
22. E. Zeidler, *Nonlinear Functional Analysis and its Applications*, vol. I (Springer-Verlag, New York, 1986)

General Linear Methods for Time-Dependent PDEs

Alexander Jaust and Jochen Schütz

Abstract The hybridized discontinuous Galerkin method has been successfully applied to time-dependent problems using implicit time integrators. These integrators stem from the 'classical' class of backward differentiation formulae (BDF) and diagonally implicit Runge–Kutta (DIRK) methods. We extend this to the class of general linear methods (GLMs) that unify multistep and multistage methods into one framework. We focus on diagonally implicit multistage integration methods (DIMSIMs) that can have the same desirable stability properties like DIRK methods while also having high stage order. The presented numerical results confirm that the applied DIMSIMs achieve expected approximation properties for linear and nonlinear problems.

Keywords General linear method · Hybridized discontinuous galerkin method Time-dependent · CFD

1 Introduction

Discontinuous Galerkin methods [12, 20, 21] have been recognized as powerful discretization methods for differential equations stemming from a variety of applications. A severe drawback of these methods is the large number of unknowns compared to the other numerical schemes. This becomes particularly problematic for steady-state problems or stiff time-dependent problems, as those problems are usually solved through implicit solution techniques that couple the unknowns globally.

The number of globally coupled unknowns may be greatly reduced by hybridization [9]. This leads to the class of so-called hybridized discontinuous Galerkin (HDG) methods [1, 17, 18, 22]. HDG has initially been developed for steady-state prob-

A. Jaust (✉) · J. Schütz
Faculty of Sciences, Hasselt University, Agoralaan Gebouw D, 3590 Diepenbeek, Belgium
e-mail: alexander.jaust@uhasselt.be

J. Schütz
e-mail: jochen.schuetz@uhasselt.be

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_4

lems; it can be applied to time-dependent problems though yielding a differential algebraic equation (DAE). The latter needs time integration schemes being able to handle stiff problems. Good results have been obtained with backward differentiation formulae (BDF) and diagonally implicit Runge–Kutta (DIRK) schemes [15, 17, 18]; an approach using multiderivative time integrators has been studied in [16].

In this work, we study the coupling of an HDG method to a diagonally implicit multistage integration method (DIMSIM) in Nordsieck representation [4, 7]. These methods are a subclass of general linear methods (GLMs) [3, 5, 8, 13]. GLMs are a generalization of multistep and multistage methods and therefore contain these schemes as special cases. Their advantage over classical time integrators such as BDF and DIRK methods is that one can obtain methods with high accuracy that are L- and A-stable while having stage order $q > 1$, being particularly important for very stiff problems. Moreover, there are techniques available to adapt the time step size and the order of the scheme [2, 13, 14] and an extension to implicit–explicit (IMEX) methods exists as well [24].

This work is structured as follows: First, we briefly introduce the HDG method and show its coupling to a GLM. Then, numerical results are presented and discussed. We end with conclusions and outlook.

2 Numerical Method

We consider partial differential equations of convection–diffusion type that can be written as

$$w_t + \nabla \cdot (f(w) - f_v(w, \nabla w)) = 0 \qquad \forall\, (x,t) \in \Omega \times [0,T] \tag{1}$$

$$w(x,0) = w_0(x) \qquad \forall\, x \in \Omega \tag{2}$$

on a domain $\Omega \subset \mathbb{R}^2$. The system consists of $m \geq 1$ equations; the functions f and f_v are given, possibly nonlinear functions. Both Euler and Navier–Stokes equations are covered by this framework. As it is frequently done, we reformulate the PDE (1) as a first-order PDE by introducing the additional unknown $\sigma := \nabla w$:

$$\sigma = \nabla w \qquad \forall\, (x,t) \in \Omega \times [0,T] \tag{3}$$

$$w_t + \nabla \cdot (f(w) - f_v(w, \sigma)) = 0 \qquad \forall\, (x,t) \in \Omega \times [0,T] \tag{4}$$

$$w(x,0) = w_0(x) \qquad \forall\, x \in \Omega. \tag{5}$$

If the system is of first order, i.e., $f_v \equiv 0$, then (3) is not needed.

2.1 The Hybridized Discontinuous Galerkin Method

For a proper discretization, the domain Ω has to be partitioned into a set of subdomains such that

$$\Omega = \bigcup_{k=1}^{N} \Omega_k.$$

The number of subdomains is denoted by N. For a hybridized discretization, we also need the set of all edges Γ. It contains intersecting subdomains $\Omega_k \cap \Omega_{k'}$ and subdomains intersecting the domain boundary $\Omega_k \cap \partial\Omega$. The number of all edges is given by $\widehat{N} = |\Gamma|$. Furthermore, we need spaces for the approximations of w, σ and the additional hybrid unknown λ on the edges. The following standard spaces are considered:

$$\begin{aligned}
H_h &:= \{q \in L^2(\Omega) \mid q_{|\Omega_k} \in \Pi^P(\Omega_k)\ \forall k = 1, \ldots, N\}^{2m} \\
V_h &:= \{q \in L^2(\Omega) \mid q_{|\Omega_k} \in \Pi^P(\Omega_k)\ \forall k = 1, \ldots, N\}^{m} \\
M_h &:= \{q \in L^2(\Gamma_l) \mid q_{|\Gamma_l} \in \Pi^P(\Gamma_l) \quad \forall k = 1, \ldots, \widehat{N},\ \Gamma_l \in \Gamma\}^{m}.
\end{aligned}$$

For a shorter notation, we also define the following abbreviations for standard scalar products on elements and edges

$$\begin{aligned}
(h_1, h_2) &:= \sum_{k=1}^{N} \int_{\Omega_k} h_1 \cdot h_2 \, dx, \\
\langle h_1, h_2 \rangle &:= \sum_{k=1}^{N} \int_{\partial\Omega_k} h_1 \cdot h_2 \, d\sigma, \quad \langle h_1, h_2 \rangle_\Gamma := \sum_{l=1}^{\widehat{N}} \int_{\Gamma_l} h_1 \cdot h_2 \, d\sigma.
\end{aligned}$$

Moreover, we use the one-side value of a quantity $u(x)$ at a point $x \in \partial\Omega_k$ or $x \in \Gamma_l$ defined as

$$u^{\pm}(x) := \lim_{\varepsilon \to 0} u(x \pm \varepsilon n)$$

where n is the outward pointing normal vector of $\partial\Omega_k$ or the normal vector defined for Γ_l. The jump operator $[\![\cdot]\!]$ is defined as

$$[\![u]\!] := (u^- - u^+)n, \quad [\![u]\!] := (u^- - u^+) \cdot n$$

for scalar quantities u (left) and for vector quantities u (right).

Applying the HDG method in a standard way yields the task of finding $\sigma_h \in H_h$, $w_h \in V_h$ and $\lambda_h \in M_h$ such that

$$(\sigma_h - \nabla w_h, \tau_h) - \langle \lambda_h - w_h^-, \tau_h^- \cdot n \rangle = 0 \quad \forall \tau_h \in H_h \tag{6}$$

$$((w_h)_t, \varphi_h) - (f(w_h) - f_v(w_h, \sigma_h), \nabla \varphi_h) + \langle (\hat{f} - \hat{f}_v) \cdot n, \varphi_h^- \rangle = 0 \quad \forall \varphi_h \in V_h \tag{7}$$

$$\langle [\![\hat{f} - \hat{f}_v]\!] \cdot n, \mu_h \rangle_\Gamma = 0 \quad \forall \mu_h \in M_h \tag{8}$$

is fulfilled for all times $t \in [0, T]$. The fluxes on element boundaries $\partial \Omega_k$ have been replaced by numerical fluxes

$$\hat{f} := f(\lambda_h) - \alpha(\lambda_h - w_h^-)n, \quad \hat{f}_v := f_v(\lambda_h, \sigma_h^-) + \beta(\lambda_h - w_h^-)n$$

with positive real parameters α and β. The parameters have to be chosen carefully to ensure stability of the scheme. For a detailed description on how boundary conditions are incorporated, we refer to [23]. Note that a time derivative of only w_h occurs in the equation. Therefore, the equations form a set of differential algebraic equations (DAEs).

The number of unknowns in (6)–(8) is larger than for the initial problem where λ_h would be absent. However, this formulation allows to apply static condensation such that the global number of unknowns can be greatly reduced [9].

In order to obtain a more compact notation, we will abbreviate the set of ansatz and test spaces by

$$\mathbb{X}_h := H_h \times V_h \times M_h$$

and the vector of unknowns by

$$\mathbf{w}_h := (\sigma_h, w_h, \lambda_h).$$

Then, we can write (6)–(8) compactly as

$$\mathscr{T}((w_h)_t, \varphi_h) + \mathscr{N}(\mathbf{w}_h; \mathbf{x}_h) = 0, \quad \forall \mathbf{x}_h \in \mathbb{X}_h \tag{9}$$

where $\mathscr{T}$ is the vector containing time derivatives and $\mathscr{N}$ represents the spatial discretization of the problem.

2.2 General Linear Methods

In this work, we discretize (9) using general linear methods [3, 5, 13, 14]. These can be seen as generalization of standard methods like multistage (such as DIRK) or multistep (such as BDF) methods. Multistage methods rely on only $r = 1$ external approximation—the solution at the previous time step—but compute $s \geq 1$ internal approximations during stages. Multistep methods rely on $r \geq 1$ external approximations that are passed from one time step to another, but have only $s = 1$ internal approximation that equals the solution at the new time step. General linear methods

allow the usage of several internal approximations $s \geq 1$ and external approximations $r \geq 1$.

In order to give a brief idea of the method, we start with an ordinary differential equation (ODE)

$$y'(t) = g(t, y), \quad y(0) = y_0 \tag{10}$$

with unknown y, time t and a given initial condition y_0. The ODE shall be solved on a uniform grid in time with $t_n := t_0 + n \cdot \Delta t$. An approximation using a GLM is obtained via

$$Y_i = \sum_{j=1}^{s} a_{ij} \Delta t G_j + \sum_{j=1}^{r} u_{ij} y_j^{[n-1]}, \quad i = 1, \ldots, s \tag{11}$$

$$y_i^{[n]} = \sum_{j=1}^{s} b_{ij} \Delta t G_j + \sum_{j=1}^{r} v_{ij} y_j^{[n-1]}, \quad i = 1, \ldots, r \tag{12}$$

as described in [5]. External approximations are stored in $y_j^{[n-1]}$ for $j = 1, \ldots, r$. Additionally, in each time step s internal approximations Y_i ($i = 1, \ldots, s$) are computed. $G_j := g(Y_j)$ is often referred to as stage derivative because, due to the ODE (10), it describes the derivative of Y_j. This is similar to Runge–Kutta methods. Once all internal approximations are known, the external approximations $y_i^{[n]}$ are updated. The collections of $Y_i, G_j, y_j^{[n-1]}$ and $y_j^{[n]}$ are often written as single vectors consisting of the data

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}, \quad G = \begin{bmatrix} G_1 \\ G_2 \\ \vdots \\ G_s \end{bmatrix}, \quad y^{[n-1]} = \begin{bmatrix} y_1^{[n-1]} \\ y_2^{[n-1]} \\ \vdots \\ y_r^{[n-1]} \end{bmatrix}, \quad y^{[n]} = \begin{bmatrix} y_1^{[n]} \\ y_2^{[n]} \\ \vdots \\ y_r^{[n]} \end{bmatrix}. \tag{13}$$

The shape of the method depends heavily on the choice of values to be stored in $y^{[n]}$ and $y^{[n-1]}$. This also depends on the method and the way it is represented. Pure multistep methods may store solutions at previous times $y_{n-r}, y_{n-r+1}, \ldots, y_{n-1}$, the corresponding derivatives $g(y_{n-r}), g(y_{n-r+1}), \ldots, g(y_{n-1})$ or both. Pure multistage methods only need to store the solution of the previous time y_{n-1}.

Order and stability of the method depend on the careful choice of coefficients a_{ij}, u_{ij}, b_{ij} and v_{ij}. The coefficients of the method can be compactly written as matrices

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix}, \quad A \in \mathbb{R}^{s \times s}, B \in \mathbb{R}^{r \times s}, U \in \mathbb{R}^{s \times r}, V \in \mathbb{R}^{r \times r}. \tag{14}$$

In this work, we focus on a special class of general linear methods that are also called diagonally implicit multistage integration methods (DIMSIMs) [4, 7, 13]. These are closely related to (singly) diagonally implicit Runge–Kutta methods in the

sense that

$$A = \begin{bmatrix} \lambda & & & \\ a_{21} & \lambda & & \\ \vdots & \ddots & \ddots & \\ a_{s1} & \dots & a_{s(s-1)} & \lambda \end{bmatrix}$$

is a lower triangular matrix with nonzero entries on the diagonal. This allows to solve a system in each stage instead of getting a much larger system in case of nonzero entries on the upper triangular part. Furthermore, it is possible to choose the other coefficients in such a way that stability properties are equal to DIRK methods; i.e., A- and L-stable DIMSIMs are available. We use DIMSIMs of order $p = 1$ to $p = 3$ that were presented in [13]. For the applied DIMSIMs, the stage order equals p. We will refer to the methods as DIMSIM1, DIMSIM2, and DIMSIM3 to distinguish between the schemes of different order. Each method has $s = p$ internal and $r = p + 1$ external approximations. These DIMSIMs are formulated in Nordsieck formulation [7, 19] which means that $y^{[n]}$ is the Nordsieck vector

$$y^{[n]} = \begin{bmatrix} y(t_n) \\ \Delta t y'(t_n) \\ \Delta t^2 y^{(2)}(t_n) \\ \vdots \\ \Delta t^r y^{(r)}(t_n) \end{bmatrix} \tag{15}$$

that stores y and its first r derivatives. Using the specific form (15) has the advantage that time step adaptation can be easily incorporated since it only requires rescaling of the Nordsieck vector. This has been successfully applied in [2, 6, 13] to ODEs. In this work, we do not pursue this any further, and leave it for future work.

Because it is extremely unhandy to compute higher derivatives of the ODE's right-hand side, in practice, one usually uses an approximation to the Nordsieck vector [7]. In the case of the first-order method with $r = 2$, it is self-starting since the Nordsieck vector is given by

$$y^{[0]} = \begin{pmatrix} y(t_0) \\ \Delta t g(y(t_0)) \end{pmatrix} = \begin{pmatrix} y_0 \\ \Delta t g(y_0) \end{pmatrix}. \tag{16}$$

Higher order methods require a different approach. In [27], the author constructed special Runge–Kutta schemes that compute an approximation to the Nordsieck vector at $t = 0$. In [13, 14], the author describes an approach where the higher order DIMSIMs are started from lower order DIMSIMs. In this work, we use an approach similar to the starting procedure of backwards differentiation formulae. We use an already available DIRK scheme of suitable order and compute r equidistant approximations to the solution at times $t_i = i \cdot \Delta t,\ i = 1, \dots, r - 1$. These values are used together with the given initial data to construct an approximation to the Nordsieck vector using interpolation.

2.3 Applying DIMSIMs to the HDG Method

In (9), the semidiscrete form of the equations is already in the shape of a DAE. Therefore, we have to solve (11)–(12) with slightly modified notation. In each stage i of the method, we compute an internal approximation by solving

$$\mathscr{T}(w_h^{n,i}, \varphi_h) = -\Delta t \sum_{j=1}^{i} a_{ij} \mathscr{N}(\mathbf{w}_h^{n,i}; \mathbf{x}_h) + \sum_{j=1}^{r} u_{ij} \mathscr{T}(y_j^{[n-1]}, \varphi_h), \quad \forall \mathbf{x}_h \in \mathbb{X}_h. \tag{17}$$

(Note that we have defined $\mathbf{x}_h = (\tau_h, \varphi_h, \mu_h)$.) Once all stage values $w_h^{n,i}$ are known, we obtain the updated solution from

$$\mathscr{T}(y_i^{[n]}, \varphi_h) = -\sum_{j=1}^{s} b_{ij} \Delta t \mathscr{N}(\mathbf{w}_h^{n,i}; \mathbf{x}_h) + \sum_{j=1}^{r} v_{ij} \mathscr{T}(y_j^{[n-1]}, \varphi_h) \tag{18}$$

which only requires the local inversion of a mass matrix on each element. Here, $y^{[n-1]}$ stores an approximation to the Nordsieck vector,

$$y^{[n-1]} = \begin{bmatrix} w_h^{n-1} \\ \Delta t \frac{\mathrm{d}}{\mathrm{d}t} w_h^{n-1} \\ \vdots \\ \Delta t^r \frac{\mathrm{d}^r}{\mathrm{d}t^r} w_h^{n-1} \end{bmatrix} + \mathscr{O}(\Delta t^{p+1}). \tag{19}$$

Note that p is the order of the applied DIMSIM.

3 Numerical Results

In this section, we present numerical results obtained from the HDG discretization with DIMSIM time integrators in order to verify the approach. The first test case is a linear convection–diffusion equation where the exact solution is known. In the second test case, the more involved Navier–Stokes equations are solved and the results are compared to other numerical experiments. The system of equations is solved using Newton's method until the absolute residual drops below 10^{-10}. The arising linear system is then solved with a restarted GMRES until the relative residual drops below 10^{-12} for the first and 10^{-10} for the second test problem.

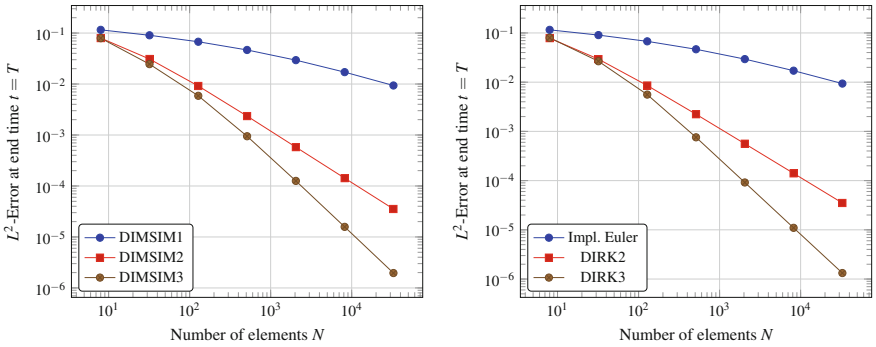


Fig. 1 Errors of DIMSIM schemes of order 1 to 3 (left) and Runge–Kutta schemes of same order (right) are presented

3.1 Linear Convection–Diffusion Equation

We first consider a 2D Gaussian that rotates on the domain $\Omega = [-0.5, 0.5]^2$ in counterclockwise direction. The same problem has been studied for HDG for different time integrators also in previous works; see, for example, [15, 17]. The flux functions are

$$f(w) = (-4x, 4y)^T, \quad f_v(w, \nabla w) = \varepsilon \nabla w$$

with given diffusion constant $\varepsilon = 10^{-3}$. The computation is run until final time $T = \frac{\pi}{4}$, and Dirichlet boundary conditions are specified everywhere. The solution to this problem is known, allowing us to compute the error of our method and check for correct order of convergence. The coarsest grid has $N = 8$ triangular elements, and the time step on this grid is $\Delta t = \frac{\pi}{16}$. We use polynomials of degree $P = 2$ for the spatial discretization such that the expect spatial order of convergence is $P + 1 = 3$. This shall not affect the order of convergence in time because we consider methods of order $p = 3$ at most.

In Fig. 1, we present the errors obtained under uniform refinement. On the left we show the solution for DIMSIM time integration, and on the right the solution for DIRK time integration with same order of accuracy in time. We observe that the methods retain the correct order of convergence in time. Moreover, the errors produced are almost identical to the ones obtained from the classical DIRK discretizations.

3.2 Navier–Stokes Equations

As second test case, we consider the compressible Navier–Stokes equations in two space dimensions. They are also of convection–diffusion type, but the fluxes are nonlinear functions and it is a system of four equations. This makes it a much more

Table 1 Values of the Strouhal number Sr and the drag coefficient from the literature

	Sr	c_D
Gopinath [10]	0.1866	1.3406
Henderson [11]	—	1.336
Williamson [25]	0.1919	—

Table 2 Values of the Strouhal number Sr and the drag coefficient for DIMSIM1

$P = 1$	Sr	c_D
$\Delta t = 1$	0.1733	1.2110
$\Delta t = 5$	0.0000	0.9723
$\Delta t = 10$	0.0000	0.9723
$P = 2$	Sr	c_D
$\Delta t = 1$	0.1733	1.2164
$\Delta t = 5$	0.1222	0.9542
$\Delta t = 10$	0.0000	0.9228
$P = 3$	Sr	c_D
$\Delta t = 1$	0.1733	1.2171
$\Delta t = 5$	0.1238	0.9492
$\Delta t = 10$	0.0000	0.9181

complicated compared to the first problem. A description of the fluxes can be found in [22].

We consider the flow around a cylinder at Reynolds number $\mathrm{Re} = 180$ and Mach number $\mathrm{Ma} = 0.2$. At these flow conditions, vortices shed from the cylinder which is known as Kármán vortex street. We compute the solution on a mesh that extends to 20 diameters around the cylinder. The mesh has $N = 2916$ elements, and it is the same that has been used in previous publications [15, 26]. The newly implemented time-stepping schemes are evaluated using three different time step sizes $\Delta t \in \{1, 5, 10\}$ and polynomials of degree $P \in \{1, 2, 3\}$. The flow field is initialized with free stream conditions. At simulation time around $t \approx 750$, the vortex street develops. We look at the fully evolved vortex street in the interval $t \in [1{,}000; 10{,}000]$ and compute the mean drag coefficient c_D and the Strouhal number Sr. These can be compared to data from the literature given in Table 1. The results from our computations are given in Tables 2, 3 and 4.

The DIMSIM1 fails to obtain an unsteady solution for low polynomial degrees and/or large time step because it is not accurate enough to catch the time-dependent features of the solution (see Table 2). We have observed similar behavior for time steps $\Delta t > 10$ even for higher order time integrators before [15]. Therefore, it is crucial to choose a discretization in time that is sufficiently accurate. This may be achieved by reducing the time step size Δt or by using time integrators of higher accuracy. Sufficiently small time steps Δt and large P finally reveal the time-dependent nature

Table 3 Values of the Strouhal number Sr and the drag coefficient for DIMSIM2

$P = 1$	Sr	c_D
$\Delta t = 1$	0.1898	1.3455
$\Delta t = 5$	0.1849	1.3449
$\Delta t = 10$	0.1733	1.3317
$P = 2$	Sr	c_D
$\Delta t = 1$	0.1898	1.3649
$\Delta t = 5$	0.1882	1.3714
$\Delta t = 10$	0.1774	1.3401
$P = 3$	Sr	c_D
$\Delta t = 1$	0.1898	1.3651
$\Delta t = 5$	0.1882	1.3727
$\Delta t = 10$	0.1774	1.3405

Table 4 Values of the Strouhal number Sr and the drag coefficient for DIMSIM3

$P = 1$	Sr	c_D
$\Delta t = 1$	0.1898	1.3448
$\Delta t = 5$	0.1832	1.3216
$\Delta t = 10$	0.1667	1.2803
$P = 2$	Sr	c_D
$\Delta t = 1$	0.1898	1.3645
$\Delta t = 5$	0.1882	1.3377
$\Delta t = 10$	0.1708	1.2840
$P = 3$	Sr	c_D
$\Delta t = 1$	0.1898	1.3649
$\Delta t = 5$	0.1882	1.3385
$\Delta t = 10$	0.1708	1.2845

of the problem. The obtained values for the mean drag coefficient and the Strouhal number are still off from the values obtained from literature, but tend to grow for decreasing Δt and therefore get closer to the reference data.

DIMSIM2 and DIMSIM3 time integrators show much better results even for large time step sizes. In all cases, the time-dependent behavior of the flow has been recognized (see Tables 3 and 4). Even for spatial discretizations with $P = 1$, the observed Strouhal number is close to the one obtained for $P > 1$. One sees that for $\Delta t = 1$, the Strouhal number rounded to four significant digits coincides for all P and both integrators. Even for a fixed larger time step and $P > 1$, the respective Strouhal numbers tend to coincide at least for the respective time integrator. This also means that the results rather depend on the resolution in time than in space. In the case of $\Delta t < 10$ and $P > 1$, the observed Strouhal numbers are in between the

Strouhal numbers reported from the literature (see Table 1) and therefore seem to be reasonable.

The weak dependency of the obtained coefficients on P can also be observed in the mean drag coefficient. The difference in the values of the coefficient of a given integrator for two different values of P is much lower than for different time step sizes. It is interesting to see that for DIMSIM3 for all P and DIMSIM2 for $P = 1$ the mean drag coefficient slightly grows with decreasing Δt. However, DIMSIM2 with $P > 1$ seems to overpredict the drag coefficient at $\Delta t = 5$ which drops down to lower values for $\Delta t = 1$. These values are very close to the one obtained by DIMSIM3 and $\Delta t = 1$. In comparison to the mean drag coefficients from literature (see Table 1), we obtain slightly larger values for $\Delta t = 1$. Nevertheless, the values are still in a reasonable range.

It shows that the values obtained for the drag coefficient and the Strouhal number are reasonable compared to data from literature if a resolution in space and especially time is sufficient. In our setting, the DIMSIM2 and DIMSIM3 time integrators show reasonable results even when used with rather large time step sizes while the DIMSIM1 time integrator suffers from its low approximation order.

4 Conclusion and Outlook

We have presented the application of general linear methods to an HDG discretization in space. The resulting system of equations is similar to the system one obtains from classical BDF of DIRK methods. The numerical experiments confirm the expected order of convergence in time and the plausibility of the results.

Future work will include the evaluation of the performance of DIMSIM time integrators for HDG schemes. In this setting, time step adaption is crucial in order to be competitive to other methods. Another interesting topic is the coupling of implicit–explicit (IMEX) general linear methods with the hybridized discontinuous Galerkin methods.

Acknowledgements This study was supported by the Special Research Fund (BOF) of Hasselt University. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government—department EWI.

References

1. T. Bui-Thanh, From Godunov to a unified hybridized discontinuous Galerkin framework for partial differential equations. J. Comput. Phys. **295**, 114–146 (2015)
2. J. Butcher, H. Podhaisky, On error estimation in general linear methods for stiff ODEs. Appl. Numer. Math. **56**(3), 345–357 (2006)

3. J.C. Butcher, On the convergence of numerical solutions to ordinary differential equations. Math. Comput. **20**(93), 1–10 (1966)
4. J.C. Butcher, Diagonally-implicit multi-stage integration methods. Appl. Numer. Math. **11**(5), 347–363 (1993)
5. J.C. Butcher, General linear methods. Acta Numer. **15**, 157–256 (2006)
6. J.C. Butcher, Z. Jackiewicz, A reliable error estimation for diagonally implicit multistage integration methods. BIT Numer. Math. **41**(4), 656–665 (2001)
7. J.C. Butcher, P. Chartier, Z. Jackiewicz, Nordsieck representation of DIMSIMs. Numer. Algorithms **16**(2), 209–230 (1997)
8. A. Cardone, Z. Jackiewicz, J.H. Verner, B. Welfert, Order conditions for general linear methods. J. Comput. Appl. Math. **290**, 44–64 (2015)
9. B. Cockburn, J. Gopalakrishnan, R. Lazarov, Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM J. Numer. Anal. **47**, 1319–1365 (2009)
10. A. Gopinath, A. Jameson, Application of the time spectral method to periodic unsteady vortex sheeding. AIAA Paper 06-0449 (2006)
11. R.D. Henderson, Details of the drag curve near the onset of vortex shedding. Phys. Fluids **7**, 2102–2104 (1995)
12. J.S. Hesthaven, T. Warburton, *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Texts in Applied Mathematics, vol. 54 (Springer, Berlin, 2008)
13. Z. Jackiewicz, Implementation of DIMSIMs for stiff differential systems. Appl. Numer. Math. **42**(1–3), 251–267 (2002)
14. Z. Jackiewicz, Construction and implementation of general linear methods for ordinary differential equations: a review. J. Sci. Comput. **25**(1), 29–49 (2005)
15. A. Jaust, J. Schütz, A temporally adaptive hybridized discontinuous Galerkin method for time-dependent compressible flows. Comput. Fluids **98**, 177–185 (2014)
16. A. Jaust, J. Schütz, D.C. Seal, Implicit multistage two-derivative discontinuous Galerkin schemes for viscous conservation laws. J. Sci. Comput. **69**, 866–891 (2016)
17. N.C. Nguyen, J. Peraire, B. Cockburn, An implicit high-order hybridizable discontinuous Galerkin method for linear convection-diffusion equations. J. Comput. Phys. **228**, 3232–3254 (2009)
18. N.C. Nguyen, J. Peraire, B. Cockburn, High-order implicit hybridizable discontinuous Galerkin methods for acoustics and elastodynamics. J. Comput. Phys. **230**, 3695–3718 (2011)
19. A. Nordsieck, On numerical integration of ordinary differential equations. Math. Comput. **16**, 22–49 (1962)
20. D.A.D. Pietro, A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*, vol. 69 (Springer Science & Business Media, New York, 2011)
21. W. Reed, T. Hill, Triangular mesh methods for the neutron transport equation. Technical report, Los Alamos Scientific Laboratory (1973)
22. J. Schütz, G. May, A hybrid mixed method for the compressible Navier-Stokes equations. J. Comput. Phys. **240**, 58–75 (2013)
23. J. Schütz, G. May, An adjoint consistency analysis for a class of hybrid mixed methods. IMA J. Numer. Anal. **34**(3), 1222–1239 (2014)
24. P.E. Vos, C. Eskilsson, A. Bolis, S. Chun, R.M. Kirby, S.J. Sherwin, A generic framework for time-stepping partial differential equations (PDEs): general linear methods, object-oriented implementation and application to fluid problems. Int. J. Comput. Fluid Dyn. **25**(3), 107–125 (2011)
25. C. Williamson, Vortex dynamics in the cylinder wake. Annu. Rev. Fluid Mech. **28**, 477–539 (1996)
26. M. Woopen, G. May, J. Schütz, Adjoint-based error estimation and mesh adaptation for hybridized discontinuous Galerkin methods. Int. J. Numer. Methods Fluids **76**, 811–834 (2014)
27. W. Wright, General linear methods with inherent Runge-Kutta stability. Ph.D. thesis, University of Auckland (2002)

An Invariant-Region-Preserving (IRP) Limiter to DG Methods for Compressible Euler Equations

Yi Jiang and Hailiang Liu

Abstract We introduce an explicit invariant-region-preserving limiter applied to DG methods for compressible Euler equations. The invariant region considered consists of positivity of density and pressure and a maximum principle of a specific entropy. The modified polynomial by the limiter preserves the cell average, lies entirely within the invariant region, and does not destroy the high order of accuracy for smooth solutions, as long as the cell average stays away from the boundary of the invariant region. Numerical tests are presented to illustrate the properties of the limiter. In particular, the tests on Riemann problems show that the limiter helps to damp the oscillations near discontinuities.

Keywords Gas dynamics · Discontinuous Galerkin method · Invariant region

2010 Mathematics Subject Classification 65M60 · 35L65 · 35L45

1 Introduction

We consider the one-dimensional version of the compressible Euler equations for the perfect gas in gas dynamics:

$$\begin{aligned} &\mathbf{w}_t + F(\mathbf{w})_x = 0, \quad t > 0,\ x \in \mathbb{R}, \\ &\mathbf{w} = (\rho, m, E)^\top, \quad F(\mathbf{w}) = (m, \rho u^2 + p, (E + p)u)^\top \end{aligned} \tag{1}$$

with

$$m = \rho u,\ E = \frac{1}{2}\rho u^2 + \frac{p}{\gamma - 1}, \tag{2}$$

Y. Jiang · H. Liu (✉)
Department of Mathematics, Iowa State University, Ames, IA 50010, USA
e-mail: hliu@iastate.edu

Y. Jiang
e-mail: yjiang1@iastate.edu

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_5

where $\gamma > 0$ is a constant ($\gamma = 1.4$ for the air), ρ is the density, u is the velocity, m is the momentum, E is the total energy, and p is the pressure; supplemented by initial data $\mathbf{w}_0(x)$. For the associated entropy function $s = \log\left(\frac{p(x)}{\rho^\gamma(x)}\right)$, it is known that

$$A = \{(\rho, m, E)^\top, \rho > 0,\ \ p > 0,\ \ s \geq s_0\} \tag{3}$$

for any $s_0 \in \mathbb{R}$ is an invariant region in the sense that if $\mathbf{w}_0(x) \in A$, then $\mathbf{w}(x, t) \in A$ for all $t > 0$ (see e.g. [3, 10]). At numerical level this set is proved to be invariant by the first-order Lax–Friedrichs scheme (see [1]), and by the first-order finite element method (see [2]), in which a larger class of hyperbolic conservation laws is considered. It is difficult, if not impossible, to preserve such set by a high-order numerical method unless some nonlinear limiter is imposed at each step while marching in time. In this work, we design such a limiter.

In recent years, an interesting mathematical literature has developed devoted to high-order maximum-principle-preserving schemes for scalar conservation equations (see [13]) and positivity-preserving schemes for hyperbolic systems including compressible Euler equations (see e.g. [7, 14, 16]). In [7] up to third-order positivity-preserving finite volume schemes are constructed based on positivity-preserving properties by the corresponding first-order schemes for both density and pressure of one- and two-dimensional compressible Euler equations. Following [7], positivity-preserving high-order DG schemes for compressible Euler equations were first introduced in [14], where the limiter in [13] is generalized. A recent work by Zhang and Shu in [15] introduced a minimum-entropy-principle-preserving limiter for high-order schemes to the compressible Euler equation. In their work, the limiter for entropy part is enforced separately from the ones for the density and pressure and is given implicitly with the limiter parameter solved by Newton's iteration.

For the isentropic gas dynamics, the invariant region is bounded by two global Riemann invariants; for which the authors have designed an explicit limiter in [4] to preserve the underlying invariant region, called an invariant-region-preserving (IRP) limiter. Our goals in this work are to design an IRP limiter for the compressible Euler system (1) and to rigorously prove that such a limiter does not destroy the high-order accuracy. Our limiter differs from that in [15] in two aspects: (i) it is given in an explicit form; (ii) the scaling reconstruction depends on a uniform parameter for the whole vector solution polynomial; in addition to the rigorous proof of the preservation of the accuracy by the limiter. As a result, the limiter preserves the positivity of density and pressure and also a maximum principle for the specific entropy [11], with reduced computational costs in numerical implementations.

2 The Limiter

We construct a novel limiter based on both the cell average (strictly in A) and the high-order polynomial approximation, which is not entirely in A; through a linear convex combination as in [13, 15].

2.1 Averaging is Contraction

For initial density $\rho_0 > 0$ and pressure $p_0 > 0$, we fix

$$s_0 = \inf_x \log\left(\frac{p_0(x)}{\rho_0^\gamma(x)}\right), \tag{4}$$

and define $q = (s_0 - s)\rho$, then the set A is equivalent to the following set:

$$\Sigma = \{\mathbf{w} : \quad \rho > 0,\ p > 0,\ q \le 0\}, \tag{5}$$

which is convex due to the concavity of p and convexity of q. By using set Σ we are able to work out an explicit limiter which has the invariant-region-preserving property. Numerically, the set of admissible states is defined as

$$\Sigma^\varepsilon = \{\mathbf{w} : \quad \rho \ge \varepsilon,\ p \ge \varepsilon,\ q \le 0\}, \tag{6}$$

with its interior denoted by

$$\Sigma_0^\varepsilon = \{\mathbf{w} : \quad \rho > \varepsilon,\ p > \varepsilon,\ q < 0\}, \tag{7}$$

where ε is a small positive number chosen (say as 10^{-13} in practice) so that q is well defined.

For any bounded interval I (or bounded domain in multi-dimensional case), we define the average of $\mathbf{w}(x)$ by

$$\bar{\mathbf{w}} = \frac{1}{|I|}\int_I \mathbf{w}(x)dx \tag{8}$$

where $|I|$ is the measure of I. Such an averaging operator is a contraction:

Lemma 1. *Let $\mathbf{w}(x)$ be non-trivial piecewise continuous vector functions. If $\mathbf{w}(x) \in \Sigma^\varepsilon$ for all $x \in I$, then $\bar{\mathbf{w}} \in \Sigma_0^\varepsilon$ for any bounded interval I.*

Proof. For the entropy part, since q is convex, using Jensen's inequality and the assumption, we have

$$q(\bar{\mathbf{w}}) = q\left(\frac{1}{|I|}\int_I \mathbf{w}(x)dx\right) \le \frac{1}{|I|}\int_I q(\mathbf{w}(x))dx \le 0. \tag{9}$$

With this we can show $q(\bar{\mathbf{w}}) < 0$. Otherwise, if $q(\bar{\mathbf{w}}) = 0$, we must have $q(\mathbf{w}(x)) = 0$ for almost all $x \in I$; that is

$$(s_0 - s(\bar{\mathbf{w}}))\bar{\rho} = (s_0 - s(\mathbf{w}(x)))\rho(x). \tag{10}$$

By taking average of this relation over I on both sides, we have for $g_1 = s\rho$,

$$s_0\bar{\rho} - g_1(\bar{\mathbf{w}}) = s_0\bar{\rho} - \frac{1}{|I|}\int_I g_1(\mathbf{w}(x))dx. \tag{11}$$

This gives

$$\frac{1}{|I|}\int_I g_1(\mathbf{w}(x))dx = g_1(\bar{\mathbf{w}}). \tag{12}$$

By taking the Taylor expansion around $\bar{\mathbf{w}}$, we have

$$g_1(\mathbf{w}(x)) = g_1(\bar{\mathbf{w}}) + \nabla_{\mathbf{w}} g_1(\bar{\mathbf{w}}) \cdot \xi + \xi^\top H_1 \xi, \quad \forall x \in I, \quad \xi := \mathbf{w}(x) - \bar{\mathbf{w}}, \tag{13}$$

which upon integration yields $\frac{1}{|I|}\int_I \xi^T H_1 \xi dx = 0$, where H_1 is the Hessian matrix of g_1. This when combined with the strict concavity of g_1 ensures that $\mathbf{w}(x) \equiv \bar{\mathbf{w}}$, which contradicts the assumption.

We can show $p(\bar{\mathbf{w}}) > \varepsilon$ by a similar contradiction argument. The density part with $\bar{\rho} > \varepsilon$ is obvious.

2.2 Reconstruction

Let $\mathbf{w}_h(x) = (\rho_h(x), m_h(x), E_h(x))^\top$ be a vector of polynomials of degree k over an interval I, which is a high-order approximation to the smooth function $\mathbf{w}(x) = (\rho(x), m(x), E(x))^\top \in \Sigma^\varepsilon$. We assume that the average $\bar{\mathbf{w}}_h \in \Sigma_0^\varepsilon$, but $\mathbf{w}_h(x)$ is not entirely located in Σ^ε for $x \in I$, then we can use the average as a reference in the following reconstruction

$$\tilde{\mathbf{w}}_h(x) = \theta \mathbf{w}_h(x) + (1-\theta)\bar{\mathbf{w}}_h, \tag{14}$$

where

$$\theta = \min\{1, \theta_1, \theta_2, \theta_3\}, \tag{15}$$

with

$$\theta_1 = \frac{\bar{\rho}_h - \varepsilon}{\bar{\rho}_h - \rho_{h,\min}}, \quad \theta_2 = \frac{p(\bar{\mathbf{w}}_h) - \varepsilon}{p(\bar{\mathbf{w}}_h) - p_{h,\min}}, \quad \theta_3 = \frac{-q(\bar{\mathbf{w}}_h)}{q_{h,\max} - q(\bar{\mathbf{w}}_h)}$$

and

$$\rho_{h,\min} = \min_{x\in I} \rho_h(x), \quad p_{h,\min} = \min_{x\in I} p(\mathbf{w}_h(x)), \quad q_{h,\max} = \max_{x\in I} q(\mathbf{w}_h(x)). \tag{16}$$

Note that $p(\bar{\mathbf{w}}_h) > p_{h,\min}$ and $q(\bar{\mathbf{w}}_h) < q_{h,\max}$ due to the concavity of p and convexity of q. Therefore $\theta_i's$ are well defined and positive, for $i = 1, 2, 3$. We can prove that this reconstruction has three desired properties, summarized in the following.

Theorem 1. *The reconstructed polynomial $\tilde{\mathbf{w}}_h(x)$ satisfies the following three properties:*

(i) *the average is preserved, i.e., $\bar{\mathbf{w}}_h = \bar{\tilde{\mathbf{w}}}_h$;*
(ii) *$\tilde{\mathbf{w}}_h(x)$ lies entirely within invariant region Σ^ε, $\forall x \in I$;*
(iii) *order of accuracy is maintained, i.e., $\|\tilde{\mathbf{w}}_h - \mathbf{w}\|_\infty \le C\|\mathbf{w}_h - \mathbf{w}\|_\infty$, provided $\|\mathbf{w}_h - \mathbf{w}\|_\infty$ is sufficient small, where C is a positive constant that only depends on $\bar{\mathbf{w}}_h$, $\mathbf{w}$, and the invariant region Σ^ε.*

Proof. (i) Since $0 < \theta \le 1$ is a uniform constant, average preservation is obvious.
(ii) If $\rho_{h,\min} \ge \varepsilon$, $p_{h,\min} \ge \varepsilon$, and $q_{h,\max} \le 0$, then $\theta = 1$, no reconstruction is needed. When $\theta = \theta_1$, we have

$$\begin{aligned}\tilde{\rho}_h(x) =& \theta_1\rho_h(x) + (1-\theta_1)\bar{\rho}_h \\ =& (\bar{\rho}_h - \varepsilon)\frac{\rho_h(x) - \rho_{h,\min}}{\bar{\rho}_h - \rho_{h,\min}} + \varepsilon \ge \varepsilon.\end{aligned} \tag{17}$$

Since $\theta_1 \le \theta_2$, we have $(p(\bar{\mathbf{w}}_h) - p_{h,\min})\theta_1 + \varepsilon \le p(\bar{\mathbf{w}}_h)$. Therefore, by the concavity of p, we have

$$\begin{aligned}p(\tilde{\mathbf{w}}_h) \ge& \theta_1 p(\mathbf{w}_h) + (1-\theta_1)p(\bar{\mathbf{w}}_h) \\ =& \theta_1(p(\mathbf{w}_h) - p(\bar{\mathbf{w}}_h)) + p(\bar{\mathbf{w}}_h) \\ \ge& \theta_1(p(\mathbf{w}_h) - p(\bar{\mathbf{w}}_h)) + (p(\bar{\mathbf{w}}_h) - p_{h,\min})\theta_1 + \varepsilon \\ =& \theta_1(p(\mathbf{w}_h) - p_{h,\min}) + \varepsilon \ge \varepsilon.\end{aligned} \tag{18}$$

For entropy part, since $\theta_1 \le \theta_3$, we have $\theta_1(q_{h,\max} - q(\bar{\mathbf{w}}_h)) \le -q(\bar{\mathbf{w}}_h)$. Therefore, by the convexity of q, we have

$$\begin{aligned}q(\tilde{\mathbf{w}}_h) <& \theta_1 q(\mathbf{w}_h) + (1-\theta_1)q(\bar{\mathbf{w}}_h) \\ =& \theta_1(q(\mathbf{w}_h) - q(\bar{\mathbf{w}}_h)) + q(\bar{\mathbf{w}}_h) \\ \le& \theta_1(q_{h,\max} - q(\bar{\mathbf{w}}_h)) + q(\bar{\mathbf{w}}_h) \le 0.\end{aligned} \tag{19}$$

In the case that $\theta = \theta_2$ or θ_3 the proof is similar.
(iii) We prove for the case $\theta = \theta_2$, the other cases are similar. In such case, we only need to prove

$$\|\tilde{\mathbf{w}}_h - \mathbf{w}_h\|_\infty \le C\|\mathbf{w}_h - \mathbf{w}\|_\infty, \tag{20}$$

from which (iii) follows by using the triangle inequality. Here and in what follows $\|\cdot\|_\infty := \max_{x\in I}|\cdot|$. We prove (20) in four steps.

Step 1. From (14) it follows that

$$\begin{aligned}\|\tilde{\mathbf{w}}_h - \mathbf{w}_h\|_\infty =& (1-\theta_2)\|\bar{\mathbf{w}}_h - \mathbf{w}_h\|_\infty \\ =& \frac{\max\limits_{x\in I}|\bar{\mathbf{w}}_h - \mathbf{w}_h(x)|}{p(\bar{\mathbf{w}}_h) - p_{h,\min}}(\varepsilon - p_{h,\min}).\end{aligned} \tag{21}$$

Step 2. The overshoot estimate. Since $\mathbf{w}(x) \in \Sigma^\varepsilon$,

$$\varepsilon - p_{h,\min} \le \max_x(p(\mathbf{w}) - p(\mathbf{w}_h)) \le C_1\|\mathbf{w} - \mathbf{w}_h\|_\infty, \quad C_1 := \|\nabla p\|_\infty. \tag{22}$$

Step 3. We map I to $[0, 1]$ by $\xi = (x-a)/(b-a)$ for $I = [a, b]$, and let $l_\alpha(\xi)$ $(\alpha = 1, \cdots, N)$ be the Lagrange interpolating polynomials at quadrature points $\hat{\xi}^\alpha \in [0, 1]$ with $N = k + 1$, then $\mathbf{w}_h(x) - \bar{\mathbf{w}}_h = \sum_{\alpha=1}^N(\mathbf{w}_h(\hat{x}^\alpha) - \bar{\mathbf{w}}_h)l_\alpha(\xi)$, where $\hat{x}^\alpha = a + (b-a)\hat{\xi}^\alpha$. Hence, we have

$$\begin{aligned}\max_{x\in I}|\bar{\mathbf{w}}_h - \mathbf{w}_h(x)| \le& \max_{\xi\in[0,1]}\sum_{\alpha=1}^N |l_\alpha(\xi)||\bar{\mathbf{w}}_h - \mathbf{w}_h(\hat{x}^\alpha)| \\ \le& C_2\max_\alpha|\bar{\mathbf{w}}_h - \mathbf{w}_h(\hat{x}^\alpha)|,\end{aligned} \tag{23}$$

where $C_2 = \Lambda_{k+1}([0, 1]) \doteq \max\limits_{\xi\in[0,1]}\sum\limits_{\alpha=1}^N |l_\alpha(\xi)|$ is the Lebesgue constant. Note that

$$\max_\alpha|\bar{\mathbf{w}}_h - \mathbf{w}_h(\hat{x}^\alpha)| \le \max_\alpha|\bar{\rho}_h - \rho_h(\hat{x}^\alpha)| + \max_\alpha|\bar{m}_h - m_h(\hat{x}^\alpha)| + \max_\alpha|\bar{E}_h - E_h(\hat{x}^\alpha)|. \tag{24}$$

Define

$$\hat{f}_{h,\min} \doteq \min_\alpha f(\mathbf{w}_h(\hat{x}^\alpha)), \quad \hat{f}_{h,\max} \doteq \max_\alpha f(\mathbf{w}_h(\hat{x}^\alpha)), \tag{25}$$

we can show that

$$\max_\alpha|\bar{f}_h - f_h(\hat{x}^\alpha))| \le \max\{\bar{f}_h - \hat{f}_{h,\min}, \hat{f}_{h,\max} - \bar{f}_h\} \le C_3(\bar{f}_h - \hat{f}_{h,min}), \tag{26}$$

where

$$C_3 = \max\left\{1, \frac{1 - \min\limits_\alpha \hat{w}_\alpha}{\min\limits_\alpha \hat{w}_\alpha}\right\}. \tag{27}$$

Here $f_h = \rho_h, m_h, E_h$. The type of estimates using C_2 and C_3 is known, see [12, Lemma 7, Appendix C], where the proof was accredited to Mark Ainsworth.

Step 4. The above three steps lead to

$$\|\tilde{\mathbf{w}}_h - \mathbf{w}_h\|_\infty \le C_1 C_2 C_3 \frac{B}{p(\bar{\mathbf{w}}_h) - p_{h,\min}} \|\mathbf{w} - \mathbf{w}_h\|_\infty, \tag{28}$$

with

$$B = \bar{\rho}_h - \hat{\rho}_{h,min} + \bar{m}_h - \hat{m}_{h,min} + \bar{E}_h - \hat{E}_{h,min}. \tag{29}$$

On one hand, we have $p_{h,min} \le \varepsilon$ since $\theta = \theta_2 \le 1$, leading to

$$p(\bar{\mathbf{w}}_h) - p_{h,\min} \ge p(\bar{\mathbf{w}}_h) - \varepsilon; \tag{30}$$

On the other hand the assumption $\theta = \theta_2 \le \theta_1$ implies

$$\bar{\rho}_h - \hat{\rho}_{h,min} \le \left(\frac{\bar{\rho}_h - \varepsilon}{p(\bar{\mathbf{w}}_h) - \varepsilon} \right) \cdot \left(p(\bar{\mathbf{w}}_h) - p_{h,\min} \right). \tag{31}$$

By the assumption on the smallness of $\|\mathbf{w}_h - \mathbf{w}\|_\infty$ we have

$$\bar{m}_h - \hat{m}_{h,min} \le 2\|m - m_h\|_\infty + \bar{m} - m_{\min} \tag{32}$$

and

$$\bar{E}_h - \hat{E}_{h,min} \le \bar{E} + 1. \tag{33}$$

where $E \ge \frac{\varepsilon}{\gamma - 1}$ is used. Collecting the above estimates we take

$$C_4 = \frac{\bar{\rho}_h - \varepsilon + 2\|m - m_h\|_\infty + \bar{m} - m_{\min} + \bar{E} + 1}{p(\bar{\mathbf{w}}_h) - \varepsilon} \tag{34}$$

to conclude the desired estimate in (iii) with $C = \Pi_{i=1}^4 C_i$.

2.3 Algorithm

Let $\mathbf{w}_h^n$ be the numerical solution generated from a high-order scheme of an abstract form

$$\mathbf{w}_h^{n+1} = \mathscr{L}(\mathbf{w}_h^n), \tag{35}$$

where $\mathbf{w}_h^n = \mathbf{w}_h^n(x) \in V_h$, which is a finite element space of piecewise polynomials of degree k over each computational cell I. Assume $\lambda = \frac{\Delta t}{h}$ is the mesh ratio, where h is the characteristic length of the mesh size.

Provided that scheme (35) has the following property: there exists λ_0, and a test set S_I in each computational cell I such that if

$$\lambda \le \lambda_0 \quad \text{and} \quad \mathbf{w}_h^n \in \Sigma^\varepsilon, \quad x \in S_I \tag{36}$$

then

$$\bar{\mathbf{w}}_h^{n+1} \in \Sigma_0^{\varepsilon}, \tag{37}$$

then the IRP limiter can be applied with I replaced by S_I in (16), i.e.,

$$\rho_{h,\min} = \min_{x\in S_I} \rho_h(x), \quad p_{h,\min} = \min_{x\in S_I} p(\mathbf{w}_h(x)), \quad q_{h,\max} = \max_{x\in S_I} q(\mathbf{w}_h(x)). \tag{38}$$

Our algorithm is given as follows:

Step 1. Initialization: take the piecewise L^2 projection of $\mathbf{w}_0$ onto V_h, such that

$$\int_I (\mathbf{w}_h^0(x) - \mathbf{w}_0(x))\phi(x)dx = 0, \quad \forall \phi \in V_h. \tag{39}$$

Also from $\mathbf{w}_0$, we compute s_0 as defined in (4) to determine the invariant region Σ^{ε}.

Step 2. Impose the modified limiter (14), (15) with (38) on $\mathbf{w}_h^n$ for $n = 0, 1, \cdots$.

Step 3. Update by the scheme:

$$\mathbf{w}_h^{n+1} = \mathscr{L}(\tilde{\mathbf{w}}_h^n). \tag{40}$$

Return to **Step 2**.

Remark 1. Indeed the limiter (14), (15) with (38) can well enhance the efficiency of computation, and we will use this modified IRP limiter in the numerical experiments. Note that with (38), (i) and (iii) in Theorem 1 remain valid, and the resulting reconstructed polynomial lies within invariant region Σ^{ε} only for $x \in S_I$.

Remark 2. Notice that Lemma 1 ensures that $\bar{\mathbf{w}}_h^0$ lies strictly within Σ_0^{ε}; therefore, the limiter is valid already at the initialization step.

Remark 3. Some sufficient conditions for (36) to ensure the cell average propagation property (37) for the DG method have been obtained for one-dimensional case [14], as well as for rectangular meshes [14, 15] and triangular meshes ([16]) in two-dimensional cases. For example, the test set S_I and the CFL condition given in [14, Theorem 2.1] is

$$S_I = \{\hat{x}^{\alpha}, \alpha = 1, \cdots, N\}, \tag{41}$$

which is a set of N-point Legendre Gauss–Lobatto quadrature on I with $2N - 3 \geq k$, and

$$\lambda \|(|u| + c)\|_{\infty} \leq \frac{1}{2}\hat{w}_1, \tag{42}$$

where $\hat{w}_1$ is the first Legendre Gauss–Lobatto quadrature weights for the interval $[-\frac{1}{2}, \frac{1}{2}]$ such that $\sum_{\alpha}^{N} \hat{w}_{\alpha} = 1$.

3 Numerical Tests

We present numerical tests for the IRP limiter applied to a general high-order DG scheme with the Lax–Friedrich numerical flux, using a proper time discretization. The semi-discrete DG scheme we take is a closed ODE system of the form

$$\frac{d}{dt}\mathbf{W} = L(\mathbf{W}), \tag{43}$$

where $\mathbf{W}$ consists of the unknown coefficients of the numerical solution in terms of the spatial basis, and L is the corresponding spatial operator.

We consider the following two types of time discretizations.

- The third-order SSP Runge–Kutta (RK3) method in [9] reads as

$$\begin{aligned}\mathbf{W}^{(1)} =& \mathbf{W}^n + \Delta t L(\mathbf{W}^n)\\ \mathbf{W}^{(2)} =& \frac{3}{4}\mathbf{W}^n + \frac{1}{4}\mathbf{W}^{(1)} + \frac{1}{4}\Delta t L\left(\mathbf{W}^{(1)}\right)\\ \mathbf{W}^{n+1} =& \frac{1}{3}\mathbf{W}^n + \frac{2}{3}\mathbf{W}^{(2)} + \frac{2}{3}\Delta t L\left(\mathbf{W}^{(2)}\right).\end{aligned} \tag{44}$$

- The third-order SSP multi-stage (MS) method in [8] reads as

$$\mathbf{W}^{n+1} = \frac{16}{27}(\mathbf{W}^n + 3\Delta t L(\mathbf{W}^n)) + \frac{11}{27}\left(\mathbf{W}^{n-3} + \frac{12}{11}\Delta t L(\mathbf{W}^{n-3}\right). \tag{45}$$

We apply the limiter at each time stage or each time step.

Remark 4. In the implementation of the third-order SSP multi-step method, we apply SSP RK3 method in the first three-time evolutions to obtain the starting values.

In all of the following examples $\gamma = 1.4$ is taken.

Example 1. *Accuracy Test*
We first test the accuracy of the IRP DG scheme. The initial condition is

$$\rho_0(x) = 1 + \frac{1}{2}\sin(2\pi x), \quad u_0(x) = 1, \quad p_0(x) = 1. \tag{46}$$

The domain is [0, 1] and the boundary condition is periodic. The exact solution is

$$\rho(x, y, t) = 1 + \frac{1}{2}\sin(2\pi(x - t)), \quad u(x, t) = 1, \quad p(x, t) = 1. \tag{47}$$

The results presented in Tables 1 and 2 show that using IRP limiter does not destroy high-order accuracy.

Table 1 Numerical accuracy study of the p^2 DG method

P^2 DG	SSP RK				SSP multi-step			
N	L^∞Error	Order	L^1Error	Order	L^∞Error	Order	L^1Error	Order
8	5.43E-04	/	5.77E-04	/	5.35E-04	/	5.70E-04	/
16	8.98E-05	2.60	8.55E-05	2.75	8.89E-05	2.59	8.53E-05	2.74
32	1.04E-05	3.11	1.09E-05	2.98	1.03E-05	3.11	1.08E-05	2.99
64	1.33E-06	2.97	1.40E-06	2.95	1.34E-06	2.94	1.39E-06	2.95
128	1.67E-07	2.99	1.75E-07	3.00	1.75E-07	2.94	1.76E-07	2.98

Table 2 Numerical accuracy study of the p^3 DG method

P^3 DG	SSP RK				SSP multi-step			
N	L^∞Error	Order	L^1Error	Order	L^∞Error	Order	L^1Error	Order
8	1.44E-05	/	1.09E-05	/	1.42E-05	/	1.08E-05	/
16	1.39E-06	3.37	7.23E-07	3.92	1.37E-06	3.37	7.07E-07	3.94
32	7.06E-08	4.30	6.14E-08	3.56	6.93E-08	4.31	5.99E-08	3.56
64	6.34E-09	3.48	3.18E-09	4.27	6.21E-09	3.48	3.03E-09	4.30
128	3.50E-10	4.18	2.12E-10	3.91	3.30E-10	4.23	1.97E-10	3.94

In the following examples, we solve (1) subject to several different Riemann initial data. We compare the numerical solution obtained from the DG scheme with IRP limiter (14), (15) with (38) and the one obtained from the DG scheme with only positivity-preserving limiter, that is, using $\theta = \min\{1, \theta_1, \theta_2\}$, where θ_1 and θ_2 are defined as in (15).

Example 2. *Lax Shock Tube Problem*
Consider the Lax initial data:

$$(\rho, m, E) = \begin{cases} (0.445, 0.311, 8.928), & x < 0, \\ (0.5, 0, 1.4275), & x > 0, \end{cases} \tag{48}$$

which induces a composite wave, a rarefaction wave followed by a contact discontinuity and then by a shock. We calculate the exact solution by following the formulas given in [6, Sect. 14.11]. The P^2-DG scheme with SSP RK3 method in time discretization is tested on $N = 100$ cells over $x \in [-2, 2]$ at final time $T = 0.5$. From Fig. 1, we see that the IRP limiter helps to damp oscillations near the discontinuities.

Example 3. *Shu–Osher Shock Tube Problem*
Consider the Shu–Osher problem:

$$(\rho, u, p) = \begin{cases} (3.857143, 2.629369, 10.3333), & x < -4, \\ (1 + 0.2 \sin 5x, 0, 1), & x \geq -4. \end{cases} \tag{49}$$

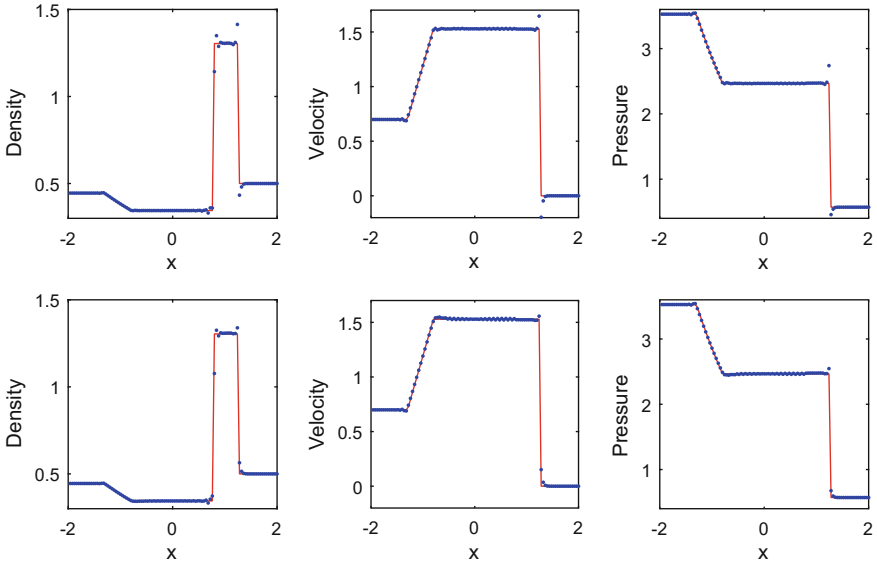


Fig. 1 Lax shock tube problem. Exact solution (solid line) versus numerical solution (dots); Top: with positive-preserving limiter; Bottom: with IRP limiter

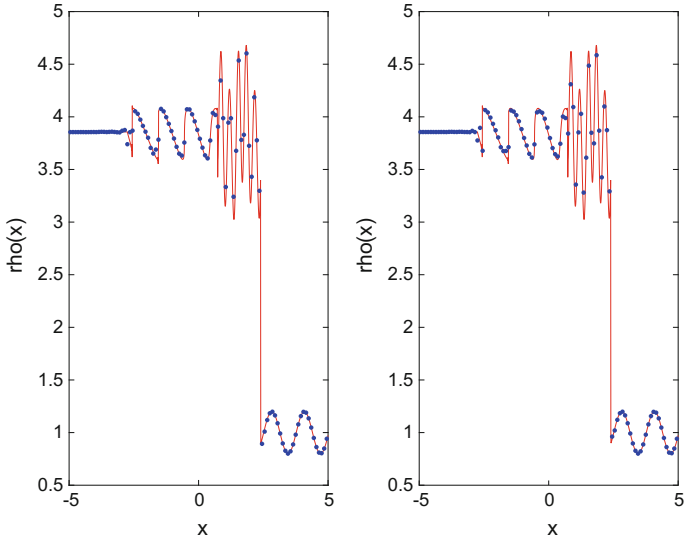


Fig. 2 Shu–Osher problem. Exact solution (solid line) versus numerical solution (dots); Left: with positive-preserving limiter; Right: with IRP limiter

The P^2-DG scheme with SSP RK3 method in time discretization is tested on $N = 100$ cells over $x \in [-5, 5]$ at final time $T = 1.8$. The reference solution is obtained from P^2-DG scheme with SSP RK3 method on $N = 2560$ cells. The results presented in Fig. 2 show that the shock is captured well.

4 Conclusion and Future Work

In this work, we introduced a novel IRP limiter for the one-dimensional compressible Euler equations. The limiter is made so that the reconstructed polynomial preserves the cell average, lies entirely within the invariant region, and does not destroy the original high-order of accuracy for smooth solutions. Moreover, this limiter is explicit and easy for computer implementation. Let us point out that the IRP limiter (14) may be applied to multi-dimensional compressible Euler equations as well if we replace I in (16) by multi-dimensional cells or test set in each cell. Implementation details are in a forthcoming paper [5]. Future work would be to investigate IRP limiters for more general hyperbolic systems or specific systems in important applications.

Acknowledgements This work was supported in part by the National Science Foundation under Grant DMS1312636 and by NSF Grant RNMS (KI-Net) 1107291.

References

1. H. Frid, Maps of convex sets and invariant regions for finite-difference systems of conservation laws. Arch. Rational Mech. Anal. **160**(3), 245–269 (2001)
2. J.-L. Guermond, B. Popov, Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J. Numer. Anal. **54**(4), 2466–2489 (2016)
3. D. Hoff, Invariant regions for systems of conservation laws. Trans. Am. Math. Soc. **289**(2), 591–610 (1985)
4. Y. Jiang, H. Liu, An invariant-region-preserving limiter for DG schemes to isentropic Euler equations. To appear in Numerical Methods Partial Differential Equation (2018)
5. Y. Jiang, H. Liu, Invariant-region-preserving DG Methods for Multi-dimensional Hyperbolic Conservation Law Systems, with an Application to Compressible Euler Equations. J. Comput. Phys. (2018). http://doi.org/10.1016/j.jcp.2018.03.004
6. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, vol. 31 (Cambridge University Press, Cambridge, 2002)
7. B. Perthame, C.-W. Shu, On positivity preserving finite volume schemes for Euler equations. Numerische Mathematik **73**, 119–130 (1996)
8. C.-W. Shu, Total-variation-diminishing time discretizations. SIAM J. Sci. Stat. Comput. **9**, 1073–1084 (1988)
9. C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)
10. J. Smoller, *Shock Waves and Reaction-Diffusion Equations*. Grundlehren der Mathematischen Wissenschaften, vol. 258 (Springer, New York, 1983)
11. E. Tadmor, A minimum entropy principle in the gas dynamics equations. Appl. Numer. Math. **2**, 211–219 (1986)

12. X. Zhang, On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier-Stokes equations. J. Comput. Phys. **328**, 301343 (2017)
13. X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws. J. Comput. Phys. **229**, 3091–3120 (2010)
14. X. Zhang, C.-W. Shu, On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. J. Comput. Phys. **229**, 8918–8934 (2010)
15. X. Zhang, C.-W. Shu, A minimum entropy principle of high order schemes for gas dynamics equations. Numerische Mathematik **121**, 545–563 (2012)
16. X. Zhang, Y. Xia, C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes. J. Sci. Comput. **50**, 29–62 (2012)

β-Schemes with Source Terms and the Convergence Analysis

Nan Jiang

Abstract The schemes concerned in this study are non-homogeneous β-schemes for $m = 2$. The homogeneous counterparts (HCPs) of the schemes were constructed by Osher and Chakravarthy (J Oscil Theory Comput Methods Compens Compact 229–274, 1986, [8]). The entire families of β-schemes are defined for $0 < \beta \leq (m\binom{2m}{m})^{(-1)}$, where m is an integer between 2 and 8. These schemes are $2m - 1$ order accurate, variation diminishing, $2m + 1$ point bandwidth, conservative approximations to the conservation laws. Although the numerical results have been shown to be very effective (Osher and Chakravarthy in J Oscil Theory Comput Methods Compens Compact 229–274, 1986, [8], Osher and Chakravarthy in SIAM J Numer Anal 21:955–984 1984, [7]), the entropy convergence of these schemes has been open. The goal of this paper is to show that, when $m = 2$, β-schemes with source terms indeed persist entropy consistency for non-homogeneous scalar convex conservation laws by using author's recent result on extended Yang's wave tracing theory (Jiang in On wave-wise entropy inequalities for high-resolution schemes with source terms II: the fully-discrete case, submitted, [4], Yang in SIAM J Numer Anal 36(1):1–31, 1999, [10]). The entropy convergence of the HCPs of these schemes was established by the author (Jiang in Int J Numer Anal Model 14(1):103–125, 2017, [6]).

Keywords β-schemes with source terms · Entropy convergence · Conservation laws

1 Introduction

We consider numerical approximations to the scalar conservation laws

N. Jiang (✉)
Department of Mathematical Sciences, University of South Dakota,
414 E. Clark St., Vermillion, SD 57069, USA
e-mail: njiang@usd.edu

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_6

$$\begin{cases} u_t + f(u)_x = q(u), \\ u(x,0) = u_0(x), \end{cases} \tag{1}$$

where $f \in C^1(\mathbb{R})$, $q \in C^1(\mathbb{R})$, and $u_0 \in BV(\mathbb{R})$. Here, BV stands for the subspace of L^1_{loc} consisting of functions with bounded total variation. For the numerical methods concerned, let $\lambda = \frac{\tau}{h}$ be fixed, where h and τ are spatial and temporal steps, respectively, and $u_k^n = u(x_k, t_n)$ be the nodal values of the piecewise constant mesh function $u_h(x,t)$ approximating the solution of (1). In this study, numerical schemes admit the conservative form

$$u_k^{n+1} = H(u_{k-p}^n, \cdots, u_{k+p}^n; \lambda) = u_k^n - \lambda(g_{k+1/2}^n - g_{k-1/2}^n) + \tau q(u_k^n), \tag{2}$$

where the numerical flux g is given by $g_{k+1/2}^n = g_{k+1/2}[u_k^n]$, and

$$g_{k+1/2}[v] = g(v_{k-p+1}, v_{k-p+2}, \cdots, v_k, \cdots, v_{k+p}), \tag{3}$$

for any data $\{v_j\}$. The function g is Lipschitz continuous with respect to its $2p$ arguments and is *consistent* with the conservation law in the sense that

$$g(u, u, \cdots, u) \equiv f(u). \tag{4}$$

In particular, the schemes that we will focus on are special cases of non-homogeneous β-schemes when $m = 2$. The HCPs of these schemes were introduced by Osher and Chakravarthy [8] in the 1980s. 1980s is very productive and influential era in terms of analysis and design of numerical methods in order to solve hyperbolic conservation laws. For instance, the emerging of the higher order flux limiter methods was from the necessity in balance of the weakness of the first-order (although convergent) schemes and the oscillations of the higher order ones. Some other approaches based on well-understood design principles were entailed by ENO and WENO, which when combined with SSP methods are truly effective that can achieve uniformly high-order accuracy for systems and multi-dimensional cases. The methods invented in the 1980s are still relevant in recently developed state-of-the-art schemes. They are often embedded in or partial steps of today's modern schemes. For example, they have been integrated into the design and implementation of uniformly high-order bound-preserving [9, 11] as well as entropy stable (ECENO) schemes [2].

The paper is organized as follows. In the remaining part of this section, we review the notions of the extremum paths, extremum traceability of a scheme, and then we establish the sufficient conditions for the extremum traceableness of the schemes, which is necessary for analyzing the entropy convergence of the schemes. In Sect. 2, we present one of the extended Yang's [4] convergence criteria, an important entropy estimate, and finally the main result.

First, we introduce the non-homogeneous β-schemes for the case of $m = 2$. Throughout the paper, to improve the readability, we use the shorthand notations of $f_k^n := f(u_k^n)$, $\Delta u_{k\pm1/2}^n = \pm(u_{k\pm1}^n - u_k^n)$, and $f_{k\pm1/2}^n := \Delta f_{k\pm1/2}^n = \pm(f_{k\pm1}^n - f_k^n)$. Also, whenever there is no ambiguity in the context, we employ the simplified nota-

tions: $u^k := u_k^{n+1}$, $u_k := u_k^n$, $f_k := f_k^n$, and $f_{k\pm1/2}^{\pm} := (f_{k\pm1/2}^n)^{\pm}$, where k and n are the spatial and temporal indexes, respectively. Let $g_{k+1/2}^E := g^E(u_k^n, u_{k+1}^n)$ be the flux of an E-scheme [7] that is characterized by the inequality: sgn $(u_{k+1}^n - u_k^n)[g_{k+1/2}^E - f(u)] \le 0$, for all u in between u_k^n and u_{k+1}^n. Then, the flux differences are defined by

$$f_{k+1/2}^+ = f_{k+1} - g_{k+1/2}^E, \quad \text{and} \quad f_{k+1/2}^- = g_{k+1/2}^E - f_k. \tag{5}$$

At the time level $t = t^n$, for all k, we define a series of local CFL numbers

$$\nu_{k+1/2}^+ = \lambda f_{k+1/2}^+ / \Delta u_{k+1/2}, \quad \text{and} \quad \nu_{k+1/2}^- = \lambda f_{k+1/2}^- / \Delta u_{k+1/2}. \tag{6}$$

Clearly, we have $\nu_{k+1/2}^+ \ge 0$ and $\nu_{k+1/2}^- \le 0$. For convenience, we also set

$$r_k^+ = f_{k-1/2}^+ / f_{k+1/2}^+, \quad \text{and} \quad r_k^- = f_{k+1/2}^- / f_{k-1/2}^-. \tag{7}$$

The operator "minmod," or "mm" in short, is defined by

$$\text{minmod}(x, y) = \text{mm}(x, y) = \begin{cases} x, & \text{if } |x| \le |y| \text{ and } xy > 0, \\ y, & \text{if } |x| > |y| \text{ and } xy > 0, \\ 0, & \text{if } xy \le 0, \end{cases} \tag{8}$$

which can be converted to, divided by x, a monotone-increasing function

$$\phi(r) = \max(0, \min(1, r)) = \begin{cases} 1, & \text{if } r \ge 1, \\ r, & \text{if } 0 \le r \le 1, \\ 0, & \text{if } r \le 0, \end{cases} \tag{9}$$

with $r = y/x$. Clearly, $\phi(r)$ has a symmetry property $\phi(r)/r = \phi(1/r)$, which is very helpful to rewrite a β-scheme into an increment form. The operator of "minmod" of three quantities is defined by

$$\text{minmod}[\, x, y, z \,] = \text{minmod}[\, \text{minmod}[\, x, y \,], z \,],$$

which is independent of the order of x, y, and z. Now, for $m = 2$ and $0 < \beta \le \frac{1}{12}$, a β-scheme with a source term (see [8] for the HCP) is given by

$$u^k = u_k - \lambda (\, g_{k+1/2} - g_{k-1/2} \,) + \tau q(u_k), \tag{10}$$

where

$$\begin{aligned} g_{k+1/2} = g_{k+1/2}^E &- (1/12 + \beta)(f_{k+3/2}^-)^{(1)} - (1/2 - 2\beta)(f_{k+1/2}^-)^{(0)} \\ &+ (1/12 - \beta)(f_{k-1/2}^-)^{(-1)} - (1/12 - \beta)(f_{k+3/2}^+)^{(1)} \\ &+ (1/2 - 2\beta)(f_{k+1/2}^+)^{(0)} + (1/12 + \beta)(f_{k-1/2}^+)^{(-1)}. \end{aligned} \tag{11}$$

The superscripts shown over the $f^{\pm}$ denote flux limited values of $f^{\pm}$ and are computed as follows:

$$(f^-_{k+3/2})^{(1)} = \text{mm}\,[f^-_{k+3/2}, b\, f^-_{k+1/2}] = \phi(r^-_{k+1}/b)\, b\, f^-_{k+1/2} \tag{12}$$

$$(f^-_{k+1/2})^{(0)} = \text{mm}\,[f^-_{k+1/2}, b\, f^-_{k+3/2}] = \phi(b\, r^-_{k+1})\, f^-_{k+1/2} \tag{13}$$

$$\begin{aligned}(f^-_{k-1/2})^{(-1)} &= \text{mm}\,[f^-_{k-1/2}, b\, f^-_{k+1/2}, b\, f^-_{k+3/2}] \\ &= \text{mm}\,[1/br^-_k, \phi(r^-_{k+1})]\, b\, f^-_{k+1/2} = \text{mm}\,[\phi(1/b\, r^-_k), r^-_{k+1}]\, b\, f^-_{k+1/2}\end{aligned} \tag{14}$$

$$\begin{aligned}(f^+_{k+3/2})^{(1)} &= \text{mm}\,[f^+_{k+3/2}, b\, f^+_{k+1/2}, b\, f^+_{k-1/2}] \\ &= \text{mm}\,[\phi(1/b\, r^+_{k+1}), r^+_k]\, b\, f^+_{k+1/2} = \text{mm}\,[1/b\, r^+_{k+1}, \phi(r^+_k)]\, b\, f^+_{k+1/2}\end{aligned} \tag{15}$$

$$(f^+_{k+1/2})^{(0)} = \text{mm}\,[f^+_{k+1/2}, b\, f^+_{k-1/2}] = \phi(b\, r^+_k)\, f^+_{k+1/2} \tag{16}$$

$$(f^+_{k-1/2})^{(-1)} = \text{mm}\,[f^+_{k-1/2}, b\, f^+_{k+1/2}] = \phi(r^+_k/b)\, b\, f^+_{k+1/2} \tag{17}$$

In the above, b is a "compression" parameter chosen in the range

$$1 < b \le 3 + 12\beta.$$

We shall assume for the remainder of the paper that the local CFL numbers satisfy $|\Delta\nu^{\pm}_{k+1/2}| \le 1$ for all $k \in \mathbb{Z}$, $f'' \ge 0$ and $q' \ge 0$. Also, we rewrite the schemes (10)–(11) in an increment form (18)–(20) below, which provides a convenient way of checking extremum traceability property of the schemes [6]. We recall that, by Yang [10], an extremum traceable scheme is total variation diminishing (TVD).

$$u^k = H(u^n_{k-p}, \cdots, u^n_{k+p}; \lambda) = u_k - C_{k-1/2}\Delta u_{k-1/2} + D_{k+1/2}\Delta u_{k+1/2} + \tau q(u_k), \tag{18}$$

with $C_{k-1/2}$ and $D_{k+1/2}$ given, respectively, by

$$\begin{aligned}C_{k-1/2} = \nu^+_{k-1/2}[\, &-(1/12-\beta)b\ \text{mm}\,[1/r^+_k \phi(1/r^+_{k+1}), 1] + (1/2-2\beta)b\,\phi(1/b\, r^+_k) \\ &+ (1/12+\beta)\phi(b/r^+_k) + 1 + (1/12-\beta)b\ \text{mm}\,[1/b\, r^+_k, \phi(r^+_{k-1})] \\ &- (1/2-2\beta)\phi(b\, r^+_{k-1}) - (1/12+\beta)b\,\phi(r^+_{k-1}/b)\,],\end{aligned} \tag{19}$$

$$\begin{aligned}D_{k+1/2} = -\nu^-_{k+1/2}[\, &1 - (1/12+\beta)b\,\phi(r^-_{k+1}/b) - (1/2-2\beta)\phi(b\, r^-_{k+1}) \\ &+ (1/12-\beta)b\ \text{mm}\,[1/b\, r^-_k, \phi(r^-_{k+1})] + (1/12+\beta)\phi(b/r^-_k) \\ &+ (1/2-2\beta)b\,\phi(1/b\, r^-_k) - (1/12-\beta)b\,\text{mm}\,[1/r^-_k \phi(1/b\, r^-_{k-1}), 1]\,].\end{aligned} \tag{20}$$

For convenience, let Υ be the set of all sequences of numbers in (0, 1) with zero limit. We use a letter with hat to represent the sequences in Υ and use the same letter

with subscripts to represent the terms in such a sequence. For $\hat{\varepsilon} \in \Upsilon$, i.e., $\hat{\varepsilon} = \{\varepsilon_l\}_{l=0}^{\infty}$ and $\lim_{l\to\infty} \varepsilon_l = 0$, if u^k is generated by the scheme

$$u^k = H_\varepsilon(u^n_{k-p}, \cdots, u^n_{k+p}; \lambda) = u_k - C_{k-1/2}\Delta u_{k-1/2} + D_{k+1/2}\Delta u_{k+1/2} + \varepsilon_l \tau q(u_k), \tag{21}$$

then we call (21) *the $\hat{\varepsilon}$-scaled form* of the scheme (18)–(20). The HCP of (18)–(20) is given by

$$u^k = \bar{H}(u^n_{k-p}, \cdots, u^n_{k+p}; \lambda) = u_k - C_{k-1/2}\Delta u_{k-1/2} + D_{k+1/2}\Delta u_{k+1/2}. \tag{22}$$

The notions of Yang's extremum paths [10] are very important concepts, which were introduced by Yang in order to track the extrema of the numerical solutions in the computational domain. The definitions of extremum paths and extremum traceability of a scheme are relevant, and they are stated here so that the paper is reasonably self-contained. Also, we will give sufficient conditions that guarantee an $\hat{\varepsilon}$-scaled form (21) to be extremum traceable. Throughout the paper, we refer to [4–6, 10] for the definitions, lemmas, and theorems that we have quoted in the context.

Let a numerical solution u be defined on the set of grid points $X := \{(x_j, t_n) : j \in \mathbb{Z}, n \in \mathbb{Z}^+\}$. A finite set of successive grid points $\{x_q, \cdots, x_r\}$ with $r \geq q$ is said to be *the stencil of a spatial maximum* or simply *an $\overline{E}$-stencil* of u at the time t_n, provided $u^n_q = \cdots = u^n_r, u^n_{q-1} < u^n_q$ and $u^n_{r+1} < u^n_r$. Notions of *$\underline{E}$-stencils* for minima and *E-stencils* for general extrema are defined similarly.

Definition 1 (Definition 2.13 [10]). A non-empty subset of X denoted by $\overline{E}_{t_n,t_m}$, $n \leq m$ is called *a ridge of the numerical solution u from t_n to t_m* if

(i) For all ν, $n \leq \nu \leq m$, the set

$$P_{\overline{E}}(\nu) := \{x_j : (x_j, t_\nu) \in \overline{E}_{t_n,t_m}\} = \{x_{q^\nu}, \cdots, x_{r^\nu}\}$$

is not empty and is an $\overline{E}$-stencil of u at t_ν.

(ii) For all ν, $n \leq \nu \leq m-1$,

$$P_{\overline{E}}(\nu) \cup P_{\overline{E}}(\nu+1) = \{x_j : \min(q^\nu, q^{\nu+1}) \leq j \leq \max(r^\nu, r^{\nu+1})\}.$$

The set $P_{\overline{E}}(\nu)$ is called *the x-projection of* $\overline{E}_{t_n,t_m}$ at t_ν. The value of u along the ridge is denoted by $V_{\overline{E}}(\nu) : V_{\overline{E}}(\nu) = u^\nu_j$ for $q^\nu \leq j \leq r^\nu$.

If, for all ν, $n \leq \nu \leq m$, the $\overline{E}$-stencil in the item (i) of the definition is replaced by an $\underline{E}$-stencil, then the set is called *a trough* of u from t_n to t_m and is denoted by $\underline{E}_{t_n,t_m}$. The related notions $P_{\underline{E}}(\nu)$ and $V_{\underline{E}}(\nu)$ are defined similarly. Ridges and troughs are also called *extremum paths*. When we do not distinguish between ridges and troughs, we use E_{t_n,t_m}, $P_E(\nu)$, and $V_E(\nu)$ for either type. We write

$$E^1_{t_n,t_m} < (\leq) E^2_{t_n,t_m}, \text{ if } \max P_{E^1}(\nu) < (\leq) \max P_{E^2}(\nu) \text{ for } n \leq \nu \leq m.$$

Definition 2 (Definition 2.14 [10]). A scheme is said to be *extremum traceable* if there exists a constant $c \geq 1$ such that for each numerical solution u of the scheme and each integer $N > 0$, there exists a finite or infinite collection of extremum paths $\{E^l_{t_0,t_N}\}_{l=l_1}^{l_2}$ with the following properties:

(i) $\{P_{E^l}(N)\}_{l=l_1}^{l_2}$ is precisely the set of E-stencils of u_j^n at the time t_N arranged in ascending spatial coordinates.

(ii) If $E^l_{t_0,t_N}$ is a ridge (trough), then $V_{E^l}(n)$ is a non-increasing (non-decreasing) function of n.

(iii) Let $P_{E^l}(n) = \{x_{q^l(n)}, \cdots, x_{r^l(n)}\}$ for $1 \leq n \leq N$. If $P_{E^l}(n) \cap P_{E^l}(n+1) = \emptyset$, then

$$| u^n_{q^l(n+1)} - u^n_{r^l(n)} | \leq c \, | V_{E^l}(n+1) - V_{E^l}(n) | \quad \text{when} \quad q^l(n+1) > r^l(n),$$

$$| u^n_{r^l(n+1)} - u^n_{q^l(n)} | \leq c \, | V_{E^l}(n+1) - V_{E^l}(n) | \quad \text{when} \quad q^l(n) > r^l(n+1).$$

(iv) If $l_2 > l_1$, then $E^l_{t_0,t_N} < E^{l+1}_{t_0,t_N}$ for $l_1 \leq l \leq l_2 - 1$.

Following the proof of Theorem 2.3 [5], we can easily obtain the sufficient conditions for the $\hat{\varepsilon}$-scaled form (21) to be extremum traceable.

Theorem 1. *The sufficient conditions for the $\hat{\varepsilon}$-scales form (21) to be extremum traceable are, for sufficiently small ε and $\varepsilon_l < \varepsilon$, the following inequalities:*

$$0 \leq C_{k+1/2}, \quad 0 \leq D_{k+1/2}, \quad 0 \leq C_{k+1/2} + D_{k+1/2} \leq 1, \text{ for all } k \tag{23}$$

hold and there is a positive constant μ such that, if u_k is a space extremum, then

$$\max \{C_{k\pm1/2}, C_{k\pm3/2}, D_{k\pm1/2}, D_{k+3/2}\} \leq \mu/4 < 1/4, \tag{24}$$

where $C_{k+1/2}$ and $D_{k+1/2}$ are given by (19)-(20).

To recast Theorem 1 in terms of the local CFL numbers, we consider a subclass of E-fluxes:

$$g^E(x, y) = \begin{cases} f(x) & \text{if } s \leq x \leq y, \\ f(y) & \text{if } x \leq y \leq s, \end{cases} \tag{25}$$

where s is a sonic point of $f(\cdot)$ ($f'(s) = 0$). It is clear that both Godunov [3] and Engquist–Osher [1] fluxes:

$$g^{God}(u_j, u_{j+1}) = \begin{cases} \min_{u_j \leq w \leq u_{j+1}} f(w) & \text{when} \quad u_j \leq u_{j+1}, \\ \max_{u_j \geq w \geq u_{j+1}} f(w) & \text{when} \quad u_j \geq u_{j+1}, \end{cases} \tag{26}$$

$$g^{EO}(u_j, u_{j+1}) = \int_0^{u_j} \max(f'(w), 0) dw + \int_0^{u_{j+1}} \min(f'(w), 0) dw + f(0), \tag{27}$$

are members of the fluxes defined by (25).

Lemma 1. (See Lemma 2.5 *[6]* for the result of HCPs). *An $\hat{\varepsilon}$-scaled form (21), with the building block given by the member of (25) and $\varepsilon_l < \varepsilon$ for sufficiently small ε, is extremum traceable provided that for all k: $\nu^+_{k+1/2} - \nu^-_{k+1/2} \le 1/3$ and when u_k is an extremum, $\lambda \max_{u_{k-2} \le w \le u_{k+2}} | f'(w) | \le \frac{1}{10}$.*

2 The Convergence of β-Schemes with Source Terms

The following separation property is necessary to apply the extended convergence criterion. Lemma 2 verifies that non-homogeneous β-schemes satisfy this separation property. The proof of Lemma 2, which does not involve the source terms, is the same as that of Lemma 3.2 [6] for the HCPs of (10)–(11).

Assumption 2. The numerical fluxes $g^n_{k+1/2}$, $-\infty < k < \infty$ satisfy

$$g^n_{k+1/2} \ge f(u^n_k) \ge g^n_{k-1/2} \text{ if } u^n_k \ge u^n_{k\pm1}; \quad g^n_{k+1/2} \le f(u^n_k) \le g^n_{k-1/2} \text{ if } u^n_k \le u^n_{k\pm1}.$$

Lemma 2. *The schemes (10)–(11), hence (18)–(20), satisfy Assumption 2.*

Let $f[w; L, R]$ be the linear function interpolating $f(w)$ at $w = L$ and $w = R$. In reference to (21), we denote $\tilde{v}_j = H_\varepsilon(v_{j-p}, \cdots, v_{j+p}; \lambda)$ and $\bar{v}_j = \frac{v_j + \tilde{v}_j}{2}$ for any collection of data $\{v_j\}$. Recall the HCPs (22) of the $\hat{\varepsilon}$-scaled form (21) are TVD [8].

Definition 3 (See Definition 2.20 [10] for the HCPs). For an $\hat{\varepsilon}$-scaled form (21), we call *an ordered pair of numbers $\{L, R\}$ a rarefying pair* if $L < R$ and $f[w; L, R] > f(w)$ when $L < w < R$. We call *a collection of data $\Gamma = \{v_j\}_{j=I-p}^{J+p}$ an ε-rarefying collection* of the $\hat{\varepsilon}$-scaled form (21) to the rarefying pair $\{L, R\}$ if, for $\varepsilon > 0$,

(i) $L = v_I \le v_{I+1} \le \cdots \le v_J = R$;
(ii) $\tilde{v}_I \le \tilde{v}_{I+1} \le \cdots \le \tilde{v}_J$, $|L - \tilde{v}_I| < \varepsilon$, $|R - \tilde{v}_J| < \varepsilon$;
(iii) Either $v_{I-1} \ge v_I$ or $v_I = v_{I+1}$ and either $v_{J+1} \le v_J$ or $v_{J-1} = v_J$.

Clearly, the conditions of (i) and (ii) imply that

$$\bar{v}_I \le \bar{v}_{I+1} \le \cdots \le \bar{v}_J, \ |L - \bar{v}_I| < \varepsilon/2, \text{ and } |R - \bar{v}_J| < \varepsilon/2.$$

We define the piecewise constant function g_Γ associated with the ε-rarefying collection Γ of an $\hat{\varepsilon}$-scaled form (21) as follows:

$$g_\Gamma(w) = g_{j+1/2}[v] \quad \text{for } w \in (\bar{v}_j, \bar{v}_{j+1}), \quad I \le j \le J-1. \tag{28}$$

Definition 4. An ε-rarefying collection $\Gamma = \{v_j\}_{j=I-2}^{J+2}$ of the $\hat{\varepsilon}$-scaled form (21) to the pair $\{L, R\}$ is called *an ε-normal collection*, provided that

$$L = v_{I-2} = v_{I-1} = v_I = v_{I+1} \le \cdots \le v_{J-1} = v_J = v_{J+1} = v_{J+2} = R. \tag{29}$$

Theorem 3 (See Theorem 2.21 [10] , Theorem 3.22 [4]). *A scheme (10)–(11), hence (18)–(20), with extremum traceable $\hat{\varepsilon}$-scaled form (21) converges for convex conservation laws (1) if, for every rarefying pair $\{L, R\}$ and ε-rarefying collection of the $\hat{\varepsilon}$-scaled form (21) to the pair, the quadrature inequality*

$$\int_L^R f[w; L, R]\, dw - \int_{\bar{v}_l}^{\bar{v}_J} g_\Gamma(w)\, dw > \delta \tag{30}$$

holds for some constant $\delta > 0$, provided that ε is sufficiently small.

For the β-schemes concerned in this study, the condition of ε-rarefying collections in Theorem 3 can be weakened by ε-normal collections.

Lemma 3. *A scheme (10)–(11), hence (18)–(20), with extremum traceable $\hat{\varepsilon}$-scaled form (21) converges for convex conservation laws (1) if for each rarefying pair $\{L, R\}$ there is a constant $\delta > 0$ such that the inequality (30) holds for all ε-normal collections of the $\hat{\varepsilon}$-scaled form (21) to the pair $\{L, R\}$.*

Proof. Let $\Lambda = \{\kappa_{P-2}, \cdots, \kappa_{Q+2}\}$ be an arbitrary ε-rarefying collection of the $\hat{\varepsilon}$-scaled form (21) to the pair $\{L, R\}$. Without loss of generality, we assume that $|\varepsilon_l \tau q| < \varepsilon$ for all l. Let

$$S' = \int_{\tilde{\kappa}_P}^{\tilde{\kappa}_Q} g_\Lambda(w)\, dw = \sum_{j=P}^{Q-1} (\bar{\kappa}_{j+1} - \bar{\kappa}_j)\, g_{j+1/2}[\kappa]. \tag{31}$$

by (i) and (iii) of Definition 3, and either κ_P or κ_{P+1} is a minimum. In either case, Assumption 2 and the condition (ii) of Definition 3 imply that

$$\begin{aligned} \varepsilon &> |L - \tilde{\kappa}_P| = |\tilde{\kappa}_P - \kappa_P| \\ &\geq \lambda |g_{P+1/2}[\kappa] - g_{P-1/2}[\kappa]| - |\varepsilon_l \tau q| \geq \lambda |g_{P\pm 1/2}[\kappa] - f(L)| - |\varepsilon_l \tau q|, \end{aligned}$$

or

$$\lambda |g_{P\pm 1/2}[\kappa] - f(L)| \leq \varepsilon + |\varepsilon_l \tau q| < 2\varepsilon. \tag{32}$$

Similarly, we have

$$\varepsilon > |R - \tilde{\kappa}_Q| \geq \lambda |g_{Q\pm 1/2}[\kappa] - f(R)| - |\varepsilon_l \tau q|,$$

or

$$\lambda |g_{Q\pm 1/2}[\kappa] - f(R)| \leq \varepsilon + |\varepsilon_l \tau q| < 2\varepsilon. \tag{33}$$

Next, we construct an ε-normal collection $\Gamma = \{v_j\}_{j=I-2}^{J+2}$ as follows. First, let $I = P - 1$ and $J = Q + 1$ and we also set $v_{I-2} = v_{I-1} = v_I = L$, $v_J = v_{J+1} = v_{J+2} = R$, and $v_j = \kappa_j$ for $I + 1 \le j \le J - 1$. Then, we have

$$g_{I\pm1/2}[v] = f(L),\ g_{J\pm1/2}[v] = f(R),\ \tilde{v}_I = L + \varepsilon_l \tau q(L),\ \text{and}\ \tilde{v}_J = R + \varepsilon_l \tau q(R). \tag{34}$$

Thus, the ε-normality of $\Gamma = \{v_j\}_{j=I-2}^{J+2}$ is justified by the non-decreasing relation of

$$\tilde{v}_I \le \tilde{v}_{I+1} \le \cdots \le \tilde{v}_J.$$

Indeed, we notice that the relationship of $\tilde{v}_{I+3} \le \tilde{v}_{I+4} \le \cdots \le \tilde{v}_{J-4} \le \tilde{v}_{J-3}$ is directly inherited from the condition (ii) of the given ε-rarefying collection of Λ

$$\tilde{\kappa}_{P+2} \le \tilde{\kappa}_{P+3} \le \cdots \le \tilde{\kappa}_{Q-3} \le \tilde{\kappa}_{Q-2}.$$

Also, using the definition of the numerical flux, we can verify that $(f^-_{P-1/2})^{(-1)} = 0$, $(f^+_{P+3/2})^{(1)} = 0$, $(f^+_{P+1/2})^{(0)} = 0$, and $(f^+_{P-1/2})^{(-1)} = 0$, which imply that $g_{P+1/2} = g_{I+3/2}$. Likewise, $(f^-_{Q+1/2})^{(1)} = 0, (f^-_{Q-1/2})^{(0)} = 0, (f^-_{Q-3/2})^{(-1)} = 0$, and $(f^+_{Q+1/2})^{(1)} = 0$ imply that $g_{Q-1/2} = g_{J-3/2}$. Thus, we have $\tilde{v}_{I+2} = \tilde{\kappa}_{P+1}$ and $\tilde{v}_{J-2} = \tilde{\kappa}_{Q-1}$ as well. Therefore, we only need to verify that

$$\tilde{v}_I \le \tilde{v}_{I+1} \le \tilde{v}_{I+2} \quad \text{and} \quad \tilde{v}_{J-2} \le \tilde{v}_{J-1} \le \tilde{v}_J.$$

We will show that $\tilde{v}_I \le \tilde{v}_{I+1}$ and $\tilde{v}_{I+1} \le \tilde{v}_{I+2}$. The proof of $\tilde{v}_{J-2} \le \tilde{v}_{J-1} \le \tilde{v}_J$ is similar, and we omit the details. Notice that the following estimate is the consequence of the definition of Γ and the Assumption 2

$$\begin{aligned}\tilde{v}_{I+1} &= v_{I+1} - \lambda(g_{I+3/2} - g_{I+1/2}) + \varepsilon_l \tau q(v_{I+1}) = v_{I+1} - \lambda(g_{I+3/2} - f(L)) + \varepsilon_l \tau q(v_{I+1}) \\ &\ge v_{I+1} + \varepsilon_l \tau q(v_{I+1}) \ge v_I + \varepsilon_l \tau q(v_I) = \tilde{v}_I.\end{aligned}$$

Also, $\tilde{v}_{I+1} \le \tilde{v}_{I+2}$ follows from the fact that $g_{P+1/2} \le f_P = f_{I+1} = f(L)$, $q' \ge 0$ and $g_{P+1/2} = g_{I+3/2}$. Indeed,

$$\tilde{v}_{I+2} - \tilde{v}_{I+1} = v_{I+2} - v_{I+1} + \varepsilon_l \tau q'(v_{I+2} - v_{I+1}) - \lambda(g_{P+1/2} - f(L)) \ge 0.$$

Secondly, let G be the Lipschitz constant of the numerical flux g, and $K = \max\{|f(L)|, |f(R)|\} + 2G(R - L)$. Denote

$$S = \int_L^R g_\Gamma(w)\, dw = \sum_{j=I}^{J-1} (\bar{v}_{j+1} - \bar{v}_j) g_{j+1/2}[v], \tag{35}$$

and then a priori estimate $|S - S'| \le 4K\varepsilon$ holds. Let δ' be a constant such that for all ε-normal collections of the $\hat{\varepsilon}$-scaled form (21) to the pair $\{L, R\}$ the inequality

(30) holds for $\delta = \delta'$. Thus, for $\delta = \delta'$, the inequality (30) also holds for the ε-normal collection $\Gamma = \{v_j\}_{j=I-2}^{J+2}$. Therefore, for $\delta = \frac{\delta'}{2}$, the inequality (30) holds for all ε-rarefying collection of the $\hat{\varepsilon}$-scaled form (21) to the pair $\{L, R\}$ for $\varepsilon \le \frac{\delta}{4K}$.

It remains to show the a-priori estimate. First, we notice that $\bar{\kappa}_j = \bar{v}_j$ for $P+1 \le j \le Q-2$, Therefore, the terms of the difference

$$S - S' = \sum_{j=I}^{J-1} (\bar{v}_{j+1} - \bar{v}_j) g_{j+1/2}[v] - \sum_{j=P}^{Q-1} (\bar{\kappa}_{j+1} - \bar{\kappa}_j)\, g_{j+1/2}[\kappa]$$

from $j = P+1$ to $j = Q-2$ are all diminished. For the remaining terms, we use the relationship of Λ and Γ and (32)–(34) to yield the following estimates.

$$|\bar{v}_{I+1} - \bar{\kappa}_{I+1}| \le \varepsilon/2 + |\varepsilon_l \tau q|/2 < \varepsilon, \quad |\bar{v}_{J-1} - \bar{\kappa}_{J-1}| \le \varepsilon/2 + |\varepsilon_l \tau q|/2 < \varepsilon, \tag{36}$$

$$|\bar{v}_{I+1} - \bar{v}_I| = \lambda/2 |g_{I+3/2} - f(L)| = \lambda/2 |g_{P+1/2} - f(L)| < \varepsilon, \tag{37}$$

$$|\bar{v}_J - \bar{v}_{J-1}| = \lambda/2 |f(R) - g_{J-3/2}| = \lambda/2 |f(R) - g_{Q-1/2}| < \varepsilon. \tag{38}$$

Finally, using the fact that $\bar{v}_{I+2} = \bar{\kappa}_{P+1}$, $\bar{v}_{J-2} = \bar{\kappa}_{Q-1}$, $g_{I+3/2}[v] = g_{P+1/2}[\kappa]$, $g_{J-3/2}[v] = g_{Q-1/2}[\kappa]$, and (36)–(38), we have the desired estimate as follows.

$$\begin{aligned}
|S - S'| &= |(\bar{v}_{I+1} - \bar{v}_I) g_{I+1/2}[v] + (\bar{v}_J - \bar{v}_{J-1}) g_{J-1/2}[v] + (\bar{v}_{I+2} - \bar{v}_{I+1}) g_{I+3/2}[v] \\
&\quad - (\bar{\kappa}_{P+1} - \bar{\kappa}_P) g_{P+1/2}[\kappa] + (\bar{v}_{J-1} - \bar{v}_{J-2}) g_{J-3/2}[v] - (\bar{\kappa}_Q - \bar{\kappa}_{Q-1}) g_{Q-1/2}[\kappa]| \\
&\le |\bar{v}_{I+1} - \bar{v}_I| |g_{I+1/2}[v]| + |\bar{v}_J - \bar{v}_{J-1}| |g_{J-1/2}[v]| \\
&\quad + |\bar{v}_{I+1} - \bar{\kappa}_P| |g_{I+3/2}[v]| + |\bar{v}_{J-1} - \bar{\kappa}_Q| |g_{Q-1/2}[\kappa]| \\
&< (\varepsilon + \varepsilon + \varepsilon + \varepsilon) K = 4K\varepsilon,
\end{aligned}$$

and the proof is completed.

For an ε-normal collection $\Gamma = \{v_j\}_{j=I-2}^{J+2}$, we denote the vertex $(v_j, f(v_j))$ by V_j and the area of convex polygon $V_{j_1} V_{j_2} \cdots V_{j_r}$ by $S_{j_1,\ldots,j_r}$. Let $\sigma_\Gamma = \max_{I-2 \le j \le J+2} |\nu^{\pm}_{j\pm 1/2}|$, and let

$$\alpha_j = \begin{cases} 0.5 & \text{if } \Delta_+ v_{j-2} = \Delta_+ v_{j+1} = 0, \\ 1 & \text{otherwise.} \end{cases}$$

When the building blocks of the schemes (10)–(11), hence (18)–(20), are the E-schemes with the fluxes defined by (25), we have the following very important inequality (39), which will enable us to prove the mail result of Theorem 4. The proof of Lemma 4 is similar to the one for the HCPs of (10)–(11) [6] and will be omitted. The Lemma 5 is Yang's original result.

Lemma 4. *Let $\Gamma = \{v_j\}_{j=I-2}^{J+2}$ be an ε-normal collection of the $\hat{\varepsilon}$-scaled form (21) to a rarefying pair $\{L, R\}$. Then, the numerical solutions of the $\hat{\varepsilon}$-scaled form (21) for convex conservation laws (1) satisfy, for sufficiently small ε and σ_Γ, the inequality*

$$\int_L^R (f[w; L, R] - g_\Gamma)dw \geq S_{I,I+1,\ldots,J} - \sum_{j=I+1}^{J-1} \alpha_j S_{j-1,j,j+1}. \tag{39}$$

Lemma 5. (Lemma 3.7 [10]) *For $I < i < J-1$, we have*

$$S_{I,I+1,\ldots,J} - \sum_{j=I+1}^{J-1} S_{j-1,j,j+1} \geq S_{I,i,i+1,J} - (S_{I,i,i+1} + S_{i,i+1,J}).$$

Let $\sigma = \lambda \max_w |f'(w)|$. For the non-homogeneous β-schemes when $m = 2$, equipped with Lemmas 1, 4, and 5, we have obtained the following entropy convergence result.

Theorem 4. *A scheme of the form (10)–(11) converges for convex conservation laws (1) if, $g^E(\cdot, \cdot)$ is a numerical flux given by (25) and, σ and ε are sufficiently small.*

Proof. For sufficiently small σ and ε, by Lemma 1, the $\hat{\varepsilon}$-scaled form (21) is extremum traceable. Now, for each ε-normal collection $\Gamma = \{v_i\}_{i=I-2}^{J+2}$ of (21) to a rarefying pair $\{L, R\}$, we claim that the following inequality holds

$$\int_{\bar{v}_I}^{\bar{v}_J} g_\Gamma(w)dw \leq \int_L^R g_\Gamma(w)dw + \varepsilon. \tag{40}$$

Indeed, first of all, we have $\bar{v}_I = L + (\varepsilon_l \tau q(L))/2$ and $\bar{v}_J = R + (\varepsilon_l \tau q(R))/2$. Also, recall that $q'(w) \geq 0$ and $g_\Gamma(w) = g_{j+\frac{1}{2}}[v]$ for $w \in (\bar{v}_j, \bar{v}_{j+1})$ and $I \leq j \leq J-1$.

Case 1. If $q(L) \geq 0$, then $q(R) \geq 0$ as well. Let c be a constant such that $|g_\Gamma(w)| \leq c$, for $w \in (R, \bar{v}_J)$, and we set $g_\Gamma(w) = -c$, when $w \in (L, \bar{v}_I)$. Then, we have

$$\int_{\bar{v}_I}^{\bar{v}_J} g_\Gamma(w)dw = \{\int_{\bar{v}_I}^{L} + \int_L^R + \int_R^{\bar{v}_J}\} g_\Gamma(w)dw \leq c\varepsilon_l \tau(q(L) + q(R))/2 + \int_L^R g_\Gamma(w)dw.$$

Case 2. If $q(L) \leq 0$, and $q(R) \geq 0$, we let c be a constant such that $|g_\Gamma(w)| \leq c$, for $w \in (R, \bar{v}_J) \cup (\bar{v}_I, L)$. Now, we have

$$\int_{\bar{v}_I}^{\bar{v}_J} g_\Gamma(w)dw = \{\int_{\bar{v}_I}^{L} + \int_L^R + \int_R^{\bar{v}_J}\} g_\Gamma(w)dw \leq c\varepsilon_l \tau(-q(L) + q(R))/2 + \int_L^R g_\Gamma(w)dw.$$

Case 3. If $q(L) \leq 0$, and $q(R) \leq 0$, we let c be a constant such that $|g_\Gamma(w)| \leq c$, for $w \in (\bar{v}_I, L)$ and set $g_\Gamma(w) = -c$, when $w \in (\bar{v}_J, R)$. We obtain

$$\int_{\bar{v}_I}^{\bar{v}_J} g_\Gamma(w)dw = \{\int_{\bar{v}_I}^{L} + \int_L^R + \int_R^{\bar{v}_J}\} g_\Gamma(w)dw \leq c\varepsilon_l \tau(-q(L) - q(R))/2 + \int_L^R g_\Gamma(w)dw.$$

In all cases, without loss of generality, for the given $\varepsilon > 0$ we let $c\varepsilon_l \tau(|q(L)| + |q(R)|)/2 < \varepsilon$ for all l. Thus, as claimed, the inequality (40) holds. Next, we set

$$d_1(\Gamma) = \max_{I \le i \le J} \min(v_i - L, R - v_i).$$

Since $J - I$ is finite, $d_1(\Gamma) = \min(v_j - L, R - v_j)$ for some j between I and J. We then let

$$d_2(\Gamma) = \max_{I \le i \le J, i \ne j} \min(v_i - L, R - v_i).$$

We also have $d_2(\Gamma) = \min(v_k - L, R - v_k)$ for some $k \ne j$ between I and J. Clearly, we can choose j and k so that $|j - k| = 1$.

To complete the proof, we argue by contradiction. Thus, we assume that for certain convex f, the scheme (10)–(11), hence (18)–(20), does not converge. By Lemma 3 and (40), there is a rarefying pair $\{L, R\}$ such that for each $\delta > 0$, $\delta' = \delta/2$, and $\varepsilon = \delta/2$, there is a ε-normal collection $\Gamma = \{v_j\}_{j=I-2}^{J+2}$ of the $\hat{\varepsilon}$-scaled form (21) to the pair that satisfies

$$\int_L^R \{f[w; L, R] - g_\Gamma(w)\} dw \le \delta' + \varepsilon = \delta.$$

It follows that there is a sequence of ε-normal collections $\{\Gamma_\nu\}_{\nu=1}^{\infty}$, where $\Gamma_\nu = \{v_j^\nu\}_{j=I^\nu-2}^{J^\nu+2}$ such that

$$\lim_{\nu \to \infty} \int_L^R \{f[w; L, R] - g_{\Gamma_\nu}(w)\} \le 0. \tag{41}$$

The following three cases exhaust all possibilities.

Case 1. $\limsup_{\nu\to\infty} d_2(\Gamma_\nu) > 0$. Set $\rho = 1/2 \limsup_{\nu\to\infty} d_2(\Gamma_\nu)$. Then, there is a subsequence of the ε-normal collections, still denoted by $\{\Gamma_\nu\}_{\nu=1}^{\infty}$, and a corresponding sequence of integers $\{i(\nu)\}_{\nu=1}^{\infty}$ such that

$$L + \rho \le v_{i(\nu)}^\nu \le v_{i(\nu)+1}^\nu \le R - \rho,$$

and $\sup_\nu \sigma_{\Gamma_\nu} \le \sigma$. For simplicity, we fix a ν and drop it from the notation. Set $\gamma = f[\frac{L+R}{2}; L, R] - f(\frac{L+R}{2})$. It is a positive constant since $\{L, R\}$ is a rarefying pair. Applying Lemmas 4 and 5, we have

$$\begin{aligned} &\int_L^R \{f[w; L, R] - g_{\Gamma_\nu}(w)\} dw \ge S_{I,i,i+1,J} - (S_{I,i,i+1} + S_{i,i+1,J}) \\ &= 1/2\{(v_i - v_I)(f[v_{i+1}; L, R] - f(v_{i+1})) + (v_J - v_{i+1})(f[v_i; L, R] - f(v_i))\} > \eta, \end{aligned} \tag{42}$$

if $\eta = 2\rho^2\gamma/(R - L)$. This contradicts (41).

Case 2. $\limsup_{\nu\to\infty} d_1(\Gamma_\nu) > \limsup_{\nu\to\infty} d_2(\Gamma_\nu) = 0$. Set $\rho = 1/2 \limsup_{\nu\to\infty} d_1(\Gamma_\nu)$. Then, there is a subsequence of the ε-normal collections, still denoted by $\{\Gamma_\nu\}_{\nu=1}^{\infty}$, and a corresponding sequence of integers $\{i^\nu\}_{\nu=1}^{\infty}$ such that $\lim_{\nu\to\infty} v_{i^\nu-1}^\nu = L$, $\lim_{\nu\to\infty} v_{i^\nu+1}^\nu = R$, and $\lim_{\nu\to\infty} v_{i^\nu}^\nu = v \in [L + \rho, R - \rho]$. We then have

$$\int_L^R (f[w; L, R] - g_{\Gamma_\nu}(w))dw \to \int_L^R (f[w; L, R] - g_\Gamma(w))dw,$$

where Γ is the following ε-normal collection: $I = 0$, $J = 4$, $v_{-2} = v_{-1} = v_0 = v_1 = L$, $v_2 = v$, and $v_3 = v_4 = v_5 = v_6 = R$. By Lemma 4, we have

$$\int_L^R (f[w; L, R] - g_\Gamma(w))dw \geq S_{1,2,3} - \alpha_2 S_{1,2,3} = 1/2 S_{1,2,3} > 0$$

for $\alpha_2 = 1/2$ since $\Delta_+ v_0 = \Delta_+ v_3 = 0$. This contradicts (41).

Case 3. $\limsup_{\nu\to\infty} d_1(\Gamma_\nu) = 0$. Then, there exists a sequence of integers $\{i^\nu\}$ with $I^\nu + 1 \leq i^\nu < J^\nu - 1$ such that $\lim_{\nu\to\infty} v_{i^\nu}^\nu = L$, $\lim_{\nu\to\infty} v_{i^\nu+1}^\nu = R$. We then have

$$\int_L^R (f[w; L, R] - g_{\Gamma_\nu}(w))dw \to \int_L^R (f[w; L, R] - g_\Gamma(w))dw,$$

where Γ is the following ε-normal collection: $I = 0$, $J = 3$, $v_{-2} = v_{-1} = v_0 = v_1 = L$, $v_2 = v_3 = v_4 = v_5 = R$. In this case, the numerical flux $g_\Gamma(w)$ becomes E-flux $g^E(L, R)$. Hence, we have

$$\int_L^R (f[w; L, R] - g_\Gamma(w))dw \geq \int_L^R (f[w; L, R] - f(w))dw.$$

The right-hand side of the inequality is a positive constant since $\{L, R\}$ is a rarefying pair. This contradicts (41) again. We have thus completed the proof of Theorem 4.

References

1. B. Engquist, S. Osher, Stable and entropy satisfying approximations for transonic flow calculations. Math. Comp. **34**, 45–75 (1980)
2. U.S. Fjordholm, S. Mishra, E. Tadmor, Arbitrarily high-order accurate entropy stable essentially nonoscillatory schemes for systems of conservation laws. SIAM J. Numer. Anal. **50**(2), 544–573 (2012)
3. S.K. Godunov, Finite-difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. Mat. Sbornik **47**, 271–306 (1959)
4. N. Jiang, *On Wavewise Entropy Inequalities for High-Resolution Schemes with Source Terms II: The Fully-Discrete Case*, submitted
5. N. Jiang, On the convergence of fully-discrete high-resolution schemes with van leer's flux limiter for conservation laws. Methods Appl. Anal. **16**(3), 403–422 (2009)
6. N. Jiang, On the convergence β-schemes. Int. J. Numer. Anal. Model. **14**(1), 103–125 (2017)
7. S. Osher, S. Chakravarthy, High resolution schemes and entropy condition. SIAM J. Numer. Anal. **21**, 955–984 (1984)
8. S. Osher, S. Chakravarthy, Very high order accurate TVD schemes. J. Oscil. Theory Comput. Methods Compens. Compact. 229–274 (1986)

9. Y. Xiangyu Hu, N.A. Adams, C-W. Shu, Positivity-preserving method for high-order conservative schemes solving compressible Euler equations, JCP **242**, 169–180 (2013)
10. H. Yang, On wavewise entropy inequalities for high resolution schemes ii: fully discrete MUSCL schemes with exact evolution in small time. SIAM. J. Numer. Anal. **36**(1), 1–31 (1999)
11. X. Zhengfu, Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: one-dimensional scalar problem. J. Math. Comp. **83**, 2213–2238 (2014)

Existence of Undercompressive Shock Wave Solutions to the Euler Equations

Buğra Kabil

Abstract The sharp-interface dynamics of compressible inviscid liquid–vapor flows with constant temperature can be described by the isothermal Euler equations using a non-monotone pressure function. The motion of the discontinuous phase boundaries is constrained besides mass conservation by the dynamical Young–Laplace law and the prescribed entropy dissipation rate. We consider the initial value problem for a two-phase configuration in multiple space dimensions, such that the smooth bulk state data are separated by a subsonic phase boundary which can be understood as a non-Laxian, undercompressive shock wave. It is proven that the associated free boundary problem admits a piecewise classical solution for short times. This strongly nonlinear problem will be formulated as an abstract combination of a hyperbolic initial boundary value problem for the hydromechanical unknowns and a parabolic evolution equation for the front position. By an iteration scheme (local-in-time), the well-posedness of the nonlinear problem is established.

Keywords Undercompressive shocks · Euler equations
Compressible liquid-vapor dynamics

1 Introduction

Compressible liquid–vapor dynamics is of major interest for the understanding of many natural processes and technical applications. In this paper, we are interested in the spatial dynamics of phase fronts as single parts of droplets or bubbles. In spite of its importance, the analytical treatment of associated mathematical models—in particular in multiple space dimensions—is still in the beginnings. Here, the focus

B. Kabil (✉)
Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Stuttgart, Germany
e-mail: Bugra.Kabil@mathematik.uni-stuttgart.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_7

will be on a sharp-interface model where the dynamics in the bulk phases is governed by an ideal compressible fluid whose temperature is supposed to be constant. Therefore, as the mathematical model we consider the isothermal Euler equations with a non-monotone Van der Waals type pressure function. The non-monotone pressure relation allows to define a liquid and a vapor phase in a straightforward way. The phase boundaries are moving sharp interfaces separating two bulk domains. The major modeling issue remains to fix the coupling conditions across the phase boundary. Basic kinematic requirements prescribe the conservation of mass and the balance of momentum, the latter expressed through the Young–Laplace law, which brings the interface curvature into the model. Close to equilibrium, it is expected that the interface moves as a subsonic transition. Mathematical speaking, this corresponds to a nonclassical undercompressive shock wave such that a further condition should be applied to ensure well-posedness. It has been suggested (see [1, 6, 14]) to prescribe the entropy dissipation across the interface in the form of an additional algebraic condition, called kinetic relation. The complete model will be presented in all necessary details in Sect. 2.

The planar case where surface tension can be neglected is by now quite well understood. The stability of phase boundaries is a consequence of the work in [4], which in fact covers a much wider situation. There are many results on the existence and stability of weak solutions, and we refer to, for example, [5, 7, 8] for Riemann problems and [10] for a general initial value problem. First results on the persistence of shock waves in arbitrary space dimension are due to [11] for classical Laxian shock waves, which are completely constrained by the (homogeneous) Rankine–Hugoniot conditions (see also [12]). Afterward, undercompressive shock waves have been analyzed in the general framework of systems of hyperbolic conservation laws in [4, 11, 16]. For the Euler equations with van der Waals pressure, it was shown that there are special solutions composed of a single Laxian shock front and one subsonic phase boundary [15, 17]. Entropy dissipation and curvature effects have been taken into account in [7] in the sense of energy estimates for the linearized system.

To our knowledge, the multidimensional well-posedness for the evolution that takes into account surface tension and general kinetic relations has not been established. In Sect. 2, the complete free boundary problem is introduced and formulated in a way that allows the analytical treatment. The main result is given with Theorem 2.1, which gives local-in-time well-posedness of a classical solution of the free boundary value problem. Moreover, the obtained solution satisfies the second law of thermodynamics, which accounts for the interfacial energy contributions of surface tension. The proof relies on the method of successive iterations in an appropriate functional setting. The essential ingredient is are uniform energy estimates and well-posedness statements for non-homogeneous linearized versions of the free boundary value problem with variable coefficients which was studied in [7, 8]. This work is based on [9] where more details can be found.

2 The Mathematical Model and the Main Result

The dynamics of a compressible, isothermal, and inviscid fluid in $d \geq 1$ space dimensions is described by the Euler equations

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0\,, \tag{1}$$

$$\partial_t (\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \nabla p(\rho) = 0\,, \tag{2}$$

where $\mathbf{u} = \mathbf{u}(x,t) \in \mathbb{R}^d$, $\rho = \rho(x,t) > 0$ are the unknown velocity and density depending on the space variable $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and time $t > 0$. In this system, the function $p = p(\rho)$ denotes a non-monotone pressure function of van der Waals type such that the fluid occurs in liquid and vapor states.

Definition 1. A pressure function $p \in C^\infty([0, \rho^*))$ for some $\rho^* > 0$ is called a **van der Waals pressure function** if there are constants $l^* > v^* > 0$ such that

$$\begin{cases} p'(\rho) > 0, & \text{if } \ 0 < \rho < v^* \qquad \text{(vapor states)}, \\ p'(\rho) < 0, & \text{if } \ v^* < \rho < l^* \qquad \text{(spinodal states)}, \\ p'(\rho) > 0, & \text{if } \ l^* < \rho < \rho^* \qquad \text{(liquid states)}. \end{cases} \tag{3}$$

Now, we consider a van der Waals fluid which is separated by a sharp interface $\Sigma(t)$ in liquid and vapor domains. Let us denote the vapor domain by $V_+(t)$, the liquid domain by $V_-(t)$ and the separating unknown interface by $\Sigma(t)$. The normal vector $\mathbf{n} = \mathbf{n}(x,t)$ to $\Sigma(t)$ is oriented such that it points into the vapor bulk domain; see Fig. 1.

The motion of this fluid is given by piecewise smooth functions

$$\rho : V_+(t) \cup V_-(t) \subset \mathbb{R}^d \times \mathbb{R}^+ \longrightarrow \mathbb{R}^+, \quad \mathbf{u} : V_+(t) \cup V_-(t) \subset \mathbb{R}^d \times \mathbb{R}^+ \longrightarrow \mathbb{R}^d$$

with

$$\rho\big|_{V_\pm(t)} = \rho^\pm, \quad \mathbf{u}\big|_{V_\pm(t)} = \mathbf{u}^\pm$$

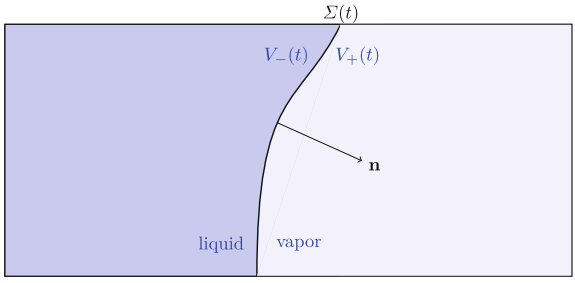


Fig. 1 Liquid–vapor interface $\Sigma(t)$ and bulk domains $V_\pm(t)$

for smooth functions $(\rho^+, \mathbf{u}^+)$ and $(\rho^-, \mathbf{u}^-)$. These satisfy the Euler equations in each domain $V_\pm(t)$ in the classical sense, that is

$$\partial_t \rho^\pm + \nabla \cdot (\rho^\pm \mathbf{u}^\pm) = 0 \quad \text{in} \quad V_\pm(t), \tag{E1}$$

$$\partial_t (\rho^\pm \mathbf{u}^\pm) + \nabla \cdot (\rho^\pm \mathbf{u}^\pm \otimes \mathbf{u}^\pm) + \nabla p(\rho^\pm) = 0 \quad \text{in} \quad V_\pm(t). \tag{E2}$$

Since the fluid occurs in two different phases, we have jumps along the interface $\Sigma(t)$, which is driven as a free boundary by conditions that express conservation of mass, balance of momentum, and the entropy dissipation at the interface. These are given by modified Rankine–Hugoniot jump conditions (see (J1)–(J2) below) and a kinetic relation (see (J3) below). The jump condition describing the conservation of momentum is a dynamical version of the Young–Laplace law for momentum and contains a non-homogeneous term. Precisely, we consider the jump conditions

$$[\rho(\mathbf{u} \cdot \mathbf{n} - \sigma)] = 0, \tag{J1}$$

$$[\rho(\mathbf{u} \cdot \mathbf{n} - \sigma)\mathbf{u} + p\mathbf{n}] = (d-1)\kappa \mathrm{s}\mathbf{n}, \tag{J2}$$

$$\left[g(\rho) + \frac{j^2}{2\rho^2} \right] = -\mathrm{B}j, \tag{J3}$$

where the brackets express the jump of some quantity f across the interface, i.e.,

$$[f] = \lim_{\epsilon \searrow 0} (f(x + \epsilon \mathbf{n}) - f(x - \epsilon \mathbf{n}))$$

for any $x \in \Sigma(t)$, $\mathrm{s} > 0$ is the surface tension, κ the mean curvature, $\mathrm{B} > 0$ the interfacial mobility constant, $\mathbf{n} \in \mathbb{R}^d$ the unit normal vector to the moving interface, $\sigma \in \mathbb{R}$ the normal speed of propagation of the interface in x. The mass transfer flux j is defined by

$$j := \lim_{\epsilon \searrow 0} (\rho(x - \epsilon \mathbf{n})(\mathbf{u}(x - \epsilon \mathbf{n}) \cdot \mathbf{n} - \sigma)) .$$

The chemical potential g is determined via $g'(\rho) = \rho^{-1} p'(\rho)$. Multiplying the kinetic relation (J3) by the mass transfer flux j and using (J1)–(J2) implies

$$[(\mathbf{u} \cdot \mathbf{n} - \sigma) E(\rho, \mathbf{u})] + [(\mathbf{u} \cdot \mathbf{n} - \sigma) p(\rho)] - \sigma j [\mathbf{u} \cdot \mathbf{n}] = -\mathrm{B}j^2,$$

where $E(\rho, \mathbf{u}) = (\rho \mathbf{u}^2)/2 + \rho\psi(1/\rho)$ is the specific energy with the Helmholtz energy ψ determined via $\psi'(1/\rho) = -p(\rho)$ and related to the chemical potential $g(\rho)$ through $p(\rho) = \rho g(\rho) + \rho\psi(1/\rho)$. This implies an exact description of entropy dissipation Ω at the interface through

$$\Omega := -\sigma \left([E(\rho, \mathbf{u})] + (d-1)\kappa \mathrm{s} \right) + [(E(\rho, \mathbf{u}) + p(\rho))\mathbf{u} \cdot \mathbf{n}] = -\mathrm{B}j^2 \leq 0 , \tag{4}$$

e.g, [13] for the derivation of (4). Inequality (4) shows that the kinetic relation (J3) is compatible with the second law of thermodynamics. Note that smooth solutions to (J1)–(J3) satisfy the extra balance law

$$\partial_t E(\rho, \mathbf{u}) + \nabla \cdot (E(\rho, \mathbf{u})\mathbf{u} + p(\rho)\mathbf{u}) = 0\,.$$

Further, we assume that the interface $\Sigma(t)$ can be represented as the graph of the front function $X \in C^2\left([0,\infty) \times \mathbb{R}^{d-1}\right)$ through

$$\Sigma(t) = \left\{x = (x_1, \ldots, x_d) \,\middle|\, x_d = X(x_1, \ldots, x_{d-1}, t)\right\}.$$

Then, the geometrical quantities in (J1)–(J3) in terms of X are written as

$$\mathbf{n} = \frac{1}{\sqrt{1 + \|\check{\nabla} X\|^2}} \left(-\check{\nabla} X, 1\right)^{\mathsf{T}}, \quad \sigma = \frac{\partial_t X}{\sqrt{1 + \|\check{\nabla} X\|^2}},$$

and

$$\kappa = \frac{1}{d-1}\check{\nabla} \cdot \left(\frac{\check{\nabla} X}{\sqrt{1 + \|\check{\nabla} X\|^2}}\right),$$

where we used $\check{\nabla} = (\partial_{x_1}, \ldots, \partial_{x_{d-1}})^{\mathsf{T}}$. We collect the first spatial components in $y = (x_1, \ldots, x_{d-1})^{\mathsf{T}}$.

Consider now the hyperplane $\left\{x \in \mathbb{R}^d \,\middle|\, x_d = 0\right\}$ and decompose the velocity $\mathbf{u}$ as $\mathbf{u} = (\mathbf{v}, u)$, where $\mathbf{v}$ denotes the tangential part and u its normal part with respect to the hyperplane. With these notations, all jump conditions (J1)–(J3) can be expressed in terms of X. We say that solutions $(\rho^+, \mathbf{u}^+)$ and $(\rho^-, \mathbf{u}^-)$ of the Euler equations (E1)–(E2) and a solution X satisfy the boundary conditions if

$$\left[\rho\,(u - \partial_t X - \mathbf{v} \cdot \check{\nabla} X)\right] = 0, \tag{5}$$

$$\left[\rho\,(u - \partial_t X - \mathbf{v} \cdot \check{\nabla} X)\mathbf{v} - p\check{\nabla} X\right] = 0, \tag{6}$$

$$\left[\rho\,(u - \partial_t X - \mathbf{v} \cdot \check{\nabla} X)u + p\right] = \mathrm{s}\check{\nabla} \cdot \check{\nabla} X, \tag{7}$$

$$\left[\left(1 + \|\check{\nabla} X\|^2\right) g + \tfrac{1}{2}(u - \partial_t X - \mathbf{v} \cdot \check{\nabla} X)^2\right] = -\mathrm{B}\,j\left(1 + \|\check{\nabla} X\|^2\right). \tag{8}$$

Altogether, we consider the Euler equations (E1)–(E2) with the boundary conditions (5)–(8) and the initial data

$$(\rho^\pm, \mathbf{u}^\pm)(x, 0) = (\rho_0^\pm, \mathbf{u}_0^\pm) \quad \text{and} \quad X(y, 0) = X_0\,. \tag{A1}$$

To state our results in a convenient way, we rewrite the Euler equations (E1)–(E2) by using $\mathbf{U}^\pm := (\rho^\pm, \mathbf{u}^\pm)$ in the quasilinear form

$$\mathbf{A}_0^{\pm}(\mathbf{U}^{\pm})\partial_t \mathbf{U}^{\pm} + \sum_{j=1}^{d} \mathbf{A}_j^{\pm}(\mathbf{U}^{\pm})\,\partial_j \mathbf{U}^{\pm} = 0 \qquad \text{in} \quad V_{\pm}(t), \tag{9}$$

for given matrices $\mathbf{A}_0(\mathbf{U}^{\pm}), \ldots, \mathbf{A}_d(\mathbf{U}^{\pm})$. The boundary conditions defined in (5)–(8) can be summarized with $\check{\Delta} = \partial_1^2 + \cdots + \partial_{d-1}^2$ in the form

$$\mathbf{b}_0(\mathbf{U}^{\pm}, X)\partial_t X + \sum_{j=1}^{d-1} \mathbf{b}_j(\mathbf{U}^{\pm}, X)\partial_j X + \mathbf{b}_{\mathrm{s}}\check{\Delta} X + \mathbf{M}(\mathbf{U}^{\pm}, X)\mathbf{U}^{\pm} = 0\,, \tag{10}$$

for given vectors $\mathbf{b}_0(\mathbf{U}^{\pm}, X)$, ..., $\mathbf{b}_{d-1}(\mathbf{U}^{\pm}, X)$, $\mathbf{b}_{\mathrm{s}}$ and a given matrix $\mathbf{M}(\mathbf{U}^{\pm}, X)$.

2.1 The Initial Boundary Value Problem

We reformulate the initial value problem (9) with the boundary conditions (10) as an initial boundary value problem by transforming the solution in the last component. Plugging

$$\underline{\mathbf{U}}^{\pm}(t, y, z) := \mathbf{U}^{\pm}(t, y, \pm z + X(t, y))$$

in (9), we obtain with $\partial_0 := \partial_t$ the equivalent equations

$$\sum_{j=0}^{d-1} \mathbf{A}_j^{\pm}(\underline{\mathbf{U}}^{\pm})\,\partial_j \underline{\mathbf{U}}^{\pm} \pm \mathbf{A}_d^{\pm}(\underline{\mathbf{U}}^{\pm})\partial_z \underline{\mathbf{U}}^{\pm} \mp \sum_{j=0}^{d-1} \partial_j X\, \mathbf{A}_j^{\pm}(\underline{\mathbf{U}}^{\pm})\,\partial_z \underline{\mathbf{U}}^{\pm} = 0 \tag{11}$$

for $(t, y, z) \in \mathbb{R}^+ \times \mathbb{R}^{d-1} \times \mathbb{R}^+$. We define $\mathbf{U} := (\underline{\mathbf{U}}^-, \underline{\mathbf{U}}^+)$,

$$\mathbf{A}_j(\mathbf{U}) := \begin{pmatrix} \mathbf{A}_j^-(\underline{\mathbf{U}}^-) & 0 \\ 0 & \mathbf{A}_j^+(\underline{\mathbf{U}}^+) \end{pmatrix}$$

and

$$\mathbf{A}_z(\mathbf{U}) := \begin{pmatrix} -\mathbf{A}_d^-(\underline{\mathbf{U}}^-) & 0 \\ 0 & \mathbf{A}_d^+(\underline{\mathbf{U}}^+) \end{pmatrix} + \sum_{j=0}^{d-1} \partial_j X \begin{pmatrix} -\mathbf{A}_j^-(\underline{\mathbf{U}}^-) & 0 \\ 0 & \mathbf{A}_j^+(\underline{\mathbf{U}}^+) \end{pmatrix}.$$

Altogether, we consider for $t > 0$ and $y \in \mathbb{R}^{d-1}$ the nonlinear system

$$\begin{cases} \sum_{j=0}^{d-1} \mathbf{A}_j(\mathbf{U})\partial_j \mathbf{U} + \mathbf{A}_z(\mathbf{U})\partial_z \mathbf{U} = 0, & \text{for} \quad z > 0, \\ \sum_{j=0}^{d-1} \mathbf{b}_j(\mathbf{U}, X)\partial_j X + \mathbf{b}_{\mathrm{s}}\check{\Delta} X + \mathbf{M}(\mathbf{U}, X)\mathbf{U} = 0, & \text{for} \quad z = 0, \end{cases} \tag{NL}$$

with initial data $\mathbf{U}(t=0)=\mathbf{U}_0$ and $X(t=0)=X_0$. Note that the associated vectors $\mathbf{b}_0(\mathbf{U}, X), ..., \mathbf{b}_{d-1}(\mathbf{U}, X), \mathbf{b}_s$ and the matrix $\mathbf{M}(\mathbf{U}, X)$ are given explicitly.

The initial data should satisfy some conditions to be compatible with the jump conditions. Let $s > \frac{d+3}{2}$. Then initial data $\mathbf{U}_0 \in H^s(\mathbb{R}^{d-1}\times\mathbb{R}^+)$ and $X_0 \in H^{s+3/2}(\mathbb{R}^{d-1})$ are said to satisfy the first compatibility condition if some functions $(\hat{\mathbf{U}}, \hat{X}) \in H^{s+1}(\mathbb{R}^d\times\mathbb{R}^+)\times\left(H^{s+1}(\mathbb{R}^d)\cap L^2(\mathbb{R}, H^{s+2}(\mathbb{R}^{d-1}))\right)$ with trace on $t=0$ exist such that

$$\mathbf{b}_0(\mathbf{U}_0, X_0)X_1 = -\sum_{j=1}^{d-1}\mathbf{b}_j(\mathbf{U}_0, X_0)\partial_j X_0 + \mathbf{b}_s\check{\Delta}X_0 + \mathbf{M}(\mathbf{U}_0, X_0)\mathbf{U}_0 \tag{12}$$

and

$$\mathbf{A}_0(\mathbf{U}_0)\mathbf{U}_1 = -\sum_{j=1}^{d-1}\mathbf{A}_j(\mathbf{U}_0)\partial_j\mathbf{U}_0 - \mathbf{A}_z(\mathbf{U}_0)\partial_z\mathbf{U}_0 \tag{13}$$

with $X_1 := \partial_t\hat{X}\big|_{t=0}$ and $\mathbf{U}_1 := \partial_t\hat{\mathbf{U}}\big|_{t=0}$ hold.

Definition 2. Let $\mathbf{U}_0 \in H^s(\mathbb{R}^{d-1}\times\mathbb{R}^+)$ and $X_0 \in H^{s+3/2}(\mathbb{R}^{d-1})$ satisfy the first compatibility condition with some functions $\hat{\mathbf{U}} \in H^{s+1}(\mathbb{R}^d\times\mathbb{R}^+)$ and $\hat{X} \in \left(H^{s+1}(\mathbb{R}^d)\cap L^2(\mathbb{R}, H^{s+2}(\mathbb{R}^{d-1}))\right)$. The initial data $(\mathbf{U}_0, X_0)$ are said to be **compatible to order** $s-1$ if the sequence

$$(\mathbf{U}_k, X_k) := \left(\partial_t^k\hat{\mathbf{U}}\big|_{t=0}, \partial_t^k\hat{X}\big|_{t=0}\right),$$

with

$$(\mathbf{U}_k, X_k) \in H^{s-k}(\mathbb{R}^{d-1}\times\mathbb{R}^+)\times H^{s+3/2-k}(\mathbb{R}^{d-1})$$

inductively defined as in (12)–(13), exists for $k=1, ..., s-1$.

The compatibility conditions will allow us to construct approximative solutions which will be used in the analysis of the iteration scheme. The resulting sequence of solutions of the iteration scheme will be bounded in H^s and will converge in L^2. The solution itself will be given in terms of the space $CH^s((0,T)\times\mathbb{R}^{d-1}\times\mathbb{R}^+)$ on a given time interval $[0,T]$ whose definition is given next.

Definition 3. By $CH^s((0,T)\times\mathbb{R}^{d-1}\times\mathbb{R}^+)$, we denote the space of functions u on $(0,T)\times\mathbb{R}^{d-1}\times\mathbb{R}^+$ such that

$$\forall j = 0, ..., s: \quad \partial_t^j u \in C^0([0,T], H^{s-j}(\mathbb{R}^{d-1}\times\mathbb{R}^+)).$$

In what follows, we consider a given subsonic planar phase boundary as a piecewise constant reference solution. Precisely, we choose a piecewise constant function

$$(\rho^+, \mathbf{u}^+) = (\rho_r, \mathbf{u}_r) \quad \text{and} \quad (\rho^-, \mathbf{u}^-) = (\rho_l, \mathbf{u}_l) \tag{14}$$

with constant values $(\rho_{r,l}, \mathbf{u}_{r,l}) \in \mathbb{R}^+ \times \mathbb{R}^d$ such that the following conditions hold

$$(J1)-(J3) \text{ are satisfied for some } \sigma \in \mathbb{R} \qquad \text{(jump conditions)}, \tag{15}$$

$$0 < \rho_r < v^* < l^* < \rho_l \qquad \text{(liquid–vapor)}, \tag{16}$$

$$0 < \frac{|\mathbf{u}_{r,l} \cdot \mathbf{n} - \sigma|}{c_{r,l}} < 1 \text{ with } c_{r,l} := \sqrt{p'(\rho_{r,l})} \qquad \text{(subsonic)}, \tag{17}$$

where v^*, l^* are given by the chosen van der Waals pressure function. Note that the existence of such a constant function (14) satisfying (15)–(17) for some $\sigma \in \mathbb{R}$ is ensured by [2, Proposition 2]. In particular, we have

$$0 < \rho_r < \rho_m < v^* < l^* < \rho_M < \rho_l \,,$$

where (ρ_m, ρ_M) are the Maxwell states such that $p(\rho_m) = p(\rho_M)$ and $g(\rho_m) = g(\rho_M)$, see [2]. We summarize the constant reference solution (14) satisfying (15)–(17) in

$$\mathbf{U}_{\text{ref}} := (\rho_r, \mathbf{u}_r, \rho_l, \mathbf{u}_l) \quad \text{and} \quad X_{\text{ref}} := 0. \tag{18}$$

We state in the following the main result about the existence of a solution to the nonlinear problem (**NL**).

Theorem 2.1 (Main Result). *Let $s > \frac{d+3}{2}$ and a subsonic reference solution $\mathbf{U}_{\text{ref}}$ as in (18) be given.*

Then there exists a constant $\delta > 0$ such that for all

$$(\mathbf{U}_0 - \mathbf{U}_{\text{ref}}) \in H^{s+1/2}(\mathbb{R}^{d-1} \times \mathbb{R}^+)$$

and all

$$X_0 \in H^{s+3/2}(\mathbb{R}^{d-1}),$$

which are compatible to order $s-1$ and satisfy

$$\|\mathbf{U}_0 - \mathbf{U}_{\text{ref}}\|_{L^\infty(\mathbb{R}^{d-1}\times\mathbb{R}^+)} + \|\check{\nabla} X_0\|_{L^\infty(\mathbb{R}^{d-1})} + \|\check{\Delta} X_0\|_{L^\infty(\mathbb{R}^{d-1})} < \delta,$$

*there exists a number $T > 0$ and a classical solution $(\mathbf{U} - \mathbf{U}_{\text{ref}}, X)$ of the nonlinear system (**NL**) with*

$$(\mathbf{U} - \mathbf{U}_{\text{ref}}) \in C H^s\left((0,T) \times \mathbb{R}^{d-1} \times \mathbb{R}^+\right) \tag{19}$$

and

$$X \in \left(H^{s+1}((0,T) \times \mathbb{R}^{d-1}) \cap L^2((0,T), H^{s+2}(\mathbb{R}^{d-1}))\right) . \tag{20}$$

Further, this solution is unique in the set of all solutions which satisfy (19) and (20).

To close the section, we reformulate Theorem 2.1 in the setting of the original problem (E1)–(E2) with the jump conditions (J1)–(J3). In particular, we characterize the solution in terms of the entropy concept. We mention that the interface is described by

$$\Sigma(t) = \left\{ x = (x_1, ..., x_d) \in \mathbb{R}^d \;\middle|\; x_d = X(x_1, ..., x_{d-1}, t) \right\}$$

and separates the vapor domain $V_+(t)$ and the liquid domain $V_-(t)$, i.e.,

$$\mathbb{R}^d = V_+(t) \cup \Sigma(t) \cup V_-(t)$$

for all $t \geq 0$.

Corollary 2.2. *Let $s > \frac{d+3}{2}$ and the initial data (A1) be given. There exists a constant $\delta > 0$ such that for all*

$$(\rho_0^{\pm} - \rho_{r,l}, \mathbf{u}_0^{\pm} - \mathbf{u}_{r,l}) \in H^{s+1/2}(V_{\pm}(0))$$

and all

$$X_0 \in H^{s+3/2}(\mathbb{R}^{d-1}),$$

which are compatible to order $s-1$ and satisfy

$$\|(\rho_0^- - \rho_r, \mathbf{u}_0^- - \mathbf{u}_r, \rho_0^+ - \rho_l, \mathbf{u}_0^+ - \mathbf{u}_l)\|_{L^\infty(\mathbb{R}^{d-1}\times\mathbb{R}^+)} + \|\check{\nabla} X_0\|_{L^\infty(\mathbb{R}^{d-1})} + \|\check{\Delta} X_0\|_{L^\infty(\mathbb{R}^{d-1})} < \delta,$$

there exists a number $T > 0$ and a classical shock wave solution

$$(\rho^{\pm} - \rho_{r,l}, \mathbf{u}^{\pm} - \mathbf{u}_{r,l}) \in C^1(V_{\pm}(t))$$

and

$$\big(X, X(t,\cdot)\big) \in C^1([0,T], \mathbb{R}^{d-1}) \times C^2(\mathbb{R}^{d-1})$$

for all $t \in [0,T]$ to the original problem (E1)–(E2) satisfying the jump conditions (J1)–(J3) and

$$\partial_t E(\rho, \mathbf{u}) + \nabla \cdot (E(\rho, \mathbf{u})\mathbf{u} + p(\rho)\mathbf{u}) \leq 0$$

in the distributional sense.

Note that Theorem 2.1 and Corollary 2.2 are only valid for restricted surface tension $\mathrm{s} \in (0, \mathrm{s}_0)$ and interfacial mobility constant $\mathrm{B} \in (0, \mathrm{B}_0)$, where $\mathrm{s}_0 > 0$ and $\mathrm{B}_0 > 0$ are given by Theorem 7 from [7].

3 Proof of the Main Result

We sketch the proof whose details can be found in [9]. The idea in the proof is to use the results about the well-posedness of the linearized version of the problem which was studied in [7, 8]. That means we will consider a linear system (see (21) below) which will be solved in each step of a successive iteration. For coefficients satisfying some assumptions (see [9] for the explicit formulation), we obtain a sequence of solutions to the system (21). For small enough initial data, this sequence will converge to the solution of the original nonlinear problem (**NL**). We refer to [3–5, 11] where this technique has been applied to standard hydrodynamical shock waves while we here deal with subsonic phase boundaries including a kinetic relation and surface tension.

In the case of vanishing surface tension, the second-order terms of the front position X (resulting from the Young–Laplace law) are not contained, which simplifies the statement of the problem. We have the case of positive surface tension. We consider system (21) for given $(\mathbf{U}^k, X^k) \in W^{1,\infty}(\mathbb{R}^+ \times \mathbb{R}^{d-1} \times \mathbb{R}^+) \times W^{1,\infty}(\mathbb{R}^+ \times \mathbb{R}^{d-1})$ and unknown $(\mathbf{U}^{k+1}, X^{k+1})$. The iteration scheme reads as

$$\begin{cases} \sum_{j=0}^{d-1} \mathbf{A}_j(\mathbf{U}^k)\partial_j \mathbf{U}^{k+1} + \mathbf{A}_z(\mathbf{U}^k)\partial_z \mathbf{U}^{k+1} = 0, \\ \sum_{j=0}^{d-1} \mathbf{b}_j(\mathbf{U}^k)\partial_j X^{k+1} + \mathbf{b}_s \check{\Delta} X^{k+1} + \mathbf{M}(\mathbf{U}^k, \check{\nabla} X^k)\mathbf{U}^{k+1} = 0, \end{cases} \tag{21}$$

with initial data $(\mathbf{U}^{k+1}(t=0), X^{k+1}(t=0)) = (\mathbf{U}_0 - \mathbf{U}_{\mathrm{ref}}, X_0)$.

One can show by induction that the sequence $(\mathbf{U}^k, X^k)$ will be bounded uniformly with respect to $k \in \mathbb{N}_0$ for small initial data and small enough T in

$$\begin{aligned} &CH^s((0,T), \times \mathbb{R}^{d-1} \times \mathbb{R}^+) \\ &\quad \times \left(H^{s+1}((0,T) \times \mathbb{R}^{d-1}) \cap L^2((0,T), H^{s+2}(\mathbb{R}^{d-1})) \right), \end{aligned}$$

see [9]. The limit

$$\lim_{k \to \infty} (\mathbf{U}^k - \mathbf{U}_{\mathrm{ref}}, X^k) = (\mathbf{U} - \mathbf{U}_{\mathrm{ref}}, X)$$

satisfies (**NL**) with initial data $(\mathbf{U}_0 - \mathbf{U}_{\mathrm{ref}}, X_0)$ according to the construction of the iterative scheme (21).

The solution $(\mathbf{U} - \mathbf{U}_{\mathrm{ref}}, X)$ of (**NL**) is unique since the limit of the Cauchy sequence is unique. □

References

1. R. Abeyratne, J.K. Knowles, Kinetic relations and the propagation of phase boundaries in solids. Arch. Rational Mech. Anal. **114**, 119–154 (1991)
2. S. Benzoni-Gavage, Stability of multi-dimensional phase transitions in a Van der Waals fluid. Nonlinear Anal.: T.M.A. **31**(1/2), 243–263 (1998)
3. S. Benzoni-Gavage, D. Serre, *Multi-dimensional Hyperbolic Partial Differential Equations: First-Order Systems and Applications* (Clarendon Press, Oxford, 2001)
4. J.-F. Coulombel, Stability of multidimensional undercompressive shock waves. Interfaces Free Boundaries **5**(4), 367–390 (2003)
5. J.-F. Coulombel, Weakly stable multidimensional shocks. Annales de l'Institut Henri Poincare (C) Non Linear Analysis **21**(4), 401–443 (2004)
6. W. Dreyer, M. Hantke, G. Warnecke, Exact solutions to the Riemann problem for compressible isothermal Euler equations for two-phase flows with and without phase transition. Q. Appl. Math. **71**(3), 509–540 (2013)
7. B. Kabil, C. Rohde, The influence of surface tension and configurational forces on the stability of liquid-vapor interfaces. Nonlinear Anal. **107**, 63–75 (2014)
8. B. Kabil, C. Rohde, Persistence of undercompressive phase boundaries for isothermal Euler equations including configurational forces and surface tension. Math. Methods Appl. Sci. **39**, 5409–5426 (2016)
9. B. Kabil, C. Rohde, Local well-posedness of a free boundary value problem for compressible liquid-vapor flow. Preprint (2016)
10. H.O. Kreiss, Initial boundary value problems for hyperbolic systems. Commun. Pure Appl. Math. **23**, 277–298 (1970)
11. G. Métivier, Stability of multidimensional shocks. *Advances in the Theory of Shock Waves*, vol. 47 (Springer, Berlin, 2001), pp. 25–103
12. R. Renardy, *An Introduction to Partial Differential Equations*. Texts in Applied Mathematics, vol. 13 (Springer, Berlin, 2004)
13. C. Rohde, C. Zeiler, A relaxation Riemann solver for compressible two-phase flow with phase transition and surface tension. Appl. Numer. Math. **95**, 267–279 (2015)
14. L. Truskinovsky, *Shock Induced Transitions and Phase Structures in General Media* (Springer, Berlin, 1993), pp. 185–229
15. Y.-G. Wang, Z. Xin, Stability and existence of multidimensional subsonic phase transitions. Acta Math. Appl. Sin. Engl. Ser. **19**(4), 529–558 (2003)
16. S.-Y. Zhang, Existence of multidimensional phase transitions in a steady Van Der Waals flow. Dyn. Partial Differ. Equ. **10**(1), 79–97 (2013)
17. S.-Y. Zhang, Y.-G. Wang, Existence of multidimensional shock waves and phase boundaries. J. Differ. Equ. **244**(7), 1571–1602 (2008)

Some Numerical Results of Regional Boundary Controllability with Output Constraints

Touria Karite, Ali Boutoulout and Fatima Zahrae El Alaoui

Abstract This paper deals with the problem of constrained controllability governed by parabolic evolution equations. The purpose is to compute the control u which steers the studied system to a final state which is supposed to be unknown between two defined bounds, only on a boundary subregion Γ of the system evolution domain Ω. The main result is proved via Lagrangian multiplier approach, and the numerical part is given on the basis of the well-known Uzawa algorithm. These results are illustrated by a numerical example.

Keywords Distributed systems · Parabolic systems · Regional controllability Lagrangian approach · Semilinear systems · Boundary subregion · Heat equation Minimum energy · Uzawa algorithm

1 Introduction

Mathematical control theory is the area of application-oriented mathematics that deals with the basic principles underlying the analysis and design of control systems. To control an object means to influence its behavior so as to achieve a desired goal. In order to implement this influence, engineers build devices that incorporate various mathematical techniques. Control theory is a field that plays a major role in nearly every modern precision device. It appears in our homes, in cars, in industry, and in almost every device we use in our life.

T. Karite (✉) · A. Boutoulout · F. Z. El Alaoui
TSI Team, MACS Laboratory, Institute of Sciences, Moulay Ismail University, Meknes, Morocco
e-mail: touria.karite@gmail.com

A. Boutoulout
e-mail: boutouloutali@yahoo.fr

F. El Alaoui
e-mail: fzelalaoui2011@yahoo.fr

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_8

The development of automatic control is strongly connected to the industrial revolution and the development of modern technology. The need to control the new discovered sources of power arose immediately. When new production techniques were developed, there were needs to keep them operating smoothly with high quality. Controlling a system is the fact to find whether or not one trajectory of a dynamical system can be steered toward another one. Many works dealing with the problem have been carried out in a wide literature; see Curtain and Pritchard [3], Curtain and Zwart [4] and references therein.

Later on, the concept of "regional analysis" has been introduced by El Jai et al. (1995) for parabolic systems and by Zerrik and Larhrissi (2000) for hyperbolic linear ones. It is commonly used to refer to control problems in which the target of our interest is not fully specified as a state, but refers only to a smaller internal region ω of the system domain Ω. After that, Zerrik et al. have extended the work to the case where ω is a part of the boundary $\partial\Omega$ of the domain. Interesting results were proven. It was particularly shown that the minimum time and the transfer cost of regional controllability are less than those of the controllability on the whole domain Ω.

Constrained or enlarged controllability is not a new concept. It was introduced by Lions [10] in 1989. In his book, he was interested in studying the constrained exact controllability, so-called CEC, for the wave equation on a closed convex set G of $L^2(\Omega) \times H^{-1}(\Omega)$. It is clear that if we choose $G = \{0, 0\}$, we retrieve the classic notion of exact controllability. Zerrik and Ghafrani thought of something closer to real problems [16]. So they work with an interval $[\alpha(\cdot), \beta(\cdot)]$ instead of the convex G. The reason motivating this choice of controllability is that the mathematical models of a studied phenomenon are obtained from measurements or approximations and they are often affected by perturbations. Later on, Boutoulout et al. studied the problem for hyperbolic linear systems in internal case [1] and also in the boundary [6]. Boutoulout and Karite extended the study to semilinear systems in internal case [8]. In this paper, we will extend the previous works to the case of boundary controllability of semilinear systems.

The following paper deals with the controllability properties of semilinear systems of parabolic equations where the control is exerted at the boundary.

Thus, let Ω be an open bounded subset of $\mathbb{R}^n (n \geq 1)$ with regular boundary $\partial\Omega$. And for a given $T > 0$, let's consider $Q = \Omega \times]0, T[$, $\Sigma = \partial\Omega \times]0, T[$ and let us consider a parabolic system excited by controls which may be applied via various types of actuators given by the following system:

$$\begin{cases} \partial_t y(x,t) - \mathcal{A}y(x,t) = \mathcal{F}y(x,t) + Bu(t) & Q \\ y(x,0) = y_0(x) & \Omega \\ \partial_\nu y(\xi,t) = 0 & \Sigma, \end{cases} \tag{1}$$

where the operator $\mathcal{A}$ is linear, second-order, and infinitesimal generator of a C_0-semigroup $(S(t))_{t\geq 0}$ on $L^2(\Omega)$ and $\mathcal{F} : H^1(\Omega) \to H^1(\Omega)$ a nonlinear operator which satisfies a Lipschitz condition in y [12, 14]. $\partial_\nu y(\xi, t)$ indicates the conormal derivative on the boundary Σ associated with the operator $\mathcal{A}$ and the unit out-

ward normal vector ν. $B \in \mathcal{L}(\mathbb{R}^m, L^2(\Omega))$, $y_0 \in L^2(\Omega)$ and $u \in \mathcal{U} = L^2(0, T; \mathbb{R}^m)$ (where m is the number of actuators).

The rest of the paper is organized as follows. In the next section, we give some definitions and notations. In Sect. 3, we present some results related to the Lagrangian method. In particular, we give details on the saddle point problem. In Sect. 4, we give an algorithm and a numerical example with simulations.

2 Preliminaries and Notations

Without loss of generality, we denote by $y_u(.)$ the solution of (1) when it is excited by a control u, we have $y_u(T) \in H^1(\Omega)$ (see [9]), and we consider:

- Γ a non-empty subregion of $\partial\Omega$.
- $\gamma_0 : H^1(\Omega) \to H^{1/2}(\partial\Omega)$ the trace operator of order zero which is linear, continuous, and surjective.
- The restriction operator

$$\begin{aligned} \chi_\Gamma : H^{1/2}(\partial\Omega) &\longrightarrow H^{1/2}(\Gamma) \\ y &\longmapsto \chi_\Gamma y = y|_\Gamma . \end{aligned}$$

Let's consider $H : \mathcal{U} \to H^1(\Omega)$ defined by:

$$\forall u \in \mathcal{U}, \quad Hu = \int_0^T S(T-s)Bu(s)ds,$$

we define also the following operator:

$$\begin{aligned} G_\Gamma : L^2(0, T; H^1(\Omega)) &\longrightarrow H^1(\Omega) \\ y(\cdot) &\longmapsto \int_0^T S(T-\tau)\mathcal{F}y(\tau)d\tau . \end{aligned} \tag{2}$$

and $b_1(\cdot), b_2(\cdot)$ be two given real functions in $H^{1/2}(\Gamma)$ such that $b_1(\cdot) \le b_2(\cdot)$ on Γ, and we set:

$$[b_1(\cdot), b_2(\cdot)] = \left\{ y \in H^{1/2}(\Gamma) \quad | \quad b_1(\cdot) \le y(\cdot) \le b_2(\cdot) \quad \text{on} \quad \Gamma \right\}.$$

Then we have the following definition, remark, and proposition:

Definition 1. *We say that* (1) *is* $[b_1(\cdot), b_2(\cdot)]$*-controllable on* Γ *if:*

$$\exists u \in \mathcal{U} \quad \textit{such that} \quad b_1(\cdot) \le \chi_\Gamma \left(\gamma_0 y_u(T)\right) \le b_2(\cdot).$$

It is clear that the system (1) *is* $[b_1(\cdot), b_2(\cdot)]$*-controllable on* Γ *if:*

$$[b_1(\cdot), b_2(\cdot)] - \{\chi_\Gamma \gamma_0 S(T) y_0\} \cap \left(Im\, \chi_\Gamma \gamma_0 G_\Gamma + Im\, \chi_\Gamma \gamma_0 H\right) \neq \emptyset.$$

Remark 1. 1. The above definition means that we are interested in the transfer of the system (1) to an unknown state between $b_1(\cdot)$ and $b_2(\cdot)$ on Γ.
2. If $b_1 = b_2$, we retrieve the regional exact controllability. So, for $b_1 \neq b_2$ the $[b_1(\cdot), b_2(\cdot)]$-controllability constitutes an extension of regional controllability.
3. A system which is controllable on Γ is $[b_1(\cdot), b_2(\cdot)]$-controllable on Γ.

The $[b_1(\cdot), b_2(\cdot)]$-controllability on Γ could be characterized by the following proposition:

Proposition 1. *The system (1) is* $[b_1(\cdot), b_2(\cdot)]$*-controllable on* Γ *if and only if*

$$\left(Ker\,\chi_\Gamma + Im\,\gamma_0 G_\Gamma + Im\,\gamma_0 H\right) \cap [b_1(\cdot), b_2(\cdot)] \neq \emptyset.$$

Proof. We suppose that the system (1) is $[b_1(\cdot), b_2(\cdot)]$-controllable on Γ which is equivalent to say that

$$\left(Im\,\chi_\Gamma\gamma_0 G_\Gamma + Im\,\chi_\Gamma\gamma_0 H\right) \cap [b_1(\cdot), b_2(\cdot)] \neq \emptyset.$$

So there exists $z \in [b_1(\cdot), b_2(\cdot)]$, $y(\cdot) \in L^2(0, T; H^1(\Omega))$ and $u \in \mathcal{U}$ such that $\chi_\Gamma\gamma_0 G_\Gamma y(\cdot) + \chi_\Gamma\gamma_0 Hu = \chi_\Gamma\gamma_0 z$, which gives $\chi_\Gamma\left(z - \gamma_0 G_\Gamma y(\cdot) - \gamma_0 Hu\right) = 0$. Let's consider $z_1 = z - \gamma_0 G_\Gamma y(\cdot) - \gamma_0 Hu$, $z_2 = \gamma_0 G_\Gamma y(\cdot)$ and $z_3 = \gamma_0 Hu$. Then: $z = z_1 + z_2 + z_3$ where $z_1 \in Ker\,\chi_\Gamma$, $z_2 \in Im\,\gamma_0 G_\Gamma$ and $z_3 \in Im\,\gamma_0 H$, which prove that $z \in \left(Ker\,\chi_\Gamma + Im\,\gamma_0 G_\Gamma + Im\,\gamma_0 H\right)$. Thus,

$$\left(Ker\,\chi_\Gamma + Im\,\gamma_0 G_\Gamma + Im\,\gamma_0 H\right) \cap [b_1(\cdot), b_2(\cdot)] \neq \emptyset.$$

Conversely, we suppose that $\left(Ker\,\chi_\Gamma + Im\,\gamma_0 G_\Gamma + Im\,\gamma_0 H\right) \cap [b_1(\cdot), b_2(\cdot)] \neq \emptyset$ which means that there exists $z \in [b_1(\cdot), b_2(\cdot)]$ such that $z \in Ker\,\chi_\Gamma + Im\,\gamma_0 G_\Gamma + Im\,\gamma_0 H$ so $z = z_1 + z_2 + z_3$, with $\chi_\Gamma z_1 = 0$, $\exists\, y \in L^2(0, T; H^1(\Omega)) \mid z_2 = \gamma_0 G_\Gamma y(\cdot)$ and $\exists\, u \in \mathcal{U} \mid z_3 = \gamma_0 Hu$, then by applying the restriction operator to z we will have $\chi_\Gamma z = \chi_\Gamma(z_1 + z_2 + z_3) = \chi_\Gamma\gamma_0 G_\Gamma y(\cdot) + \chi_\Gamma\gamma_0 Hu$, which gives $\chi_\Gamma z \in \left(Im\,\gamma_0 G_\Gamma + Im\,\gamma_0 H\right)$. And, we have

$$\left(Im\,\gamma_0 G_\Gamma + Im\,\gamma_0 H\right) \cap [b_1(\cdot), b_2(\cdot)] \neq \emptyset.$$

Thus (1) is $[b_1(\cdot), b_2(\cdot)]$-controllable on Γ.

Now, Let us recall that an actuator is conventionally defined by a couple (D, f), where D is a non-empty closed part of $\overline{\Omega}$, and it represents the geometric support of the actuator. And $f \in L^2(D)$ defines the spatial distribution of the action on the support D.
In the case of a pointwise actuator (internal or boundary) $D = \{b\}$ and $f = \delta(b - .)$, where δ is the Dirac mass concentrated in b, and the actuator is then denoted by (b, δ_b). For definitions and properties of strategic actuators, we refer to [5, 15].

Definition 2. *The actuator* (D, f) *is said to be* $[b_1(\cdot), b_2(\cdot)]$*-strategic on* Γ *if the excited system is* $[b_1(\cdot), b_2(\cdot)]$*-controllable on* Γ.

3 Lagrangian Approach

We consider the problem:

$$\begin{cases} \inf \dfrac{1}{2}\|u\|^2 \\ u \in \mathcal{U}_{ad}, \end{cases} \tag{3}$$

where $\mathcal{U}_{ad} = \{u \in \mathcal{U} \mid \chi_\Gamma \gamma_0 y_u(T) \in [b_1(\cdot), b_2(\cdot)]\}\cdot$ And the system (1) is excited by one zonal control.
The following proposition gives a useful characterization of the solution.

Theorem 1. *If the system (1) is* $[b_1(\cdot), b_2(\cdot)]$*-controllable on* Γ *then the solution of (3) is given by:*

$$u^* = -(\chi_\Gamma \gamma_0 H)^* \lambda^*, \tag{4}$$

where $\lambda^* \in H^{1/2}(\Gamma)$ *satisfies:*

$$\begin{cases} R_\Gamma \lambda^* + z^* = \chi_\Gamma \gamma_0 \left[S(T)y_0 + G_\Gamma y(\cdot)\right] \\ z^* = P_{[b_1(\cdot), b_2(\cdot)]}(\rho\lambda^* + z^*), \end{cases} \tag{5}$$

while $P_{[b_1(\cdot), b_2(\cdot)]} : H^{1/2}(\Gamma) \to [b_1(\cdot), b_2(\cdot)]$ *denotes the projection operator,* $\rho > 0$ *and* $R_\Gamma = (\chi_\Gamma \gamma_0 H)(\chi_\Gamma \gamma_0 H)^*\cdot$

Proof. If the system (1) is $[b_1(\cdot), b_2(\cdot)]$-controllable on Γ then $\mathcal{U}_{ad} \neq \emptyset$, and the problem (3) has a unique solution.
Problem (3) is equivalent to the saddle point problem:

$$\begin{cases} \inf \dfrac{1}{2}\|u\|^2 \\ (u, z) \in Z, \end{cases} \tag{6}$$

where $Z = \left\{(u, z) \in \mathcal{U} \times [b_1(\cdot), b_2(\cdot)] \mid \chi_\Gamma \gamma_0 y_u(T) - z = 0\right\}\cdot$
To study this constraints, we will use a Lagrangian functional and steer the problem (6) to a saddle point problem.
We associate to the problem (6) the Lagrangian functional defined by:

$$\forall (u, z, \lambda) \in \mathcal{U} \times [b_1(\cdot), b_2(\cdot)] \times H^{1/2}(\Gamma) \quad L(u, z, \lambda) = \frac{1}{2}\|u\|^2 + \langle \lambda, \chi_\Gamma \gamma_0 y_u(T) - z\rangle_{H^{1/2}(\Gamma)}\cdot \tag{7}$$

The set $\mathcal{U} \times [b_1(\cdot), b_2(\cdot)]$ is non-empty, closed, and convex. The functional L satisfies conditions:
• $(u, z) \mapsto L(u, z, \lambda)$ is convex and lower semicontinuous for all $\lambda \in H^{1/2}(\Gamma)$.
• $\lambda \mapsto L(u, z, \lambda)$ is concave and upper semicontinuous for all $(u, z) \in \mathcal{U} \times [b_1(\cdot), b_2(\cdot)]$.

Moreover, there exists $\lambda_0 \in H^{1/2}(\Gamma)$ such that:

$$\lim_{\|(u,z)\|\to+\infty} L(u, z, \lambda_0) = +\infty, \tag{8}$$

and there exists $(u_0, z_0) \in \mathcal{U} \times [b_1(\cdot), b_2(\cdot)]$ such that

$$\lim_{\|\lambda\|\to+\infty} L(u_0, z_0, \lambda) = -\infty\cdot \tag{9}$$

Then, the functional L admits a saddle point. For more details, we refer to [11].
Let (u^*, z^*, λ^*) be a saddle point of L and prove that u^* is the solution of (3). We have:

$$L(u^*, z^*, \lambda) \le L(u^*, z^*, \lambda^*) \le L(u, z, \lambda^*) \quad \forall\, (u, z, \lambda) \in \mathcal{U} \times [b_1(\cdot), b_2(\cdot)] \times H^{1/2}(\Gamma) \tag{10}$$

From the first inequality of (10), we have:

$$\langle \lambda, \chi_\Gamma \gamma_0 y_{u^*}(T) - z^* \rangle_{H^{1/2}(\Gamma)} \le \langle \lambda^*, \chi_\Gamma \gamma_0 y_{u^*}(T) - z^* \rangle_{H^{1/2}(\Gamma)} \quad \forall\, \lambda \in H^{1/2}(\Gamma),$$

which implies $\chi_\Gamma \gamma_0 y_{u^*}(T) = z^*$ and hence $\chi_\Gamma \gamma_0 y_{u^*}(T) \in [b_1(\cdot), b_2(\cdot)]$.
The second inequality of (10) means that for all $u \in \mathcal{U}$ and $z \in [b_1(\cdot), b_2(\cdot)]$, we have:

$$\frac{1}{2}\|u^*\|^2 + \langle \lambda^*, \chi_\Gamma \gamma_0 y_{u^*}(T) - z^* \rangle_{H^{1/2}(\Gamma)} \le \frac{1}{2}\|u\|^2 + \langle \lambda^*, \chi_\Gamma \gamma_0 y_u(T) - z \rangle_{H^{1/2}(\Gamma)}$$

for all $(u, z) \in \mathcal{U} \times [b_1(\cdot), b_2(\cdot)]$. Since $\chi_\Gamma \gamma_0 y_{u^*}(T) = z^*$, it follows that:

$$\frac{1}{2}\|u^*\|^2 \le \frac{1}{2}\|u\|^2 + \langle \lambda^*, \chi_\Gamma \gamma_0 y_u(T) - z \rangle_{H^{1/2}(\Gamma)} \ \forall\, (u, z) \in \mathcal{U} \times [b_1(\cdot), b_2(\cdot)].$$

Taking $z = \chi_\Gamma \gamma_0 y_u(T) \in [b_1(\cdot), b_2(\cdot)]$, we obtain:

$$\frac{1}{2}\|u^*\|^2 \le \frac{1}{2}\|u\|^2,$$

which implies that u^* is of minimum energy.
The following assumptions hold, if (u^*, z^*, λ^*) is a saddle point of L:

$$\langle u^*, u - u^* \rangle + \langle \lambda^*, \chi_\Gamma \gamma_0 H(u - u^*) \rangle = 0 \quad \forall\, u \in \mathcal{U}, \tag{11}$$

$$-\langle \lambda^*, z - z^* \rangle \ge 0 \quad \forall\, z \in [b_1(\cdot), b_2(\cdot)], \tag{12}$$

$$\langle \lambda - \lambda^*, \chi_\Gamma \gamma_0 y_{u^*}(T) - z^* \rangle = 0 \quad \forall\, \lambda \in H^{1/2}(\Gamma)\cdot \tag{13}$$

For more details about the saddle point and its theory, we refer to [2, 7, 13].
From (11), we deduce (4).

The equation (13) is equivalent to:

$$\chi_\Gamma \gamma_0 y_{u^*}(T) = z^*.$$

And since $y_{u^*}(T) = S(T)y_0 + G_\Gamma y(\cdot) + Hu^*$, we have:

$$\chi_\Gamma \gamma_0 \left[S(T)y_0 + G_\Gamma y(\cdot) + Hu^*\right] - z^* = 0.$$

Then $\chi_\Gamma \gamma_0 \left[S(T)y_0 + G_\Gamma y(\cdot)\right] + \chi_\Gamma \gamma_0 Hu^* = z^*.$
With (4), we have:

$$\chi_\Gamma \gamma_0 \left[S(T)y_0 + G_\Gamma y(\cdot)\right] - \left(\chi_\Gamma \gamma_0 H\right)\left(\chi_\Gamma \gamma_0 H\right)^* \lambda^* = z^*,$$

with $R_\Gamma = \left(\chi_\Gamma \gamma_0 H\right)\left(\chi_\Gamma \gamma_0 H\right)^*$, we obtain the first equation of (5).
And from inequality (12), we obtain:

$$\langle (\rho\lambda^* + z^*) - z^*, z - z^* \rangle_{H^{1/2}(\Gamma)} \leq 0 \qquad \forall\, z \in [b_1(\cdot), b_2(\cdot)]\ ,\ \rho > 0,$$

which is equivalent to the second equation of (5).

Corollary 1. *If the system (1) is exactly controllable on Γ and ρ suitably chosen, then the system (5) has only one solution (λ^*, z^*).*

Proof. The regional exact controllability on Γ implies that $(\chi_\Gamma \gamma_0 H)^*$ and R_Γ are bijective. So if (u^*, z^*, λ^*) is a saddle point of L then the system (5) is equivalent to

$$\begin{cases} R_\Gamma \lambda^* + z^* = \chi_\Gamma \gamma_0 \left[S(T)y_0 + G_\Gamma y(\cdot)\right] \\ z^* = P_{[b_1(\cdot), b_2(\cdot)]} \left[\left(-\rho R_\Gamma^{-1} z^* + \rho R_\Gamma^{-1} \chi_\Gamma \gamma_0 \left(S(T)y_0 + G_\Gamma y(\cdot)\right)\right) + z^*\right]. \end{cases} \tag{14}$$

It follows that z^* is a fixed point of the function

$$\begin{array}{rccl} F_\rho : & [b_1(\cdot), b_2(\cdot)] & \longrightarrow & [b_1(\cdot), b_2(\cdot)] \\ & y & \longmapsto & P_{[b_1(\cdot), b_2(\cdot)]} \left[\left(-\rho R_\Gamma^{-1} y + \rho R_\Gamma^{-1} \chi_\Gamma \gamma_0 \left(S(T)y_0 + G_\Gamma y(\cdot)\right)\right) + y\right]. \end{array} \tag{15}$$

The operator R_Γ is coercive, i.e.,

$$\exists m > 0 \quad \text{such that} \quad \langle R_\Gamma^{-1} y, y \rangle \geq m \|y\|^2 \quad \forall\, y \in H^{1/2}(\Gamma).$$

It follows that

$$\begin{aligned}\|F_\rho(y)-F_\rho(z)\|^2_{H^{1/2}(\Gamma)} &\le \|-\rho R_\Gamma^{-1}(y-z)+(y-z)\|^2\\ &\le \rho^2\|R_\Gamma^{-1}\|^2\|y-z\|^2+\|y-z\|^2-2\rho\langle R_\Gamma^{-1}(y-z),(y-z)\rangle\\ &\le \rho^2\|R_\Gamma^{-1}\|^2\|y-z\|^2+\|y-z\|^2-2\rho m\|y-z\|^2\\ &\le (1+\rho^2\|R_\Gamma^{-1}\|^2-2\rho m)\|y-z\|^2 \quad \forall\, y,z\in[b_1(\cdot),b_2(\cdot)]\end{aligned}$$

and we deduce that if $0<\rho<\dfrac{2m}{\|R_\Gamma^{-1}\|^2}$, then F_ρ is a contraction mapping. This implies the uniqueness of z^* and λ^*.

Remark 2. If $b_1(\cdot)=b_2(\cdot)$, we obtain the notion of controllability on Γ, and the solution of (3) is given by:

$$u^*=\left(\chi_\Gamma\gamma_0 H\right)^* R_\Gamma^{-1}\left[b_1(\cdot)-\chi_\Gamma\gamma_0\left(S(T)y_0+R_\Gamma y(\cdot)\right)\right].$$

4 Numerical Approach and Simulations

From the previous proposition (1), it follows that the solution of the problem (3) arises to compute the saddle points of L which is equivalent to solving the problem

$$\inf_{(u,z)\in\mathcal{U}\times[b_1(\cdot),b_2(\cdot)]}\Big(\sup_{\lambda\in H^{1/2}(\Gamma)}L(u,z,\lambda)\Big)\cdot \tag{16}$$

Finally, we develop the following algorithm based on the algorithm of Uzawa type (see [7]):

Algorithm

Step 1: initialization.

- Fix the two functions $b_1(\cdot) and b_2(\cdot)$,
- Choose two functions $(z_0,\lambda_1)\in[b_1(.),b_2(.)]\times H^{1/2}(\Gamma)$,
- Choose a region Γ, time T and the position b of the actuator,
- Choose a threshold accuracy ϵ small enough.

Step 2: Until $\|\lambda_{n+1}-\lambda_n\|\le\epsilon$, repeat

- Solve equation $u_n=-(\chi_\Gamma\gamma_0 H)^*\lambda_n$,
- Solve equation $z_n=P_{[b_1(.),b_2(.)]}(\rho\lambda_n+z_{n-1})$,
- Calculate λ_{n+1} by the formula $\lambda_{n+1}=\lambda_n+(\chi_\Gamma\gamma_0 y_{u_n}(T)-z_n)$.

Step 3 : Let (u^*,z^*,λ^*) be a saddle point of L, then the sequence (u_n) converges to u^* solution of the problem (3) and the sequence (z_n) converges to z^*.

To validate this algorithm, we use the following example:

Example
We consider a two-dimensional system defined on $\Omega =]0, 1[\times]0, 1[$, described by the following parabolic system:

$$\begin{cases} \dfrac{\partial y(x,t)}{\partial t} - \Delta y(x,t) - cy^2(x,t) = \delta(x-b)u(t) & \Omega\times]0,T[\\ y(x,0) = 0 & \Omega \\ \dfrac{\partial y}{\partial \nu}(\xi,t) = 0 & \partial\Omega\times]0,T[\ , \end{cases} \tag{17}$$

where $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0.10 \\ 0.15 \end{pmatrix}$, $\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$, c = 0.01, T = 2 and $\Gamma =]0, 1[\times\{0\}$.
Let's consider $b_1(x_1, x_2) = \dfrac{1}{5}(2x_1(x_1-1) + x_2(x_2-1))$, $b_2(x_1, x_2) = -\dfrac{1}{3}(x_1(x_1-1) + x_2(x_2-1))$ and $\epsilon = 10^{-4}$. Applying the previous algorithm, we obtain the following results:

$$u_n = \sum_{i=0}^{\infty}\sum_{k,l=0}^{\infty}\left(\int_0^T e^{\gamma_{kl}(T-t)}\psi_i(t)dt\right)\varphi_{kl}(b_1,b_2)\langle\lambda_n, \varphi_{kl}\rangle_{L^2(\Gamma)}\psi_i, \tag{18}$$

with $(\psi_i)_i$ is a complete set of eigenfunctions in $L^2(0, T; \mathbb{R}^m)$ associated with eigenvalues β_i. And $(\varphi_{kl})_{kl}$ is a complete set of eigenfunctions in $(L^2(\Omega))^2$ associated with eigenvalues γ_{kl}.

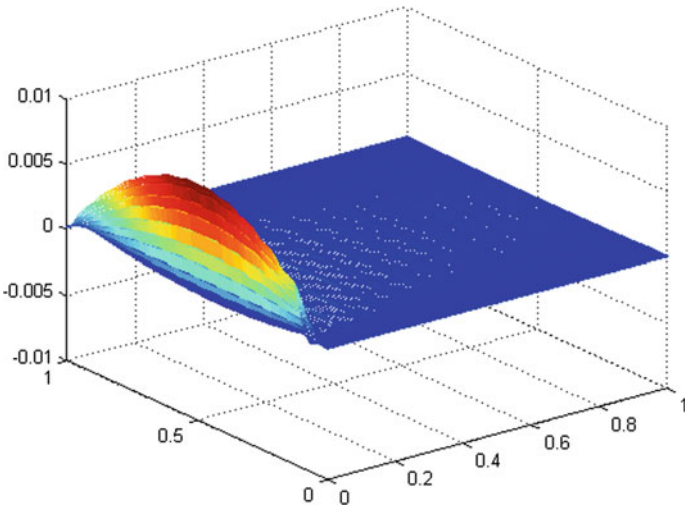


Fig. 1 Representation of $\chi_\Gamma\gamma_0 z_d$

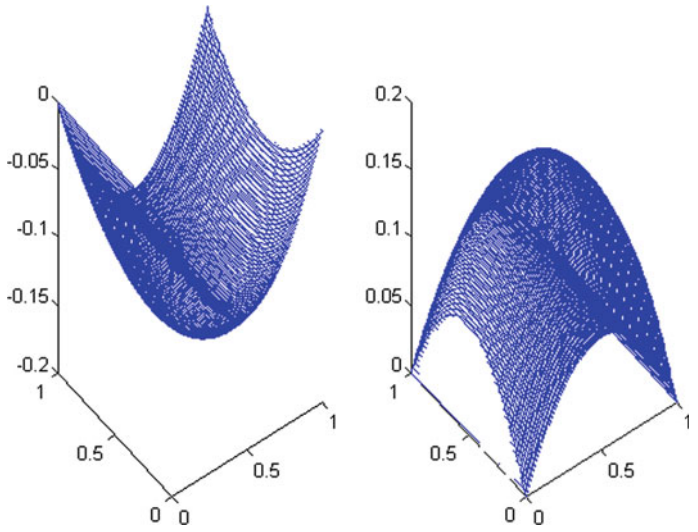


Fig. 2 Two functions $b_1(\cdot)$ and $b_2(\cdot)$

Figure 1 represents the restriction of the trace of the desired state on the whole domain Ω and Fig. 2 represents the two chosen functions $b_1(x)$ and $b_2(x)$ in Ω.

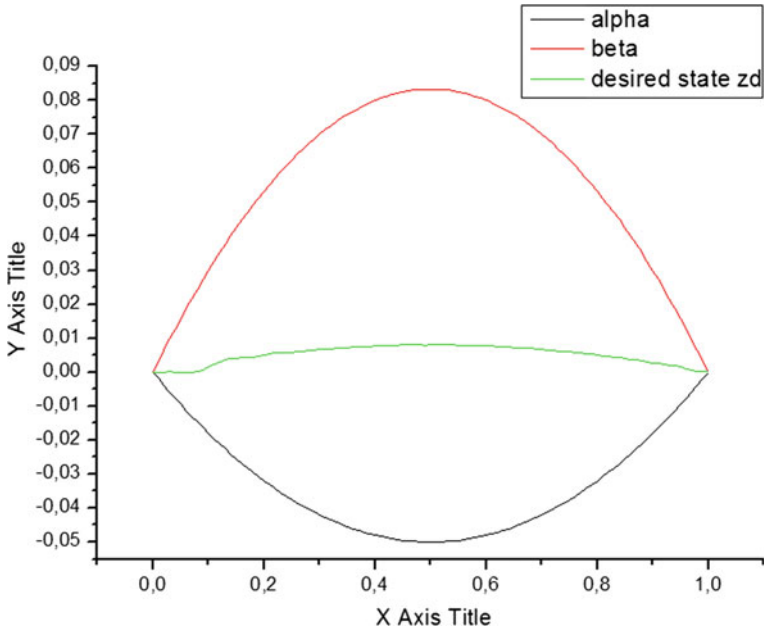


Fig. 3 Projection of the desired state z_d on the boundary Γ

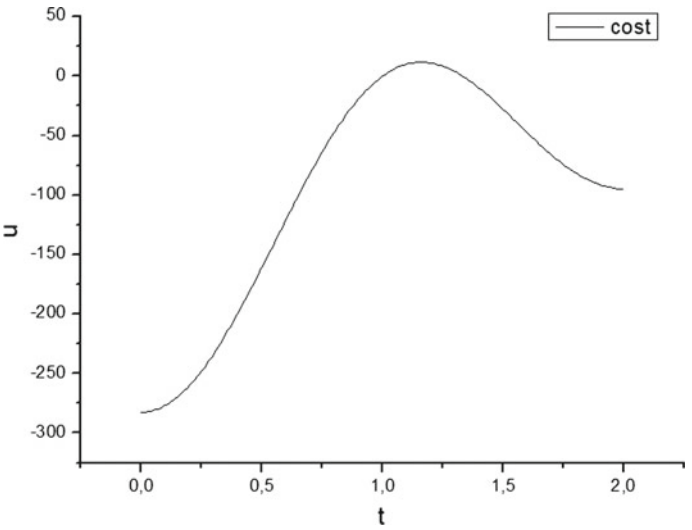


Fig. 4 Transfer cost

In Fig. 3, we represent the projection of the desired state on the boundary $\Gamma = \{0\}\times]0, 1[$. And in the Fig. 4, we see the evolution of the transfer cost function in time T.

$$z_n(x) = \begin{cases} b_1(x) & \text{if} \quad \rho\lambda_n(x) + z_{n-1}(x) \leq b_1(x) \quad \text{a.e} \\ \rho\lambda_n(x) + z_{n-1}(x) & \text{if} \quad b_1(x) \leq \rho\lambda_n(x) + z_{n-1}(x) \leq b_2(x) \quad \text{a.e} \\ b_2(x) & \text{if} \quad \rho\lambda_n(x) + z_{n-1}(x) \geq b_2(x) \quad \text{a.e}\cdot \end{cases} \tag{19}$$

And finally, we compute λ_{n+1} by the following formula:

$$\lambda_{n+1} = \lambda_n + (\chi_\Gamma \gamma_0 y_{u_n}(T) - z_n).$$

We remark that the obtained state with the used algorithm is between the two given functions $b_1(\cdot)$ and $b_2(\cdot)$ with the error $\epsilon = 4.325 \times 10^{-7}$ which validate the used method.

5 Conclusion

We developed an extension of the regional controllability to a situation encountered in many real situations where the problem is to bring the state of a system between two prescribed functions on a part of the boundary. Other methods are subject to

calculate the optimal control such as Hilbert Uniqueness Method, penalty methods, and variational methods which will be developed in future works.

Acknowledgements The work has been carried out with a grant from Hassan II Academy of Sciences and Technology.

References

1. A. Boutoulout, H. Bourray, F.Z. El Alaoui, L. Ezzahri, Constrained controllability for distributed hyperbolic systems. Math. Sci. Lett. **3**(3), 207–214 (2014)
2. F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods* (Springer Verlag, New York, 1991)
3. R.F. Curtain, A.J. Pritchard, *Infinite Dimensional Linear Systems Theory* (Springer-Verlag, 1978)
4. R.F. Curtain, H. Zwart, *An Introduction to Infinite Dimensional Linear Systems Theory. Texts in applied mathematics* (Springer-Verlag, 1995)
5. A. El Jai, A.J. Pritchard, *Sensors and Actuators in Distributed Systems Analysis* (Wiley, New York, 1988)
6. L. Ezzahri, A. Boutoulout, F.Z. El Alaoui, Boundary constrained controllability for hyperbolic systems. Inf. Sci. Lett. **4**(2), 85–91 (2015)
7. M. Fortin, R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-value Problems*, vol 15 (North-Holland, 1983)
8. T. Karite, A. Boutoulout, Regional constrained controllability for parabolic semilinear systems. Int. J. Pure Appl. Math. **113**(1), 113–129 (2017)
9. J.L. Lions, E. Magenes, *Problèmes Aux Limites Non Homogènes et Applications* vol. 1 (Dunod, 1968)
10. J.L. Lions, Sur la contrôlabilité exacte élargie, in *Partial Differential Equations and the Calculus of Variations*. Progress in Nonlinear Differential Equations and Their Applications, vol. 1 (1989), pp. 703–727
11. A. Matei, Weak solvability via Lagrange multipliers for two frictional contact models. Ann. Univ. Buchar. (Math. Series) **4**(LXII), 179–191 (2013)
12. A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations* (Springer-Verlag, New York, 1990)
13. R.T. Rockafellar, Lagrange multipliers and optimality. SIAM Rev. **35**(2), 183–238 (1993)
14. E. Zeidler, *Applied Functional Analysis: Applications to Mathematical Physics* (Springer-Verlag, New York, 1995)
15. E. Zerrik, A. El Jai, A. Boutoulout, Actuators and regional boundary controllability of parabolic system. Int. J. Syst. Sci. **31**(1), 73–82 (2000)
16. E. Zerrik, F. Ghafrani, Minimum energy control subject to output constraints numerical approach. IEE Proc.-Control Theory Appl. **149**(1), 105–110 (2002)

Water Hammer Modeling for Water Networks via Hyperbolic PDEs and Switched DAEs

Rukhsana Kausar and Stephan Trenn

Abstract In water distribution network, instantaneous changes in valve and pump settings introduce jumps and sometimes impulses. In particular, a particular impulsive phenomenon which occurs due to sudden closing of valve is the so-called water hammer. It is classically modeled as a system of hyperbolic partial differential equations (PDEs). We observed that under some suitable assumptions the PDEs usually used to describe water flows can be simplified to differential algebraic equations (DAEs). The idea is to model water hammer phenomenon in the switched DAEs framework due to its special feature of studying such impulsive effects. To compare these two modeling techniques, a system of hyperbolic PDE model and the switched DAE model for a simple setup consisting of two reservoirs, six pipes, and three valves is presented. The aim of this contribution is to present results of both models as motivation for the claim that a switched DAE modeling framework is suitable for describing a water hammer.

Keywords Water hammer · Solution theory · Switched system · Dirac impulse

1 Introduction

The occurrence of hydraulic transients in the operation of water distribution network is inevitable. Such transients are planned or accidental changes of the network configuration. These sudden structural changes can have dramatic effects in flow regimes, ranging from pump defects to catastrophic pipeline failures. The flow of water in pipes is usually modeled as system of nonlinear hyperbolic balance laws

R. Kausar (✉)
Department of Mathematics, TU Kaiserslautern, Kaiserslautern, Germany
e-mail: kausar@mathematik.uni-kl.de

S. Trenn
Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands
e-mail: S.Trenn@rug.nl

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_9

(i.e., partial differential equations—PDEs); see, for example, [4], where the sudden structural changes lead to large peaks and fast transients in the solution.

We propose to model such fast transients in the framework of switched differential algebraic equation (switched DAE). This framework was originally introduced for modeling electrical circuits [12] and allows a precise mathematical description of peaks and fast transients in the form of Dirac impulses and jumps.

Our focus in this paper is on the so-called water hammer, which results from sudden changes of velocity in pipelines and can cause large pressures magnitudes. It is usually created by rapidly closing valves, shutting off or restarting pumps. Our goal is to show that these pressure peaks occurring in the PDE simulations can be well approximated by a suitable switched DAE model.

The paper is organized as follows. In Sect. 2, the water network and its components are defined as a graph and the mathematical models of the pipes and other components (like reservoir and valves) are introduced. In Sect. 3, we study in detail a simple water network which exhibits a water hammer; in particular, we derive the corresponding PDE model as well as a switched DAE model. In Sect. 4, we describe the solution theory used in solving our sample network problem. In Sect. 5, a numerical comparison of the PDE and switched DAE model is presented.

2 Mathematical Model

The structure of a water network can be modeled as a graph $G = (\mathbb{V}, \mathbb{E})$ where $\mathbb{V}$ is the set of nodes and $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ is the set of edges. Each edge $e \in \mathbb{E}$ corresponds to a pipe of the water network, and the nodes $v \in \mathbb{V}$ are the connections or endpoints of pipes, including junctions, pumps, valves, or reservoirs. We denote by $\gamma_v^-(\gamma_v^+)$ the set of all indices of edges $e_i \in \mathbb{E}$ outgoing (ingoing) from (to) the node $v \in \mathbb{V}$; see Fig. 1 for an illustration of this notation.

In the model of water network elements, the two main physical quantities pressure and flow are involved. Those values at the endpoints of the pipes are related to each

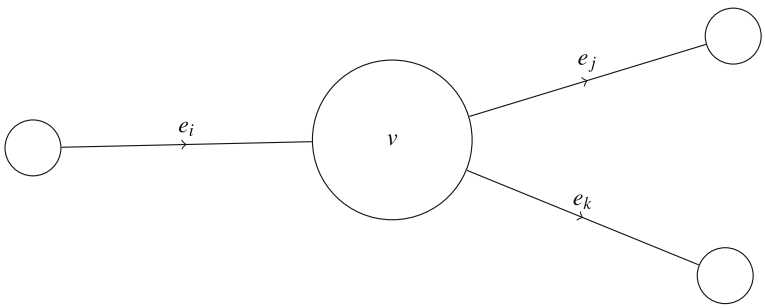


Fig. 1 A node v with three incident edges e_i, e_j, e_k; here, $\gamma_v^+ = \{i\}$ and $\gamma_v^- = \{j, k\}$

other corresponding to the type of node. Furthermore, the modeling of the flow in the pipes also involves density of the water. Usually, water is assumed to be incompressible; i.e., the density is assumed to be constant. However, our focus is on modeling the water hammer effect and for this it is necessary to take into account the (slight) compressibility of water.

2.1 Models of Water Flow in Pipe

One can model water flow in a pipe in two different ways depending on whether the compressibility of water is taken into account or not. In order to study transient phenomena like water hammer, it is necessary to model compressibility; in particular, density and mass flow become space-dependent quantities. On the other hand, to understand the qualitative behavior, in particular, in large networks, it often suffices to model water as incompressible fluid. We will briefly introduce both models in the following.

2.1.1 Compressible Flow in a Pipe

Following [1, 13] we use the following pressure law for compressible fluids:

$$P(\rho) = P_a + K\frac{\rho - \rho_a}{\rho_a}, \tag{1}$$

where $K > 0$ is the so-called *bulk modulus*, $P_a > 0$ is the atmospheric pressure, and $\rho_a > 0$ is the density at atmospheric pressure. The bulk modulus is related to the speed of sound $c > 0$ as follows:

$$c^2 = \frac{\partial P}{\partial \rho} = K/\rho_a. \tag{2}$$

Note that $\beta := 1/K$ is the so-called compressibility coefficient. We consider a completely filled pipe of length $L > 0$ with mass density $\rho(x,t) > 0$ and mass flux $q(x,t) \in \mathbb{R}$ both defined on $[0, L] \times \mathbb{R}_+$. The compressible flow of water in the pipe can be modeled by the balance law of the following form [3, Sect. 2]:

$$\begin{aligned} \partial_t \rho + \partial_x q &= 0, \\ \partial_t q + \partial_x \left(\frac{q^2}{\rho} + P(\rho)\right) &= -c_f \frac{q\,|q|}{2D\rho}, \end{aligned} \tag{3}$$

with the pressure law $P : \mathbb{R}_+ \to \mathbb{R}_+$ given by (1) and where $c_f > 0$ is the friction against the pipe wall and $D > 0$ is the diameter of the pipe. The initial condition for (3) is:

$$q(x,0) = q_0(x) \text{ and } P(\rho(x,0)) = p_0(x) \quad x \in [0,L], \tag{4}$$

for some initial flow $q_0 : [0, L] \to \mathbb{R}$ and some initial pressure $p_0 : [0, L] \to \mathbb{R}_+$. Note that the initial condition is given implicitly in terms of the pressure and not explicitly in terms of the density. The reason is that the pressure is the more relevant physical quantity, in particular, when the pipes are coupled with other water network elements. When the individual pipes are connected with other elements of the overall water distribution network, additional boundary and so-called coupling condition will be imposed.

2.1.2 Coupling Conditions at Intersection Nodes

The balance law (3) has to be completed by initial, boundary, and coupling conditions across the whole network. Suppose the initial data $P_l(\rho_l(x,0)) = p_{l,0}$ and $q_l(x,0) = q_{l,0}$ are given for each pipe l in the network, where ρ_l, q_l, and P_l denote density, flow, and pressure along each pipe edge e_l. Admissible boundaries must be chosen in accordance with the characteristics. Preservation of mass yields the coupling condition

$$\sum_{l \in \gamma_v^+} q_l(L,t) = \sum_{l \in \gamma_v^-} q_l(0,t). \tag{5}$$

and consistency of pressure yields

$$p_i(L,t) = p_j(0,t), \quad \forall i \in \gamma_v^+, \quad j \in \gamma_v^-, \quad \forall v \in \mathbb{V}. \tag{6}$$

Condition (5) is an analogue of Kirchoff's current law for electrical circuits.

2.1.3 Quasi-Stationary Water Flow Model

After some initial transient behavior, the water flow in the pipe may be assumed to get stationary; i.e., the flow is location-independent and we write $Q(t) = \frac{q(x,t)}{A}$ (mass flux is mass flow per unit area), where $A = \pi D^2/4$ is the area of the pipe. Furthermore, the density is assumed constant in space and time; i.e., $\rho(x,t) = \rho$ for $(x,t) \in [0,L] \times \mathbb{R}_+$ and the pressure variable $p(x,t)$ is not coupled to the density via (1) anymore (in particular, water is considered incompressible). The remaining dynamical behavior in the variables $Q(t)$, $P_0(t) = p(0,t)$ and $P_L(t) = p(L,t)$ can be described by the following ODE [2, 5, 6]:

$$\frac{dQ}{dt} + \frac{A}{L}(P_L - P_0) + \frac{c_f Q\,|Q|}{2DA\rho_a} = 0. \tag{7}$$

2.2 Other Network Elements

2.2.1 Reservoir

A reservoir is a node in the water network graph with arbitrary mass flow but with given pressure. For example, if a node v_i is designated as reservoir then pressure at this node will be set as constant.

2.2.2 Valve

A valve is a control element which can be opened or closed and is located at one end of an edge. A closed valve here is modeled as a boundary condition at the corresponding end of the pipe in the form of a prescribed zero flow (instead of the corresponding pressure consistency 6). As an example, assume $e_i, e_j \in \mathbb{E}$ are connected at junction node v, and a valve is located at the end of pipe e_j, then if the valve is open we just have the coupling conditions (5) and (6); in case the valve is closed instead of (6), we have the boundary condition $q_j(L, t) = 0$ and hence, due to (5), also $q_i(0, t) = 0$ (if more than two pipes are incident with v, then there may still be a nonzero flow through the node even if the valve is closed).

3 Analysis of a Simple Water Network

We want to study the water hammer effect on a simple water network consisting of two reservoirs located at nodes v_{R_1} and v_{R_2}, with given pressure $p_{v_{R_1}}$ and $p_{v_{R_2}}$, respectively, and six pipes each of length L. Three valves V_1, V_2 and V_3 are located at the end of pipes 4 and 5 and at the beginning of pipe 6, respectively, as shown in Fig. 2. We assume here that these three valves are opened and closed synchronously; the asynchronous case is ongoing research.

3.1 PDE Mode1

Each pipe is modeled by system of balance laws given by (3) with pressure law (1) and for pipe i and will look as follows,

$$\begin{aligned} \partial_t \rho_i + \partial_x q_i &= 0, \\ \partial_t q_i + \partial_x \left(\frac{q_i^2}{\rho_i} + P(\rho_i) \right) &= -c_{f_i} \frac{q_i \, |q_i|}{2 D_i \rho_i}. \end{aligned} \tag{8}$$

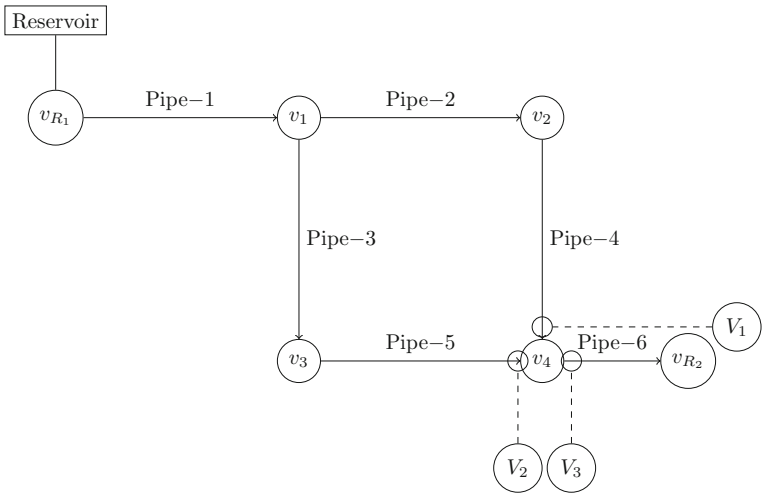


Fig. 2 Simple water network with two reservoirs at nodes v_{R_1} and v_{R_2}, six pipes, and three synchronous valves at node v_4

For the sake of simplicity, we are using identical friction factors and diameters, i.e., $c_{f_i} = c_f, D_i = D, \forall i \in \{1, \cdots, 6\}$. Denote with $P_i(x,t) = P(\rho_i(x,t))$ the pressure in the ith pipe.

In contrast to [7], we present here a water hammer on a network with multiple valves so we need to take more coupling conditions into account: The vertices v_1, v_2, v_3 are coupling vertices and modeled by (6) and (5). At node v_4 valves are present at each incident pipe and it is assumed that they are initially open and simultaneously closed at $t = t_S$, resulting in the time-varying boundary condition:

$$\begin{cases} (5), (6) \quad \text{at} \quad v_4 \qquad t \in (0, t_S), \\ q_{V_1} = q_{V_2} = q_{V_3} = 0, \quad t > t_S. \end{cases} \tag{9}$$

In the following, the pressure at the valves is denoted by $p_{V_1}(t) = P_4(L,t)$, $p_{V_2}(t) = P_5(L,t)$ and $p_{V_3}(t) = P_6(0,t)$, respectively; moreover, $q_{V_1}(t) = q_4(L,t)$, $q_{V_2}(t) = q_5(L,t)$, $q_{V_3}(t) = q_6(0,t)$.

3.2 Switched DAE Framework

The quasi-stationary model (7) together with the corresponding coupling conditions for a setup as shown in Fig. 2 leads to a switched DAE of the form,

$$E_\sigma \dot{x} = A_\sigma x + f + g_\sigma(x), \tag{10}$$

with $x = (Q_1, Q_2, Q_3, Q_4, Q_5, Q_6, P_1, P_2, P_3, P_4, P_{V_1}, P_{V_2}, P_{V_3})^\top$ and

$$\sigma(t) = \begin{cases} 1, & t \in [0, t_S), \quad V_1, V_2, V_3 \text{ open}, \\ 2, & t \geq t_S, \qquad V_1, V_2, V_3 \text{ closed}. \end{cases}$$

The equations of the network when $t \in [0, t_S)$ are given as follows,

$$-\frac{dQ_1}{dt} = c_1(P_{v_1} - P_{R_1}) + c_2 Q_1 \mid Q_1 \mid, \tag{11a}$$

$$-\frac{dQ_2}{dt} = c_1(P_{v_2} - P_{v_1}) + c_2 Q_2 \mid Q_2 \mid, \tag{11b}$$

$$-\frac{dQ_3}{dt} = c_1(P_{v_3} - P_{v_1}) + c_2 Q_3 \mid Q_3 \mid, \tag{11c}$$

$$-\frac{dQ_4}{dt} = c_1(P_{V_1} - P_{v_2}) + c_2 Q_4 \mid Q_4 \mid, \tag{11d}$$

$$-\frac{dQ_5}{dt} = c_1(P_{V_2} - P_{v_3}) + c_2 Q_5 \mid Q_5 \mid, \tag{11e}$$

$$-\frac{dQ_6}{dt} = c_1(P_{R_2} - P_{V_3}) + c_2 Q_6 \mid Q_6 \mid, \tag{11f}$$

$$Q_1 - Q_2 - Q_3 = 0 \tag{11g}$$

$$Q_3 - Q_5 = 0 \tag{11h}$$

$$Q_2 - Q_4 = 0 \tag{11i}$$

$$Q_5 + Q_4 - Q_6 = 0, \tag{11j}$$

$$P_{V_1} - P_4 = 0, \tag{11k}$$

$$P_{V_2} - P_4 = 0, \tag{11l}$$

$$P_{V_3} - P_4 = 0, \tag{11m}$$

where $c_1 = \frac{A}{L} > 0$ and $c_2 = \frac{c_f}{2DA\rho_a} > 0$. For $t \geq t_S$ Eqs. (11k), (11l), (11m) will be replaced by

$$Q_4 = 0, \quad Q_5 = 0, \quad Q_6 = 0. \tag{12}$$

In terms of the nonswitched DAE (10), we have

$$E_p = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_p = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -c_1 & c_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -c_1 & 0 & c_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -c_1 & 0 & 0 & c_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -c_1 & 0 & 0 & c_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -c_1 \\ 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1\text{-}s_p & 0 & 0 & 0 & 0 & -s_p & s_p & 0 & 0 & \\ 0 & 0 & 0 & 0 & 1\text{-}s_p & 0 & 0 & 0 & -s_p & 0 & s_p & 0 & \\ 0 & 0 & 0 & 0 & 0 & 1\text{-}s_p & 0 & 0 & -s_p & 0 & 0 & s_p & \end{bmatrix},$$

$$f = \begin{pmatrix} -P_{v_{R_1}} \\ 0 \\ 0 \\ 0 \\ 0 \\ P_{v_{R_2}} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad g_p(x) = \begin{pmatrix} c_2 Q_1 |Q_1| \\ c_2 Q_2 |Q_2| \\ c_2 Q_3 |Q_3| \\ c_2 Q_4 |Q_4| \\ c_2 Q_5 |Q_5| \\ c_2 Q_6 |Q_6| \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{13}$$

where $p = 1, 2$ and $s_1 = 1$ and $s_2 = 0$.

4 Discussion on Switched DAEs

Note that the switched DAE (10) contains a nonlinear term $g_\sigma(x)$; therefore, the distributional solution framework [10, 11] cannot be applied directly. Nonlinear switched DAEs were investigated in [9], but this approach excludes Dirac impulses in x by definition, because if a Dirac impulse occurs in the solution x of (10) (which we actually desire to capture the water hammer effect) then it is unclear how $g_\sigma(x)$ has to be evaluated in general (e.g., what is the sine of a Dirac impulse). Here we have a special structure which we can write in the following form

$$g(x) = \mathscr{N}\overline{g}(\mathscr{M}x),$$

$$\mathscr{M} = \begin{bmatrix} I_{\{6\times 6\}} & O_{\{6\times 7\}} \end{bmatrix}, \quad \mathscr{N} = \mathscr{M}^\top.$$

$$\overline{g_i}(Q_i) = -c_2 Q_i |Q_i| \quad i = 1, \cdots, 6.$$

This special structure allows us to extend the distributional solution theory from the linear case to the nonlinear case, c.f. [7]. To keep it simple here, consider the individual equation $-\frac{dQ_i}{dt} = c_1(P_{v_i} - P_{v_{R_1}}) + c_2 Q_i \mid Q_i \mid$ and let us denote by $Q_i(t_s^-)$, $Q_i(t_s^+)$ the flow before and after the switching time t_s. When the valves are closed all flows become zero, in particular, $Q_i(t_s^+) = 0$ and since in general $Q_i(t_s^-) \neq 0$ there will be Dirac impulse in $\frac{dQ_i}{dt}$ at the switching time t_s. In fact, the impulse part of $\frac{dQ_i}{dt}$ at t_s is given by

$$\frac{dQ_i}{dt}[t_s] = Q_i(t_s^+) - Q_i(t_s^-)\delta_{t_s} = -Q_i(t_s^-)\delta_s$$

and for $t > t_s$ we have $\frac{dQ_i}{dt} = 0$ because Q_i is identically zero. Altogether we can conclude from (11) together with (12) that for $t \geq t_s$:

$$\begin{aligned}
P_{v_1} &= \frac{1}{c_1} Q_1(t_s^-)\delta_{t_s} + P_{R_1} = \frac{1}{c_1} Q_1(t_s^-)\delta_{t_s} + P_{R_1}, \\
P_{v_2} &= \frac{1}{c_1} Q_2(t_s^-)\delta_{t_s} + P_{v_1} = \frac{1}{c_1}(Q_2(t_s^-) + Q_1(t_s^-))\delta_{t_s} + P_{R_1}, \\
P_{v_3} &= \frac{1}{c_1} Q_3(t_s^-)\delta_{t_s} + P_{v_1} = \frac{1}{c_1}(Q_3(t_s^-) + Q_1(t_s^-))\delta_{t_s} + P_{R_1}, \\
P_{V_1} &= \frac{1}{c_1} Q_4(t_s^-)\delta_{t_s} + P_{v_2} = \frac{1}{c_1}(Q_4(t_s^-) + Q_3(t_s^-) + Q_1(t_s^-))\delta_{t_s} + P_{R_1}, \\
P_{V_2} &= \frac{1}{c_1} Q_5(t_s^-)\delta_{t_s} + P_{v_3} = \frac{1}{c_1}(Q_5(t_s^-) + Q_3(t_s^-) + Q_1(t_s^-))\delta_{t_s} + P_{R_1}, \\
P_{V_3} &= \frac{1}{c_1} Q_6(t_s^-)\delta_{t_s} + P_{R_2} = \frac{1}{c_1} Q_6(t_s^-)\delta_{t_s} + P_{R_2}.
\end{aligned} \tag{14}$$

The coefficient in front of δ_{t_s} determines the impulse length. For $t > t_s$ it is clear that all pressures will settle down as

$$p_{v_1} = p_{v_2} = p_{v_3} = p_{V_1} = p_{V_2} = P_{R_1}, \quad p_{V_3} = P_{R_2}.$$

5 Comparison of both Modeling Approaches

Our focus here is to observe the jump and Dirac impulse in the pressure, due to the instantaneous closure of valves located at V_1, V_2. In particular, we assume that the PDE solution on $[0, t_S)$ is stationary; i.e., $q_i(t, x)$ $i = 1, \cdots, 6$ is approximately constant in time and space (or in other words, when the valves are closed the dynamics in all pipe have approximately settled down). For numerical simulations, we use Flux-Corrected Transport (FCT) scheme and artificial viscosity (<0.25) where solution is nonsmooth. Figure 3 shows the results for the pressure value at V_1 (similar plots result also for the pressure at V_2) over the time interval $[3s, 8s]$ with initial values

$$q_i(0, x) \equiv 0, \quad \rho_i(0, x) \equiv 1 \times 10^3$$

and pipes parameters:

$$\begin{aligned}
&P_a = 1.01 \times 10^6, \quad \beta = \frac{1}{K} = 4 \times 10^{-9}, \quad \rho_a = 1000, \\
&L = 5, \qquad\qquad D = 0.5, \qquad\qquad c_f = 0.02.
\end{aligned}$$

We have chosen a moderate ratio between length and diameter of pipe, so that the water hammer effect is better visible. The parameters P_a, ρ_a and β are physical

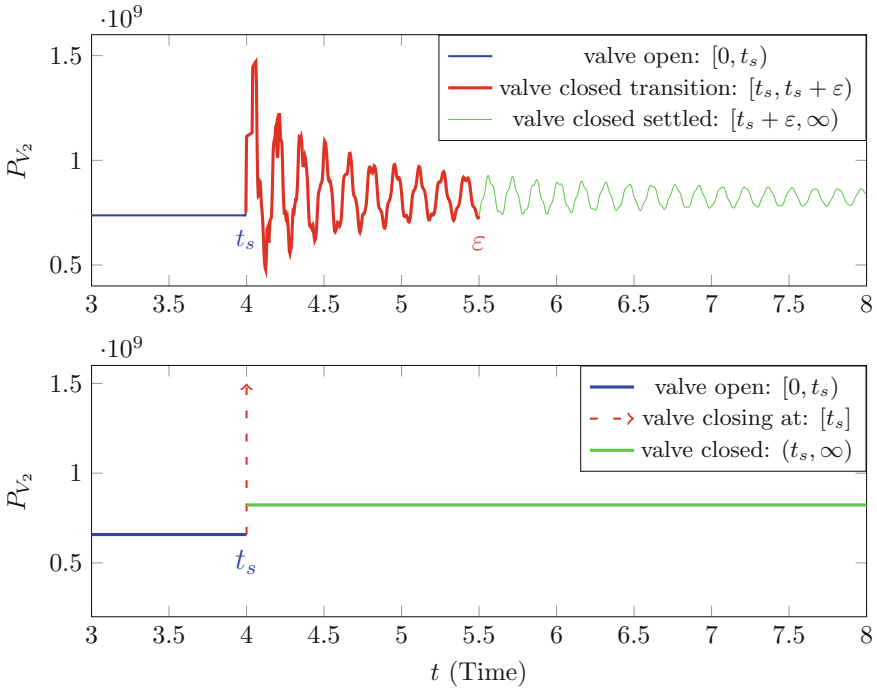


Fig. 3 Comparison of pressure profile PDE models (p_{V_2}) (above) and switched DAE model (P_{V_2}) (below), profile for p_{V_1} is approximately symmetrical

parameters, and c_f is chosen via the so-called moody chart; see, for example, [8]. Figure 3 clearly shows a strong pressure spike just after the switching time $t_S = 4s$; the pressure oscillatoryly settles to a new pressure value say $\overline{P}_R^1$. The same behavior occurs for P_{V_2} which settles to $\overline{P}_R^1$. Instead of running the simulation for a very long time, we just chose a settling time $\varepsilon > 0$ and took the average of the pressures on the interval $(t_S + \varepsilon, T]$ where $T > t_S + \varepsilon$ is our overall simulation time, i.e., with $x = L$

$$\overline{P}_R^1 := \frac{1}{T - (t_S + \varepsilon)} \int_{t_S+\varepsilon}^{T} p_{V_1}(x, t)\, \mathrm{d}t.$$

$$\overline{P}_R^2 := \frac{1}{T - (t_S + \varepsilon)} \int_{t_S+\varepsilon}^{T} p_{V_2}(x, t)\, \mathrm{d}t.$$

With

$$\varepsilon = 1.5, \quad T = 8$$

we obtain

$$\overline{P}_R^1 \approx \overline{P}_R^2 \approx 8.23 \times 10^8.$$

Table 1 Comparison of pressure at valves V_1 and V_2 for PDE and switched DAE model

β	$\overline{P}_R^1$	$\overline{P}_R^2$	$\frac{\lvert\overline{P}_R^1 - P_{v_{R_1}}(t_S^+)\rvert}{P_{v_{R_1}}(t_S^+)}$	$\frac{\lvert\overline{P}_R^2 - P_{v_{R_1}}(t_S^+)\rvert}{P_{v_{R_1}} t_S^+)}$
$15.0 \cdot 10^{-9}$	$8.1613 \cdot 10^8$	$8.2494 \cdot 10^8$	$8.3 \cdot 10^{-03}$	$2.4 \cdot 10^{-03}$
$9.0 \cdot 10^{-9}$	$8.2644 \cdot 10^8$	$8.2419 \cdot 10^8$	$4.2 \cdot 10^{-03}$	$1.4 \cdot 10^{-03}$
$4.0 \cdot 10^{-9}$	$8.2401 \cdot 10^8$	$8.2408 \cdot 10^8$	$1.2 \cdot 10^{-03}$	$1.3 \cdot 10^{-03}$
$5.0 \cdot 10^{-10}$	$8.2329 \cdot 10^8$	$8.2352 \cdot 10^8$	$3.5 \cdot 10^{-04}$	$6.3 \cdot 10^{-04}$
$2.0 \cdot 10^{-10}$	$8.2317 \cdot 10^8$	$8.2348 \cdot 10^8$	$2.6 \cdot 10^{-04}$	$5.8 \cdot 10^{-04}$

The value predicted by the switched DAE solution for $t > t_s$ from (14) is,

$$P_{v_{R_1}}(t_S^+) = P_{R_1} \approx 8.23 \times 10^8.$$

In Table 1, the relative error between $\overline{P}_R^i$, $i = \{1, 2\}$ and $P_{v_{R_1}}(t_S^+)$ is presented for decreasing compressibility coefficients β. In order to compare the peak in P_{V_1}, P_{V_2} just after the valve is closed with the Dirac impulse $P_{V_1}[t_S]$ and $P_{V_2}[t_S]$ in response to the switching time, we recall that a Dirac impulse δ_{t_s} at $t_s > 0$ can be approximated by a sequence of functions $t \mapsto \delta_{t_s}^\varepsilon(t)$ such that $\delta^\varepsilon(t) = 0$ for $t \neq [t_s, t_s + \varepsilon]$ and $\int_{t_s}^{t_s+\varepsilon} \delta_{t_s}^\varepsilon(t)\, \mathrm{d}t = 1$. We therefore make the ansatz for p_{V_1} and P_{V_2},

$$p_{V_1} \approx \overline{P}^{\,\mathrm{imp}_{t_S}^1} \delta^\varepsilon(t) + \overline{P}_R^1, \qquad p_{V_2} \approx \overline{P}^{\,\mathrm{imp}_{t_S}^2} \delta^\varepsilon(t) + \overline{P}_R^2 \quad t \in (t_S, T].$$

Hence, we can approximate the magnitude of the "smoothed-out" Dirac impulse occurring in the PDE model as follows:

$$\overline{P}^{\,\mathrm{imp}_{t_S}^1} := \int_{t_S}^{t_S+\varepsilon} p_{V_1} - \overline{P}_R^1 \, \mathrm{d}t.$$

analogously for p_{V_2},

$$\overline{P}^{\,\mathrm{imp}_{t_S}^2} := \int_{t_S}^{t_S+\varepsilon} p_{V_2} - \overline{P}_R^2 \, \mathrm{d}t.$$

The Dirac impulse induced by the switched DAE is defined from (14), i.e.,

$$P_{V_1}[t_S] = \frac{1}{c_1}(Q_4(t_s^-) + Q_3(t_s^-) + Q_1(t_s^-))\delta_{t_s} =: P^{\,\mathrm{imp}_{t_S}^1} \delta_{t_s},$$

$$P_{V_2}[t_S] = \frac{1}{c_1}(Q_5(t_s^-) + Q_3(t_s^-) + Q_1(t_s^-))\delta_{t_s} =: P^{\,\mathrm{imp}_{t_S}^2} \delta_{t_s}.$$

Table 2 Impulse length comparison

β	$\overline{P}^{\,\mathrm{imp}^1_{t_S}}$	$\overline{P}^{\,\mathrm{imp}^2_{t_S}}$	$P^{\,\mathrm{imp}^1_{t_S}}$	$P^{\,\mathrm{imp}^2_{t_S}}$	$\frac{\lvert \overline{P}^{\,\mathrm{imp}^1_{t_S}} - P^{\,\mathrm{imp}^1_{t_S}} \rvert}{P^{\,\mathrm{imp}^1_{t_S}}}$	$\frac{\lvert \overline{P}^{\,\mathrm{imp}^2_{t_S}} - P^{\,\mathrm{imp}^2_{t_S}} \rvert}{P^{\,\mathrm{imp}^2_{t_S}}}$
$15.0 \cdot 10^{-9}$	$5.7821 \cdot 10^7$	$5.7831 \cdot 10^7$	$5.1137 \cdot 10^7$	$5.1137 \cdot 10^7$	0.1307	0.1309
$9.0 \cdot 10^{-9}$	$3.3944 \cdot 10^7$	$3.3951 \cdot 10^7$	$3.8590 \cdot 10^7$	$3.8590 \cdot 10^7$	0.1204	0.1202
$4.0 \cdot 10^{-9}$	$3.0906 \cdot 10^7$	$3.0918 \cdot 10^7$	$2.8407 \cdot 10^7$	$2.8407 \cdot 10^7$	0.0880	0.0884
$5.0 \cdot 10^{-10}$	$2.0299 \cdot 10^7$	$2.0292 \cdot 10^7$	$2.1096 \cdot 10^7$	$2.1096 \cdot 10^7$	0.0378	0.0381
$2.0 \cdot 10^{-10}$	$1.8450 \cdot 10^7$	$1.8457 \cdot 10^7$	$1.8482 \cdot 10^7$	$1.8482 \cdot 10^7$	0.0017	0.0014

A comparison between $\overline{P}^{\,\mathrm{imp}^1_{t_S}}$ with $P^{\,\mathrm{imp}^1_{t_S}}$ and $\overline{P}^{\,\mathrm{imp}^2_{t_S}}$ with $P^{\,\mathrm{imp}^2_{t_S}}$ for different values of the compressibility coefficient β is presented in Table 2. For large β the approximation is not very accurate; however, for decreasing compressibility the accuracy of the approximation improves.

Similar as for the PDE simulations, we assume that the DAE is stationary before we switch, i.e., $\frac{dQ_i}{dt}(t_s^-) = 0$ for $i \in \{1, \cdots, 6\}$ before closing of the valve. It should be noted that although the compressibility coefficient β does not affect the parameters of the switched DAE model, it does affect the initial value q_0, because this is chosen to match the stationary solution of the balance law (8) considered on $[0, t_S)$ which depends on β.

6 Conclusion

We have presented a switched DAE model for water hammer on a simple setup, which we compared with a compressible nonlinear system of balance laws. With the support of numerical simulations of the PDE model, we justified our conjecture that a switched DAE model is a good approximation for the PDE model with small compressibility coefficient. In future, we will focus on a formal proof of convergence as well as the treatment of larger networks with asynchronously closed valves.

Acknowledgement We are thankful to Jochen Kall for fruitful discussions concerning the PDE simulations of earlier versions of this work.

References

1. S. Adami, X.Y. Hu, N.A. Adams, Simulating three-dimensional turbulence with SPH, in Center for Turbulence Research. Proceedings of the Summer Program 2012 (2012), pp. 177–185
2. M.H. Chaudhry, L. Mays, *Computer Modeling of Free-Surface and Pressurized Flows* (Springer Science & Business Media, 2012)

3. M. Herty, J. Mohring, V. Sachers, A new model for gas flow in pipe networks. Math. Methods Appl. Sci. **33**, 845–855 (2010)
4. J. Izquierdo, R. Pérez, P.L. Iglesias, Mathematical models and methods in the water industry. Math. Comput. Model. **39**, 1353–1374 (2004)
5. L. Jansen, J. Pade, *Global unique solvability for a quasi-stationary water network model*. Preprint series: Institut für Mathematik, Humboldt-Universität zu Berlin (ISSN 0863-0976), 2013-11, (2013)
6. L. Jansen, C. Tischendorf, A unified (P)DAE modeling approach for flow networks, in *Progress in Differential-Algebraic Equations: Deskriptor 2013*, ed. by S. Schöps, A. Bartel, M. Günther, W.E.J. ter Maten, C.P. Müller (Springer, Berlin, Heidelberg, 2014), pp. 127–151
7. J. Kall, R. Kausar, S. Trenn, Modeling water hammers via PDEs and switched DAEs with numerical justification, in Proceedings of IFAC World Congress 2017, Toulouse, (2017)
8. B.E. Larock, R.W. Jeppson, G.Z. Watters, *Hydraulics of Pipeline Systems* (CRC press, 1999)
9. D. Liberzon, S. Trenn, Switched nonlinear differential algebraic equations: solution theory, Lyapunov functions, and stability. Automatica **48**, 954–963 (2012)
10. S. Trenn, Distributional differential algebraic equations. PhD thesis, Institut für Mathematik, Technische Universität Ilmenau, Universitätsverlag Ilmenau, Germany, 2009
11. S. Trenn, Regularity of distributional differential algebraic equations. Math. Control Signals Syst. **21**, 229–264 (2009)
12. S. Trenn, Switched differential algebraic equations, in *Dynamics and Control of Switched Electronic Systems-Advanced Perspectives for Modeling, Simulation and Control of Power Converters*, ch. 6, ed. by F. Vasca, L. Iannelli, (Springer, London, 2012), pp. 189–216
13. E.B. Wylie, V.L. Streeter, *Fluid Transients* (McGraw-Hill International Book Co., New York, 1978)

Stability Criteria for Some System of Delay Differential Equations

Yuya Kiri and Yoshihiro Ueda

Abstract In this paper, we study a system of linear differential equations with interaction effects of delay and derive some necessary and sufficient conditions concerned with the absolutely stable. We had already known many results about the scalar equations. On the other hand, the results of the delay systems are not many because the corresponding characteristic equation is too complicated. Under this situation, we introduce the simple and useful method to get the stability criteria and apply to some general system of delay differential equations.

Keywords System of delay differential equations · Absolute stability

AMS Subject Classifications 34K20 · 34A30

1 Introduction

We analyze the following system of delay differential equations:

$$\begin{aligned} u_1'(t) + a_1 u_1(t) + \alpha_1 u_1(t-\tau_{11}) + \beta_n u_n(t-\tau_{1n}) &= 0, \\ u_2'(t) + a_2 u_2(t) + \alpha_2 u_2(t-\tau_{22}) + \beta_1 u_1(t-\tau_{21}) &= 0, \\ &\vdots \\ u_n'(t) + a_n u_n(t) + \alpha_n u_n(t-\tau_{nn}) + \beta_{n-1} u_{n-1}(t-\tau_{nn-1}) &= 0, \end{aligned} \tag{1}$$

Y. Kiri
Graduate School of Maritime Sciences, Kobe University, Fukaeminamimachi 5-1-1, Higashinada-ku, Kobe 658-0022, Japan
e-mail: 167w304w@stu.kobe-u.ac.jp

Y. Ueda (✉)
Faculty of Maritime Sciences, Kobe University, Fukaeminamimachi 5-1-1, Higashinada-ku, Kobe 658-0022, Japan
e-mail: ueda@maritime.kobe-u.ac.jp

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_10

where a_j, α_j and β_j are complex numbers, and time delays τ_{jk} are nonnegative numbers for $1 \leq j, k \leq n$. The system (1) has interaction delay terms, which appears population models of Lotka–Volterra type, neural network models, and also traffic models (see [1, 3–5, 10–12, 14, 17] and also references therein).

Our purpose of this paper is to derive the stability condition for the system (1). A delay system is called absolutely stable if it is asymptotically stable for all values of the delays and conditionally stable if it is asymptotically stable for all values for the delays in some intervals. It is well known that the stability profile of the system (1) is determined completely by the roots of its associated characteristic equation. Our characteristic function is defined by

$$G(\lambda) = \det \begin{pmatrix} \lambda + \gamma_1 & 0 & \cdots & 0 & \beta_n e^{-\lambda \tau_{1n}} \\ \beta_1 e^{-\lambda \tau_{21}} & \lambda + \gamma_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda + \gamma_{n-1} & 0 \\ 0 & 0 & \cdots & \beta_{n-1} e^{-\lambda \tau_{nn-1}} & \lambda + \gamma_n \end{pmatrix},$$

where we define $\gamma_j := a_j + \alpha_j e^{-\lambda \tau_{jj}}$ for $1 \leq j \leq n$. Then $G(\lambda) = 0$ is a characteristic equation of (1) and $\lambda \in \mathbb{C}$ is a corresponding characteristic root, which is called an eigenvalue. Especially, if we suppose that $\tau_{1n} = \tau_{nn}$ and $\tau_{j+1j} = \tau_{jj}$ for $1 \leq j \leq n-1$, then the function $G(\lambda)$ is represented as $G(\lambda) = \det(\lambda I + A + Be^{-\lambda T})$, where A, B and T are coefficient matrices defined by

$$A = \begin{pmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{pmatrix}, \quad B = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 & \beta_n \\ \beta_1 & \alpha_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \beta_{n-1} & \alpha_n \end{pmatrix}, \quad T = \begin{pmatrix} \tau_{11} & & & \\ & \tau_{22} & & \\ & & \ddots & \\ & & & \tau_{nn} \end{pmatrix},$$

and $e^{-\lambda T}$ denotes the exponential matrix. As mentioned before, it is well known that the solution of (1) is asymptotically stable if and only if all of our eigenvalues lie in the left half of the complex plane (see, e.g., [6, 8, 15]). Therefore, our goal is to construct not only a necessary condition but also some sufficient condition that the real parts of all eigenvalue are negative.

There are a lot of results concerned with the stability for the delay differential equations. In 1950, Hayes in [7] studied a scalar delay differential equation and derived the necessary and sufficient condition for the conditional stability. Bellman and Cooke in [2] also considered the same problem. For the scalar equation, the properties of solutions are well known. However, the analysis for systems of delay differential equations is complicated and there are many open problems. Under this situation, we treat the general system (1) and try to derive a new criterion for the absolute stability.

We have several known results for the system (1) in some specific situation. For example, Liu in [9] and Lu and Wang in [10] studied the system (1) with $n = 2$

and $\alpha_j = 0$ and obtained some stability criteria. Especially, the authors would like to mention the result in Suzuki and Matsunaga [16]. They succeeded to derive the sharp criteria of the conditionally stable for the system (1) with $\alpha_j = 0$. However, we can not apply the method introduced in [16] to the system (1). Thus, we need to find the different approach to get the stability criteria for (1).

On the other hand, Peralta and Ueda in [13] recently introduced the new approach to the system (1). They considered (1) with $n = 2$ and derived the criterion of the absolutely stable by using the energy method. Furthermore, they succeeded to apply their approach to the following system of partial differential equations with delay effect.

$$\begin{aligned}
\partial_t u_1(t,x) + \partial_x u_3(t,x) + a_1 u_1(t,x) + \alpha_1 u_1(t-\tau_1,x) + \beta_2 u_2(t-\tau_2,x) &= 0,\\
\partial_t u_2(t,x) + \partial_x u_4(t,x) + a_2 u_2(t,x) + \alpha_2 u_2(t-\tau_2,x) + \beta_1 u_1(t-\tau_1,x) &= 0,\\
\partial_t u_3(t,x) + \partial_x u_1(t,x) &= 0,\\
\partial_t u_4(t,x) + \partial_x u_2(t,x) &= 0.
\end{aligned}$$

Their result tells us that the stability criteria for the system (1) is applicable to some problems of partial differential equations.

2 Stability Criteria

In this section, we show the new criteria of the absolutely stable for the system (1). Our first main result is stated as follows.

Theorem 1. *If the coefficients of the system* (1) *satisfy the following condition:*

$$\mathrm{Re}(a_j) - |\alpha_j| > 0, \quad 1 \le j \le n, \tag{2}$$

$$\prod_{j=1}^{n} (\mathrm{Re}(a_j) - |\alpha_j|) > \prod_{j=1}^{n} |\beta_j|, \tag{3}$$

then the system (1) *is absolutely stable.*

Remark 1. Theorem 1 includes the known results obtained in Liu [9] and Peralta and Ueda [13].

Proof of Theorem 1. Let $\lambda = x + iy$ with $x, y \in \mathbb{R}$, and assume that $x \ge 0$. Then we derive a contradiction. We can rewrite the characteristic equation $G(\lambda) = 0$ that

$$\prod_{j=1}^{n} (\lambda + a_j + \alpha_j e^{-\lambda \tau_{jj}}) + \beta_n e^{-\lambda \tau_{1n}} \prod_{j=1}^{n-1} \beta_j e^{-\lambda \tau_{j+1\,j}} = 0,$$

and hence, we obtain

$$\prod_{j=1}^{n} |\lambda + a_j + \alpha_j e^{-\lambda \tau_{jj}}| = \prod_{j=1}^{n} |\beta_j e^{-\lambda \tau}|, \tag{4}$$

where we define $\tau := \tau_{1n} + \sum_{j=1}^{n-1} \tau_{j+1\,j}$. Because of $x \geq 0$, we have $e^{-x\tau_{jk}} \leq 1$ for $1 \leq j, k \leq n$. Thus, this yields

$$\prod_{j=1}^{n} |\beta_j e^{-\lambda \tau}| = \prod_{j=1}^{n} |\beta_j| e^{-x\tau} \leq \prod_{j=1}^{n} |\beta_j|. \tag{5}$$

On the other hand, since $x \geq 0$ and (2), we can compute that

$$\begin{aligned} |\lambda + a_j + \alpha_j e^{-\lambda \tau_{jj}}| &\geq |\lambda + a_j| - |\alpha_j e^{-\lambda \tau_{jj}}| \\ &\geq x + \mathrm{Re}(a_j) - |\alpha_j| e^{-x\tau_{jj}} \\ &\geq \mathrm{Re}(a_j) - |\alpha_j| > 0 \end{aligned}$$

for an arbitrary j with $1 \leq j \leq n$. Therefore, we obtain

$$\prod_{j=1}^{n} |\lambda + a_j + \alpha_j e^{-\lambda \tau_{jj}}| \geq \prod_{j=1}^{n} (\mathrm{Re}(a_j) - |\alpha_j|). \tag{6}$$

Finally, substituting (5) and (6) into (4), we arrive at

$$\prod_{j=1}^{n} (\mathrm{Re}(a_j) - |\alpha_j|) \leq \prod_{j=1}^{n} |\beta_j|.$$

However, this inequality is a contradiction under the condition (3). Consequently, we conclude that the real part of the eigenvalues must be negative under the condition (2), (3). This completes the proof. □

We next show some conditions for the coefficients of the system (1), which lead the instability phenomena.

Theorem 2. *If the coefficients of the system* (1) *satisfy the following condition* (i) *or* (ii) *or* (iii)*:*
(i) *In the case* $\beta_j = 0$ *for some* j *with* $1 \leq j \leq n$*, assume that* $\mathrm{Im}(a_k) = 0$ *and* $|a_k| < |\alpha_k|$ *for some* k *with* $1 \leq k \leq n$.
(ii) *In the case* $\beta_j \neq 0$ *and* $\alpha_j = 0$ *for any* j *with* $1 \leq j \leq n$*, assume that* $\mathrm{Im}(a_j) = 0$ *for any* j *with* $1 \leq j \leq n$*, and*

$$\prod_{j=1}^{n} |a_j| < \prod_{j=1}^{n} |\beta_j|. \tag{7}$$

(iii) *In the case* $\beta_j \neq 0$ *for any* j *with* $1 \leq j \leq n$, *assume that* $a_j = a$ *for any* j *with* $1 \leq j \leq n$, *where* a *is an arbitrary fixed real number. Moreover, assume that*

$$\prod_{j=1}^{n} |a + \alpha_j| < \prod_{j=1}^{n} |\beta_j|. \tag{8}$$

Then the system (1) *is not absolutely stable.*

Proof. Throughout this proof, we suppose $\tau_{jk} = \tau$ for $1 \leq j, k \leq n$, where τ is a nonnegative number. Then, our characteristic equation is rewritten as

$$\prod_{j=1}^{n} (\lambda + a_j + \alpha_j e^{-\tau\lambda}) + e^{-n\tau\lambda} \prod_{j=1}^{n} \beta_j = 0. \tag{9}$$

Our purpose of this proof is to obtain the root λ of (9) such that $\mathrm{Re}\lambda > 0$.

We first prove in the case of condition (i). In the case that $\beta_j = 0$ for some j with $1 \leq j \leq n$, our characteristic equation is reduced to $\prod_{j=1}^{n} (\lambda + a_j + \alpha_j e^{-\tau\lambda}) = 0$. Then at least one eigenvalue λ satisfies

$$\lambda + a_k + \alpha_k e^{-\tau\lambda} = 0. \tag{10}$$

for some k with $1 \leq k \leq n$. We consider the solution of (10). We first show that (10) has a purely imaginary root $\lambda = i\omega_0$ with some $\tau = \tau_0$. Indeed, we put $f(\lambda) := \lambda + a_k$ and then obtain $|f(0)| = |a_k|$ and $|f(i\omega)| \to \infty$ as $\omega \to \infty$. Thus, under the assumption $|a_k| < |\alpha_k|$, there is a positive number ω_0 such that $|f(i\omega_0)| = |\alpha_k|$. This tells us that there exists a positive number θ such that $f(i\omega_0) = -\alpha_k e^{-i\theta}$, and hence, $i\omega_0 + a_k + \alpha_k e^{-i\theta} = 0$. Therefore, by the choice of τ_0 such that $\omega_0\tau_0 = \theta + 2\pi m$ with $m \in \mathbb{N}_0$, the pair $(i\omega_0, \tau_0)$ becomes a solution of (10). We note that we can take τ_0 suitably large.

Next, we get the root λ of (10) with $\mathrm{Re}\lambda > 0$. We define $g(\lambda, \tau) := \lambda + a_k + \alpha_k e^{-\tau\lambda}$. Then $g_\lambda(\lambda, \tau) = 1 - \alpha_k \tau e^{-\tau\lambda}$ and we can take $m \in \mathbb{N}_0$ such that

$$g_\lambda(i\omega_0, \tau_0) = 1 + \tau_0(i\omega_0 + a_k) \neq 0.$$

Thus, by the implicit function theorem, we have a solution $\lambda(\tau)$ of (10) around τ_0. Furthermore, the equality

$$\lambda'(\tau_0) = -\frac{g_\tau(i\omega_0, \tau_0)}{g_\lambda(i\omega_0, \tau_0)} = -\frac{i\omega_0(i\omega_0 + a_k)}{1 + \tau_0(i\omega_0 + a_k)}$$

gives us that

$$\mathrm{Re}\lambda'(\tau_0) = \frac{\omega_0(\omega_0 + \mathrm{Im}(a_k))}{(1 + \tau_0 \mathrm{Re}(a_k))^2 + \tau_0^2(\omega_0 + \mathrm{Im}(a_k))^2}.$$

Therefore, we can obtain $\mathrm{Re}\lambda'(\tau_0) \neq 0$ if ω_0 satisfies $\omega_0 + \mathrm{Im}(a_k) \neq 0$, and then there is τ_1 such that $\mathrm{Re}\lambda(\tau_1) > 0$.

We next consider in the case of condition (ii) and (iii). We also show that (9) has a purely imaginary root $\lambda = i\tilde{\omega}_0$ with some $\tau = \tilde{\tau}_0$. We put $\tilde{f}(\lambda) := \prod_{j=1}^{n}(\lambda + a_j + \alpha_j e^{-\tau\lambda})$, and then we have $|\tilde{f}(0)| = \prod_{j=1}^{n}|a_j + \alpha_j|$ and $|\tilde{f}(i\omega)| \to \infty$ as $\omega \to \infty$. Therefore, under the assumption (8) and (7), there exists a positive number $\tilde{\omega}_0$ such that $|\tilde{f}(i\tilde{\omega}_0)| = \prod_{j=1}^{n}|\beta_j|$. This means that there exists a positive number $\tilde{\theta}$ such that $\tilde{f}(i\omega_0) = -e^{-i\tilde{\theta}}\prod_{j=1}^{n}\beta_j$. This fact gives us that

$$\prod_{j=1}^{n}(i\tilde{\omega}_0 + a_j + \alpha_j e^{-i\tau\tilde{\omega}_0}) + e^{-i\tilde{\theta}}\prod_{j=1}^{n}\beta_j = 0.$$

Hence, we can choose $\tilde{\tau}_0$ such that $n\tilde{\tau}_0\tilde{\omega}_0 = \tilde{\theta} + 2\pi\tilde{m}$ with $\tilde{m} \in \mathbb{N}_0$. Then the pair $(i\tilde{\omega}_0, \tilde{\tau}_0)$ satisfies the equation (9). We also remark that we can take $\tilde{\tau}_0$ suitably large.

Based on the above argument, we try to find the root λ of (9) with $\mathrm{Re}\lambda > 0$. We define

$$\tilde{g}(\lambda, \tau) := \prod_{j=1}^{n}(\lambda + a_j + \alpha_j e^{-\tau\lambda}) + e^{-n\tau\lambda}\prod_{j=1}^{n}\beta_j.$$

Then we compute that

$$\begin{aligned}\tilde{g}_\lambda(\lambda, \tau) &= \sum_{k=1}^{n}(1 - \tau\alpha_k e^{-\tau\lambda})\prod_{j\neq k}(\lambda + a_j + \alpha_j e^{-\tau\lambda}) - n\tau e^{-n\tau\lambda}\prod_{j=1}^{n}\beta_j \\ &= \sum_{k=1}^{n}\frac{1 - \tau\alpha_k e^{-\tau\lambda}}{\lambda + a_k + \alpha_k e^{-\tau\lambda}}\{\tilde{g}(\lambda, \tau) - e^{-n\tau\lambda}\prod_{j=1}^{n}\beta_j\} - n\tau e^{-n\tau\lambda}\prod_{j=1}^{n}\beta_j.\end{aligned}$$

Therefore

$$\begin{aligned}\tilde{g}_\lambda(i\tilde{\omega}_0, \tilde{\tau}_0) &= -\{\sum_{k=1}^{n}\frac{1 - \tilde{\tau}_0\alpha_k e^{-i\tilde{\tau}_0\tilde{\omega}_0}}{i\tilde{\omega}_0 + a_k + \alpha_k e^{-i\tilde{\tau}_0\tilde{\omega}_0}} + n\tilde{\tau}_0\}e^{-in\tilde{\tau}_0\tilde{\omega}}\prod_{j=1}^{n}\beta_j \\ &= -\Phi_n e^{-in\tilde{\tau}_0\tilde{\omega}_0}\prod_{j=1}^{n}\beta_j,\end{aligned}$$

where

$$\Phi_n := \sum_{k=1}^{n}\frac{1 + \tilde{\tau}_0(i\tilde{\omega}_0 + a_k)}{i\tilde{\omega}_0 + a_k + \alpha_k e^{-i\tilde{\tau}_0\tilde{\omega}_0}},$$

and we can take $\tilde{m} \in \mathbb{N}_0$ which satisfies $\tilde{g}_\lambda(i\tilde{\omega}_0, \tilde{\tau}_0) \neq 0$. Thus, using the implicit function theorem, we get a solution $\lambda(\tau)$ of (9) around $\tilde{\tau}_0$. Moreover, we calculate that

$$\tilde{g}_\tau(i\tilde{\omega}_0, \tilde{\tau}_0) = -\Psi_n e^{-in\tilde{\tau}_0\tilde{\omega}_0} \prod_{j=1}^{n} \beta_j,$$

where

$$\Psi_n := \sum_{k=1}^{n} \frac{i\tilde{\omega}_0(i\tilde{\omega}_0 + a_k)}{i\tilde{\omega}_0 + a_k + \alpha_k e^{-i\tilde{\tau}_0\tilde{\omega}_0}},$$

and then

$$\mathrm{Re}\lambda'(\tilde{\tau}_0) = -\mathrm{Re}\Big(\frac{\Psi_n}{\Phi_n}\Big) = -\frac{1}{|\Phi_n|^2}\mathrm{Re}(\Phi_n\bar{\Psi}_n).$$

Here we can compute

$$\begin{aligned}\Phi_n\bar{\Psi}_n = \sum_{j,k=1}^{n} &\frac{-i\tilde{\omega}_0(\overline{i\tilde{\omega}_0 + a_k})}{(i\tilde{\omega}_0 + a_j + \alpha_j e^{-i\tilde{\tau}_0\tilde{\omega}_0})(\overline{i\tilde{\omega}_0 + a_k + \alpha_k e^{-i\tilde{\tau}_0\tilde{\omega}_0}})} \\ &- i\tilde{\tau}_0\tilde{\omega}_0\Big|\sum_{j=1}^{n} \frac{i\tilde{\omega}_0 + a_j}{i\tilde{\omega}_0 + a_j + \alpha_j e^{-i\tilde{\tau}_0\tilde{\omega}_0}}\Big|^2.\end{aligned} \tag{11}$$

Now we consider in the case of condition (ii). Then we get from (11) that

$$\Phi_n\bar{\Psi}_n = \sum_{j=1}^{n} \frac{-in\tilde{\omega}_0}{i\tilde{\omega}_0 + a_j} - in^2\tilde{\tau}_0\tilde{\omega}_0,$$

and hence, we obtain

$$\mathrm{Re}(\Phi_n\bar{\Psi}_n) = \sum_{j=1}^{n} \frac{-n\tilde{\omega}_0(\tilde{\omega}_0 + \mathrm{Im}(a_j))}{\mathrm{Re}(a_j)^2 + (\tilde{\omega}_0 + \mathrm{Im}(a_j))^2}.$$

Consequently, we get $\mathrm{Re}\lambda'(\tilde{\tau}_0) \neq 0$ if $\mathrm{Im}(a_j) = 0$ for $1 \leq j \leq n$, and then there exists $\tilde{\tau}_1$ such that $\mathrm{Re}\lambda(\tilde{\tau}_1) > 0$.

Finally, we consider in the case of condition (iii). By the assumption $a_j = a$ for $1 \leq j \leq n$, we get

$$\Phi_n\bar{\Psi}_n = \frac{-\tilde{\omega}_0(\tilde{\omega}_0 + \mathrm{Im}(a) + i\mathrm{Re}(a))}{|1 + \tilde{\tau}_0(i\tilde{\omega}_0 + a)|^2}|\Phi_n|^2 - i\frac{\tilde{\tau}_0}{\tilde{\omega}_0}|\Psi_n|^2.$$

Thus, we obtain $\mathrm{Re}\lambda'(\tilde{\tau}_0) \neq 0$ if $\tilde{\omega}_0$ satisfies $\tilde{\omega}_0 + \mathrm{Im}(a) \neq 0$, and then there is $\tilde{\tau}_1$ such that $\mathrm{Re}\lambda(\tilde{\tau}_1) > 0$. Hence, we complete the proof of Theorem 2.

At last of this article, we make a summary that Theorem 1 is very useful to check that the concrete model of delay equation is absolutely stable or not. However, Theorem 2 is not enough to conclude the optimality of the stability condition in Theorem 1. This situation makes one open problem.

Acknowledgements The second author is partially supported by Grant-in-Aid for Young Scientists (B) No. 25800078 from Japan Society for the Promotion of Science.

References

1. P. Baldi, A. Atiya, How delays affect neural dynamics and learning. IEEE Trans. Neural Netw. **5**, 612–621 (1994)
2. R. Bellman, K.L. Cooke, *Differential-Difference Equations* (Academic Press, New York, 1963)
3. S.A. Campbell, Stability and bifurcation of a simple neural network with multiple time delays. Fields Inst. Commun. **21**, 65–79 (1999)
4. L.E. El'sgol'ts, S.B. Norkin, *Introduction to the Theory and Application of Differential Equations with Deviating Arguments*. Mathematics in Science and Engineering, vol. 105 (Academic Press, New York, 1973)
5. K. Gopalsamy, Harmless delays in a periodic ecosystem. J. Aust. Math. Soc. B **25**(3), 349–365 (1984)
6. J.K. Hale, *Theory of Functional Differential Equations*. Applied Mathematics Sciences, vol. 3 (Springer, New York, 1977)
7. N.D. Hayes, Roots of the transcendental equation associated with a certain difference-differential equation. J. Lond. Math. Soc. **25**, 226–232 (1950)
8. J.K. Hale, S.M.V. Lunel, *Introduction to Functional Differential Equations*. Applied Mathematical Sciences, vol. 99 (Springer, New York, 1993)
9. L.H. Lu, Numerical stability of the θ-methods for systems of differential equations with several delay terms. J. Comput. Appl. Math. **34**(3), 291–304 (1991)
10. Z. Lu, W. Wang, Global stability for two-species Lotka-Volterra systems with delay. J. Math. Anal. Appl. **208**(1), 277–280 (1997)
11. G. Orosz, G. Stépán, Subcritical Hopf bifurcations in a car-following model with reaction-time delay. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **462**(2073), 2643–2670 (2006)
12. G. Orosz, R.E. Wilson, G. Stépán, Traffic jams: dynamics and control. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **368**(1928), 4455–4479 (2010)
13. G. Peralta, Y. Ueda, Stability condition for a system of delay-differential equations and its application
14. S. Ruan, Absolute stability, conditional stability and bifurcation in Kolmogorov-type predator-pray systems with discrete delays. Q. Appl. Math. **59**(1), 159–173 (2001)
15. G. Stépán, *Retarded Dynamical Systems: Stability and Characteristic Functions*. Pitman Research Notes in Mathematics Series, vol. 210 (Longman Scientific & Technical, Harlow; copublished in the United States with John Wiley & Sons, Inc., New York, 1989)
16. M. Suzuki, H. Mastunaga, Stability criteria for a class of linear differential equations with off-diagonal delays. Discret. Contin. Dyn. Syst. **24**(4), 1381–1391 (2009)
17. X.P. Yan, Y.D. Chu, Stability and bifurcation analysis for a delayed Lotka–Volterra predator-prey system. J. Comput. Appl. Math. **196**(1), 198–210 (2006)

Bound-Preserving Reconstruction of Tensor Quantities for Remap in ALE Fluid Dynamics

Matej Klima, Milan Kucharik, Mikhail Shashkov and Jan Velechovsky

Abstract We analyze several new and existing approaches for limiting tensor quantities in the context of deviatoric stress remapping in an ALE numerical simulation of elastic flow. Remapping and limiting of the tensor component-by-component are shown to violate radial symmetry of derived variables such as elastic energy or force. Therefore, we have extended the symmetry-preserving Vector Image Polygon algorithm, originally designed for limiting vector variables. This limiter constrains the vector (in our case a vector of independent tensor components) within the convex hull formed by the vectors from surrounding cells—an equivalent of the discrete maximum principle in scalar variables. We compare this method with a limiter designed specifically for deviatoric stress limiting which aims to constrain the J_2 invariant that is proportional to the specific elastic energy and scale the tensor accordingly. We also propose a method which involves remapping and limiting the J_2 invariant

M. Klima (✉) · M. Kucharik
Faculty of Nuclear Sciences and Physical Engineering,
Department of Physical Electronics, Czech Technical University in Prague,
Brehova 7, 115 19 Praha, Czech Republic
e-mail: klimamat@fjfi.cvut.cz

M. Kucharik
e-mail: kucharik@newton.fjfi.cvut.cz

M. Shashkov
Los Alamos National Laboratory, XCP-4 Group, MS-F644, P.O. Box 1663,
Los Alamos, NM 87545, USA
e-mail: shashkov@lanl.gov

J. Velechovsky
Los Alamos National Laboratory, T-3 Group, MS-B216, P.O. Box 1663,
Los Alamos, NM 87545, USA
e-mail: jan@lanl.gov

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_11

independently using known scalar techniques. The deviatoric stress tensor is then scaled to match this remapped invariant, which guarantees conservation in terms of elastic energy.

Keywords ALE · Remapping · Limiter · Stress tensor · Symmetry preservation

MSC2010: 65D05 · 65M99 · 74B05

1 Introduction

The reconstruction of material quantities from discrete values defined on a computational mesh is a key part of high-order numerical schemes for fluid dynamics. For demanding simulations where both high-pressure gradients and shear flows occur simultaneously, such as in the field of laser–plasma interactions, the Arbitrary Lagrangian–Eulerian (ALE) framework [3, 6] is often used. As its name suggests, it allows for arbitrary movement of the computational mesh. We focus on the indirect ALE formulation which utilizes pure Lagrangian steps [2] advancing the solution and mesh in time.

If needed, mesh smoothing and subsequent quantity remapping are performed to preserve sufficient geometric quality of the mesh. In the remapping step, the monotonicity of the reconstructed fields is often ensured by slope limiters. These have been formulated originally for scalar and later extended to vector quantities. However, reconstructing and limiting of tensor variables are still a relatively unexplored territory with only a few specialized methods that have been proposed recently [11]. The design principles of such methods are objectivity (frame invariance) and preservation of bounds and tensor invariants.

The simplest approach presented in this paper involves piecewise linear reconstruction of the tensor components using a known limiter scheme for scalar variables (such as the Barth–Jespersen limiter [1]) applied component-wise. This method is known to violate the solution symmetry for radially symmetric problems. Our alternative scheme is inspired by the Vector Image Polygon limiter [5], constraining the tensor components within a convex hull constructed in the tensor component space.

Another approach was proposed specifically for stress tensor limiting [11], constraining its second invariant, and scaling the tensor in a way that is frame invariant and preserves local extrema and symmetry. We propose an extension of this method, based on limiting/remapping the tensor components and the J_2 invariant separately. The remapped tensor is then scaled to match the remapped J_2 value—as it is proportional to the elastic energy density, which implies that the conservation of energy will not be violated.

Properties of the particular methods are demonstrated on a selected numerical test of static remapping of a tensor quantity with a radially symmetric distribution.

2 Governing Equations—the Lagrangian Step

We solve the time-dependent Euler equations in Lagrangian form, extended to a general elastic–plastic continuum [10, 13]:

$$\rho \frac{dv}{dt} - \nabla \cdot \mathbf{u} = 0, \tag{1}$$

$$\rho \frac{d\mathbf{u}}{dt} - \nabla \cdot \sigma = 0, \tag{2}$$

$$\rho \frac{dE}{dt} - \nabla \cdot (\sigma \mathbf{u}) = 0, \tag{3}$$

where ρ represents density, $v = \frac{1}{\rho}$ specific volume, σ the Cauchy stress tensor, $\mathbf{u}$ velocity vector, and $E = \varepsilon + \frac{1}{2}\mathbf{u}^2$ specific total energy with ε being the specific internal energy. The Cauchy stress tensor is symmetric and can be decomposed as

$$\sigma = -p\,I + S, \tag{4}$$

where p is hydrodynamic pressure, I the identity matrix, and S the deviatoric stress tensor. For the closure of the system, the Mie–Gruneisen equation of state [8] is used.

The system is solved by a numerical scheme based on [10]. A compatible discretization [2] is used in which the movement of the computational mesh is calculated nodal force vectors while the discrete stress tensor is defined in cell centers. The cell-to-node subzonal forces are calculated first as

$$\mathbf{F}_{p,c} = l_1\,\sigma_c\,\mathbf{n}_1 + l_2\,\sigma_c\,\mathbf{n}_2, \tag{5}$$

and then combined to yield the total nodal force,

$$\mathbf{F}_p = \sum_{c \in N(p)} \mathbf{F}_{p,c}, \tag{6}$$

where $N(p)$ is a set containing all neighboring cells of node p. l_1, l_2 is equal to the half of the respective cell edge length, and $\mathbf{n}_1, \mathbf{n}_2$ are the unit normal vectors. See Fig. 1 for details.

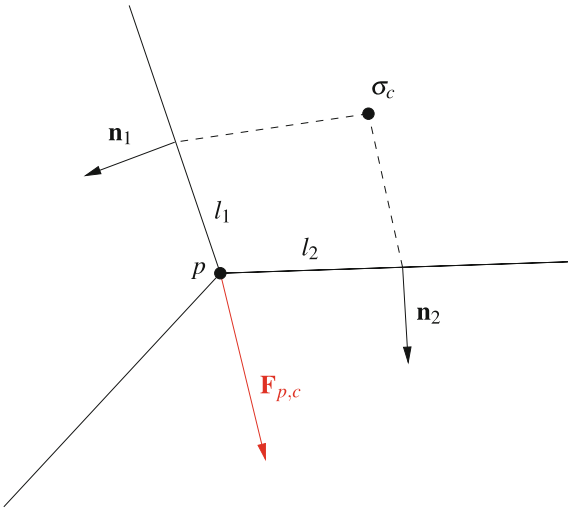


Fig. 1 Cell c to node p subzonal elastic force $\mathbf{F}_{p,c}$ construction with half-edge lengths l_1, l_2, normals $\mathbf{n}_1, \mathbf{n}_2$ and the cell-centered stress tensor σ_c

3 Remapping of the Deviatoric Stress Tensor

In this section, we propose several methods for remapping the deviatoric stress tensor. In two-dimensional planar geometry, it has the following shape:

$$S = \begin{pmatrix} S_{xx} & S_{xy} & 0 \\ S_{xy} & S_{yy} & 0 \\ 0 & 0 & -(S_{xx} + S_{yy}) \end{pmatrix}. \tag{7}$$

It is necessary to use the full 3×3 representation [10], where the third diagonal term enforces the deviatoric property $\mathrm{tr}\,(S) = 0$. The characteristic equation of the tensor defines the three invariants:

$$\lambda^3 + J_1\lambda^2 + J_2\lambda + J_3 = 0, \tag{8}$$

$$J_1 = \mathrm{tr}\,(S) = 0,\ J_2 = \frac{1}{2}(S : S) = \frac{1}{2}\mathrm{tr}(S^T S),\ J_3 = \det(S). \tag{9}$$

We are interested especially in the J_2 invariant, as it is proportional to the elastic energy density:

$$e_{\text{elast.}} = \frac{1}{2\mu} J_2, \tag{10}$$

where μ is the shear modulus, a material constant.

There are several properties, we would like the remapper to have. The first is preservation of bounds—for a tensor variable this is not readily defined but we can use one of the derived quantities. In the case of deviatoric stress, our remapper

should preserve the bounds of elastic energy [11]. The total elastic energy should also be conserved. We propose an extra criterion of preserving the elastic force radial symmetry. As a vector quantity, the elastic forces (6) are easier to analyze.

In the following subsections we describe several approaches to deviatoric stress remapping.

3.1 Component-Wise Remap and Limiting of Tensor S

The simplest way of remapping the deviatoric stress tensor is to treat the individual components of the tensor as independent scalar variables. The tensor components are remapped similarly to average pressure, where the pressure–volume work is remapped:

$$\tilde{S}^c \tilde{V}^c = S^c V^c + \sum_{c' \in \mathcal{N}(c)} \left(F^S_{c' \cap \tilde{c}} - F^S_{c \cap \tilde{c}'} \right), \quad F^S_{c' \cap \tilde{c}} = \iint_{c' \cap \tilde{c}} S(\mathbf{x}) dV. \tag{11}$$

The tensor reconstruction $S(\mathbf{x})$ can be expressed in terms of the independent tensor components as:

$$\begin{pmatrix} S_{xx} \\ S_{xy} \\ S_{yy} \end{pmatrix} (\mathbf{x}) = \begin{pmatrix} S^c_{xx} \\ S^c_{xy} \\ S^c_{yy} \end{pmatrix} + (\mathbf{x} - \mathbf{x_c}) \begin{pmatrix} \psi_{xx} \nabla S_{xx} \\ \psi_{xy} \nabla S_{xy} \\ \psi_{yy} \nabla S_{yy} \end{pmatrix}, \tag{12}$$

where ∇S is the tensor gradient and its components can be obtained using the least squares optimization [4, 7] on all neighboring cells. $\mathbf{x_c}$ is the geometric centroid of the computational cell and ψ_{xx} is a scalar limiting coefficient. In particular, the Barth–Jespersen procedure [1, 7] is used here:

$$\psi^p_{xx} = \begin{cases} \min\left(\frac{S^{\max}_{xx} - S^c_{xx}}{S^p_{xx} - S^c_{xx}}, 1.0 \right) & \text{if} \quad S^p_{xx} > S^c_{xx} \\ \min\left(\frac{S^{\min}_{xx} - S^c_{xx}}{S^p_{xx} - S^c_{xx}}, 1.0 \right) & \text{if} \quad S^p_{xx} < S^c_{xx} \\ 1.0 & \text{otherwise} \end{cases}, \tag{13}$$

$$\psi_{xx} = \min_{p \in \mathcal{P}(c)} \left(\psi^p_{xx} \right), \quad S^p_{xx} = S^c_{xx} + (\mathbf{x}_p - \mathbf{x}_c) \nabla S^c_{xx}, \tag{14}$$

where $\mathcal{P}(c)$ is the set of all vertices of the cell c, $\mathbf{x}_p$ is the position of the vertex p, and S^p_{xx} is the unlimited reconstruction in the corresponding point. $S^{\max}_{xx}$ and $S^{\min}_{xx}$ are tensor component maximum and minimum calculated on the same 9-cell stencil as is used for the gradient computation. The same procedure is also used for the other independent tensor components S_{xy} and S_{yy}.

3.2 VIP Limiter for Tensors

The component-wise limiting approach is simple to implement but it also has several disadvantages. For simplicity, let us apply this method on vectors first. Due to the independent limiting of vector components, vectors can be unnecessarily rotated, distorting the directional symmetry. In the tensor case, this can manifest as deformation of the tensor principal directions. Component-wise limiting also does not guarantee the validity of the discrete maximum principle for vector magnitudes. This is more complex for tensors, but similarly the J_2 invariant monotonicity is not preserved [11].

To solve these issues, the Vector Image Polygon limiter was proposed [5] and adapted for the vector magnitude monotonicity problem [12]. It constrains the reconstructed vector within the convex hull formed by the values in the neighboring cells. An example and comparison with the component-wise method is shown in Fig. 2—an extreme case is displayed, where all values reconstructed in vertices lie outside the convex hull in the tensor component space, but are considered valid by the component-wise limiter.

We propose applying this method as a scalar slope limiter using (S_{xx}, S_{xy}, S_{yy}) as the 3D tensor component space:

$$S(\mathbf{x}) = S^c + \psi_{VIP}(\mathbf{x} - \mathbf{x}_c)\nabla S^c, \tag{15}$$

$$\psi_{VIP} = \min_{p \in \mathscr{P}(c)} \left(\frac{\|S^{VIP} - S^c\|}{\|S^p - S^c\|} \right), \; S^p = S^c + (\mathbf{x}_p - \mathbf{x}_c)\nabla S. \tag{16}$$

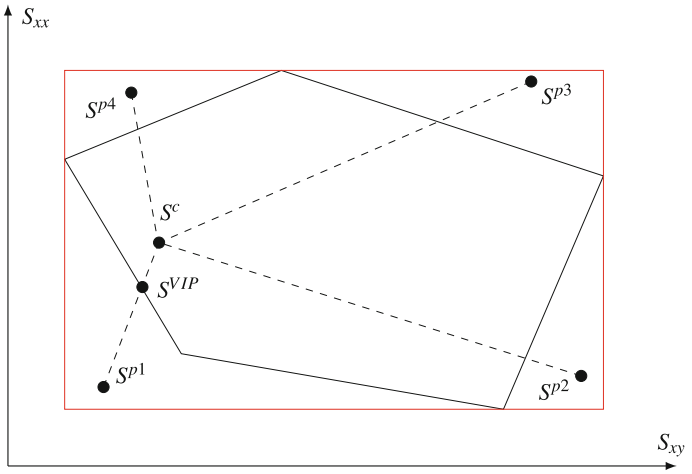


Fig. 2 A simplified 2D schematic of the VIP algorithm for tensors with unlimited reconstructed values in cell nodes $S^{p1} \ldots S^{p4}$, cell-centered average value S^c and the closest limited value S^{VIP}, compared with component-wise limiting set (red)

The construction of the convex hull in three dimensions with few points is relatively simple, but the limiter requires a robust intersection algorithm as the hull often degenerates to a planar case which needs to be treated separately. In the non-degenerate 3D case, an iterative line–polyhedron intersection is calculated.

The main disadvantages of this algorithm are complexity and more diffusion compared to component-wise limiting. It noticeably reduces the overall order of accuracy below second-order.

3.3 J_2 Invariant Scaling Limiter

This limiter was formulated specifically for the deviatoric stress limiting in [11]. It is based on an assumption that the monotonicity of J_2 invariant (proportional to elastic energy) is more important than monotonicity of tensor components. The monotonicity condition can be described as:

$$J_2^{\min} - J_2^c \leq J_2^p - J_2^c \leq J_2^{\max} - J_2^c \quad \forall p \in \mathscr{P}(c), \tag{17}$$

where $J_2^{\min}$ and $J_2^{\max}$ are again determined on the set of neighboring cells. Single scaling factor is then used for the reconstructed tensor:

$$\psi = \sqrt{\psi_{J_2} + (1 - \psi_{J_2})\frac{J_2^c}{J_2^p}}, \quad \psi_{J_2} = \textit{Barth-Jespersen}\left[J_2(S)\right], \tag{18}$$

$$S(\mathbf{x}) = \psi\left(S^c + (\mathbf{x} - \mathbf{x_c})\nabla S\right). \tag{19}$$

This approach is relatively fast, simple to implement, and the monotonicity of J_2 is guaranteed by design. However, as it has been developed in a different context, its effect on elastic forces has not been investigated in literature previously.

3.4 J_2 Invariant-Based Scalar Slope Limiter

An alternative to previous approach is also presented in [11]. The design goals are similar, but it uses the formalism of a slope limiter:

$$S(\mathbf{x}) = S^c + (\mathbf{x} - \mathbf{x_c})\psi\nabla S, \quad \psi = \textit{Barth-Jespersen}\left[\sqrt{J_2(S)}\right] \tag{20}$$

Our test show that its behavior is almost indistinguishable from the previous case while being slightly more resource intensive.

3.5 Independent Remap of S and J_2

The previously described algorithms were intended mainly to reduce symmetry distortion by using tensor-specific limiting techniques. Here we propose a different approach for deviatoric stress remapping—the J_2 invariant is remapped independently of S:

$$\tilde{J}_2^c \tilde{V}^c = J_2^c V^c + \sum_{c' \in \mathcal{N}(c)} F^{J2}_{c' \cap \tilde{c}} - F^{J2}_{c \cap \tilde{c}'}. \tag{21}$$

A scalar limiter is then used in the J_2 reconstruction. This is equivalent to remapping the elastic energy density (10) which is a conservative quantity. Then, S is remapped component-wise (11) without limiting and the resulting tensor is scaled by multiplying by the ratio of the remapped invariant $\tilde{J}_2^c$ and $J_2(\tilde{S})$ calculated from remapped S:

$$\tilde{\tilde{S}} = \tilde{S} \sqrt{\frac{\tilde{J}_2^c}{J_2(\tilde{S})}}. \tag{22}$$

This formulation guarantees conservation of total elastic energy as well as its monotonicity. The component-wise remap of S primarily determines the principal directions of the tensor (not its J_2 invariant) and according to our observation, low-order (donor) remapping is sufficient here with negligible impacts on the overall accuracy.

4 Numerical Results—Cyclic Remapping of a Nonlinear Radial Distribution of the Deviatoric Stress Tensor

We demonstrate the performance of different deviatoric stress remapping methods on a simple static test case—a distribution of the stress tensor is initialized and repeatedly remapped without any influence of the hydrodynamics. The artificial rezoning motion was inspired by the "tensor-product" cyclic rezoning [9] and is defined as follows:

$$r_n = r_l + \left[\frac{r_0 - r_l}{r_r - r_l}(1 - d_n) + \left(\frac{r_0 - r_l}{r_r - r_l} \right)^3 d_n \right] (r_r - r_l), \tag{23}$$

$$\varphi_n = \varphi_0,\ d_n = \frac{1}{2} \sin\left(\frac{\pi n}{n_{\max}} \right),\ r_l = 0.1,\ r_r = 1.0,$$

where r_0, φ_0 are the initial nodal polar coordinates, n is the current remapping step, and $n_{\max}$ is the total number of remapping steps. This represents a cyclic movement of nodes in the radial direction where the initial ($n = 0$) and final ($n = n_{\max}$) grids are identical; see Fig. 3.

On such grid, the deviatoric stress tensor is initialized as follows:

$$S_{i,j} = \begin{cases} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \text{for } i \leq \frac{n_i}{2} \\ \begin{pmatrix} -\cos(2\varphi_{i,j}) & -\sin(2\varphi_{i,j}) \\ -\sin(2\varphi_{i,j}) & \cos(2\varphi_{i,j}) \end{pmatrix} & \text{for } i > \frac{n_i}{2} \end{cases}, \tag{24}$$

where i, j are the radial and axial indices, and n_i is the number of cells in the radial direction. This distribution generates a radial discontinuity with a peak in the elastic force (as shown in Fig. 4) and piecewise constant J_2 invariant distribution.

A comparison of the final elastic force distribution after the cyclic remapping is shown in Fig. 5. If no limiter is used, there are visible undershoots of the remapped quantity. However, limiting components independently does not solve the problem and adds asymmetry. The VIP limiter performs well, but is slightly more diffusive than the other alternatives. Our approach to tensor remap seems to shift the position of the peak, but preserves monotonicity of forces perfectly in this test.

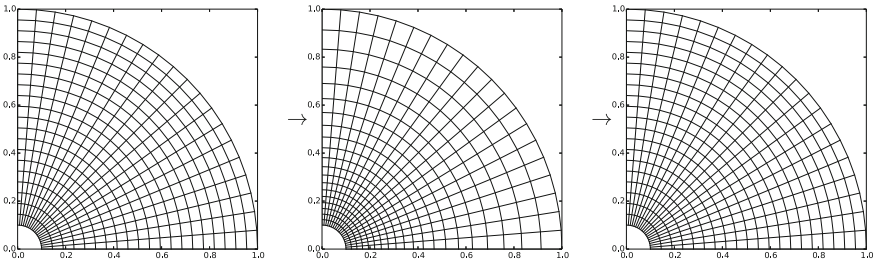


Fig. 3 Polar grid sequence—different steps of the cyclic rezoning movement, 20 × 20 mesh

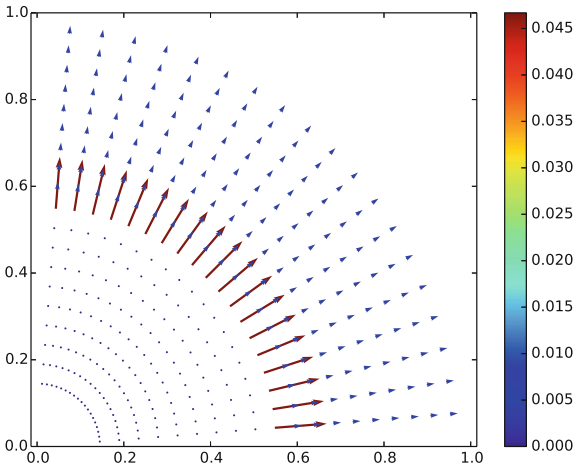


Fig. 4 Initial elastic force distribution shown in the internal nodes of the 20 × 20 mesh, force vectors are colored according to magnitude

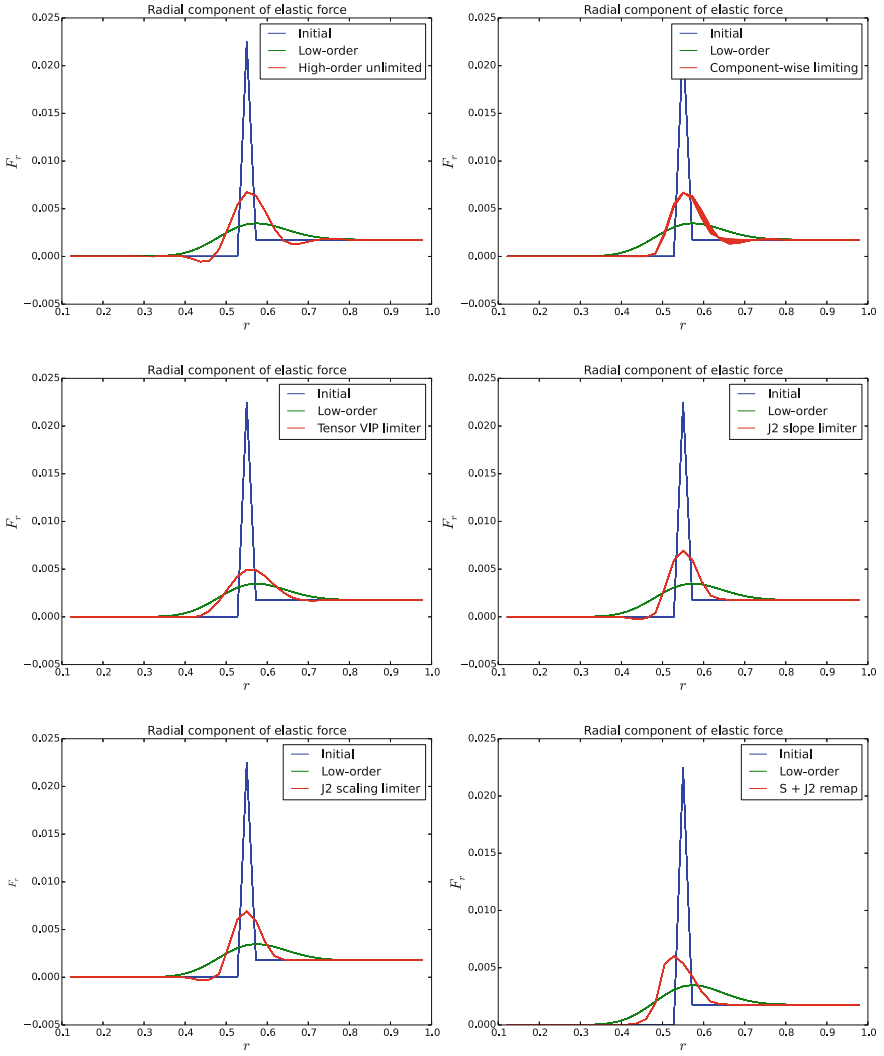


Fig. 5 Radial component of the total nodal elastic force produced by the deviatoric stress tensor after cyclic remapping, compared by different remapping methods. 40×40 mesh, $n_{\max} = 80$

Figure 6 shows the radial distribution of the J_2 invariant. Here, the asymmetry generated by component-wise limiting is even stronger. The VIP limiter does not guarantee monotonicity of the elastic energy. All other methods are based on constraining the J_2 invariant directly and therefore are successful in this task. Our remapping method preserves symmetry, does not violate energy conservation, and is the only one which limits both J_2 and elastic forces correctly in this idealized test.

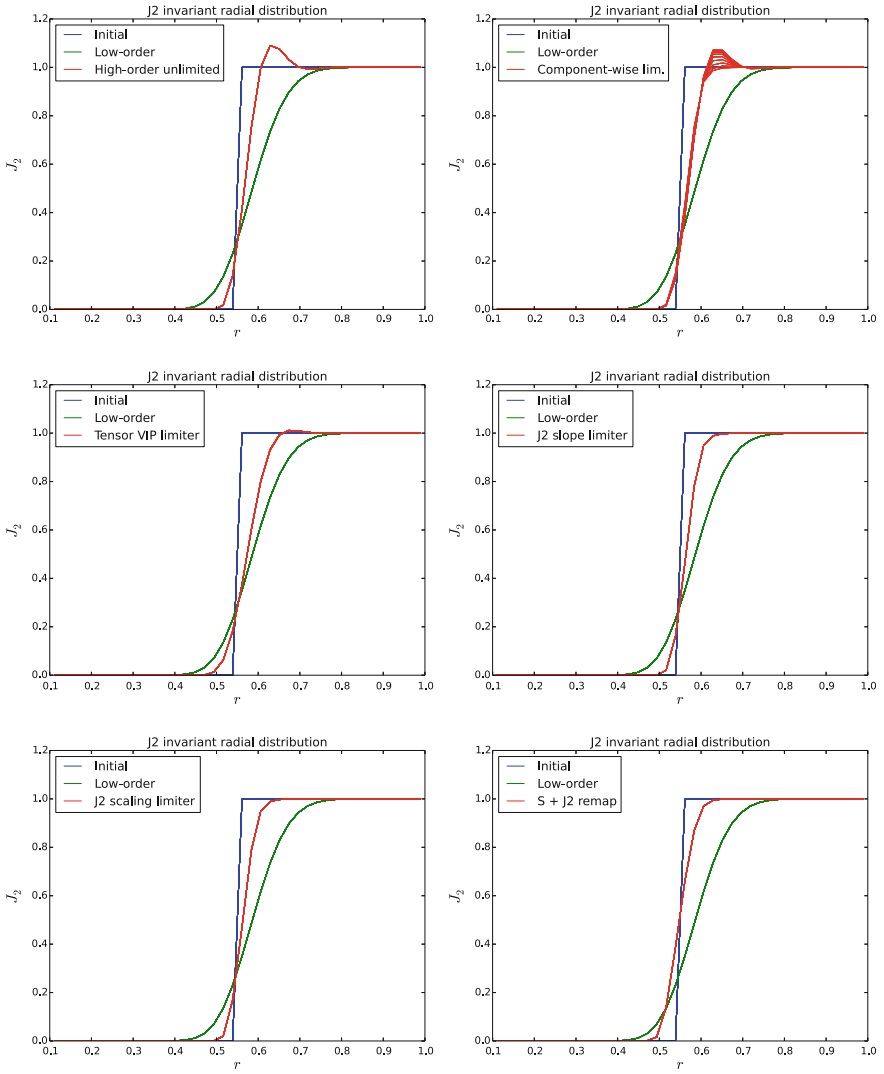


Fig. 6 Radial distribution of the J_2 invariant of the deviatoric stress tensor after cyclic remapping, compared by different remapping methods. 40×40 mesh, $n_{\max} = 80$

Table 1 illustrates the computational efficiency of all limiting methods for the cyclic remapping case. We can see that most high-order methods with limiting perform similarly, except for the VIP-based method which is much more expensive.

Table 1 Simulation times for deviatoric stress tensor cyclic remapping compared by different remapping methods, 40×40 mesh, $n_{\max} = 80$, run single-thread on an Intel^{TM} Core i5-4300M processor

Low-order	High-order unlim.	Comp.-wise lim.	VIP	J_2 scaling	J_2 slope lim.	$S + J_2$ remap
$0.3s$	$1.4s$	$1.7s$	$6.6s$	$1.7s$	$1.9s$	$1.7s$

5 Conclusion

Several methods of reconstructing a tensor quantity are proposed in this paper, focusing on the flux-form remap of the deviatoric stress in the context of an indirect ALE simulation. We show that using a scalar reconstruction method for each independent tensor component does not guarantee the monotonicity preservation of elastic forces and energy while distorting the symmetry of the solution severely.

We have implemented a modified Vector Image Polygon limiter for tensors, showing the viability of this approach. It is, however, a resource-intensive and complex method that produces more diffusive results. Specialized methods constraining the second invariant of the tensor are much faster and less diffusive but also reduce the force overshoots less.

We propose a new method for remapping the deviatoric stress, where the tensor and its second invariant are remapped independently. The tensor is then scaled to match the remapped invariant. Without much overhead, this method preserves monotonicity and guarantees the conservation of the elastic energy.

Future work includes testing the reconstruction methods in a full elastic–plastic simulations and possibly developing methods that work for general tensors.

Acknowledgements This work was performed under the auspices of the National Nuclear Security Administration of the US Department of Energy at Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396 and supported by the DOE Advanced Simulation and Computing (ASC) program. The authors acknowledge the partial support of the DOE Office of Science ASCR Program. This work was partially supported by the Czech Technical University grant SGS16/247/OHK4/3T/14, the Czech Science Foundation project 14-21318S and by the Czech Ministry of Education project RVO 68407700.

References

1. T.J. Barth, Numerical methods for gasdynamic systems on unstructured meshes, in *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws, Proceedings of the International School on Theory and Numerics for Conservation Laws*, ed. by C. Rohde, D. Kroner, M. Ohlberger. Lecture Notes in Computational Science and Engineering (Springer, Berlin, 1997). ISBN 3-540-65081-4
2. E.J. Caramana, D.E. Burton, M.J. Shashkov, P.P. Whalen, The construction of compatible hydrodynamics algorithms utilizing conservation of total energy. J. Comput. Phys. **146**(1), 227–262 (1998)

3. C.W. Hirt, A.A. Amsden, J.L. Cook, An arbitrary Lagrangian-Eulerian computing method for all flow speeds. J. Comput. Phys. **14**(3), 227–253 (1974)
4. M. Kucharik, *Arbitrary Lagrangian-Eulerian (ALE) Methods in Plasma Physics*. PhD thesis, Czech Technical University in Prague (2006)
5. G. Luttwak, J. Falcovitz, Slope limiting for vectors: a novel vector limiting algorithm. Int. J. Numer. Methods Fluids **65**(11–12), 1365–1375 (2011)
6. L.G. Margolin, Introduction to "An arbitrary Lagrangian-Eulerian computing method for all flow speeds". J. Comput. Phys. **135**(2), 198–202 (1997)
7. D.J. Mavriplis, Revisiting the least-squares procedure for gradient reconstruction on unstructured meshes, in *AIAA 2003–3986 2003, 16th AIAA Computational Fluid Dynamics Conference, June 23–26, Orlando, Florida* (2003)
8. R. Menikoff, Equations of state and fluid dynamics. Technical Report LA-UR-07-3989, Los Alamos National Laboratory (2007)
9. L.G. Margolin, M. Shashkov, Second-order sign-preserving remapping on general grids. Technical Report LA-UR-02-525, Los Alamos National Laboratory (2002)
10. P.-H. Maire, R. Abgrall, J. Breil, R. Loubere, B. Rebourcet, A nominally second-order cell-centered Lagrangian scheme for simulating elasticplastic flows on two-dimensional unstructured grids. J. Comput. Phys. **235**, 626–665 (2013)
11. S.K. Sambasivan, M. Shashkov, D.E. Burton, Exploration of new limiter schemes for stress tensors in Lagrangian and ALE hydrocodes. Comput. Fluids **83**, 98–114 (2013)
12. J. Velechovsky, M. Kucharik, R. Liska, M. Shashkov, Symmetry-preserving momentum remap for ALE hydrodynamics. J. Phys.: Conf. Ser. **454**, 012003 (2013). IOP Publishing
13. M.L. Wilkins, Calculation of elastic-plastic flow. Technical Report UCRL-7322, California. University Livermore. Lawrence Radiation Laboratory (1963)

On Computing Compressible Euler Equations with Gravity

Christian Klingenberg and Andrea Thomann

Abstract We present a well-balanced finite volume solver for the compressible Euler equations with gravity. The Riemann solver used in the finite volume method is approximated by a so-called relaxation Riemann solution. Besides the well-balanced property, the scheme is also positivity preserving regarding the density and internal energy. The scheme is able to capture not only isothermal and polytropic stationary solutions of the hydrostatic equilibrium but also to preserve more general steady states up to machine precision. The scheme is tested on numerical examples including the preservation of arbitrary steady states and the evolution of small perturbations of stationary solutions to demonstrate the properties of the designed scheme.

Keywords Well-balanced scheme · Suliciu relaxation · Euler equations with gravity

1 Introduction

When solving the two- or three-dimensional Euler equations with gravity via a finite volume discretization, we are faced with several challenges. Firstly, we need a discretization which works well at both low and high Mach numbers for the homogeneous system. Secondly, we need a discretization which maintains hydrostatic equilibria to machine precision. Finally, when combining these two methods, we need to find a scheme that is numerically stable in more than one space dimension.

C. Klingenberg (✉) · A. Thomann
Universität Würzburg, Campus Hubland Nord, Emil-Fischer-Straße 40,
97074 Würzburg, Germany
e-mail: klingen@mathematik.uni-wuerzburg.de

A. Thomann
e-mail: andrea.thomann@mathematik.uni-wuerzburg.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_12

Solutions to the first challenge can be found in the literature; see, e.g., in [2]. Solutions to the second challenge can be found, e.g., in [3, 4].

We found experimentally, when combining these type of approaches, that typically instabilities arise when computing in more than one spacial dimension. In our numerical experiments, we found that one well-balanced method in particular was more stable than others. In this contribution, we shall report on this method. It is based on a relaxation approach leading to a positivity and entropy preserving scheme which is therefore especially useful in applications. In addition, it can be extended to higher order of accuracy and to higher dimensions.

Consider the system of compressible Euler equations with gravity in one space dimension given by the following set of equations

$$\begin{aligned} \partial_t \rho + \partial_x \rho u &= 0, \\ \partial_t \rho u + \partial_x (\rho u^2 + p) &= -\rho \partial_x \Phi, \\ \partial_t E + \partial_x (E + p) u &= -\rho u \partial_x \Phi. \end{aligned} \tag{1}$$

Here, $\rho > 0$ denotes the density, u the velocity, p the pressure and $E = \rho e + \frac{1}{2}\rho u^2$ the total energy, where $e > 0$ is the internal energy. The function Φ which is a continuous function from $\mathbb{R}$ to $\mathbb{R}$ denotes the gravitational potential. The pressure is described by a general pressure law which depends on the internal energy and specific volume $\tau = \frac{1}{\rho}$. We require for the solution $w = (\rho, \rho u, E)$ the density and the internal energy to be positive. That means the state vector w must belong to the set $\left\{ w \in \mathbb{R}^3 \mid \rho > 0,\ e > 0 \right\}$.

The paper is organized as follows. In Sect. 2, the relaxation method we use is described. The approximate Riemann solver which is designed to have the well-balanced property is presented in Sect. 3. Section 4 is devoted to the associated numerical scheme which is tested in Sect. 5 to verify the well-balancing property.

2 Relaxation

We consider the following relaxation model derived in [4] where a Suliciu-type relaxation approach is used, see [1],

$$\begin{aligned} \partial_t \rho + \partial_x \rho u &= 0, \\ \partial_t \rho u + \partial_x (\rho u^2 + \pi) &= -\rho \partial_x Z, \\ \partial_t E + \partial_x (E + \pi) u &= -\rho u \partial_x Z, \\ \partial_t \rho \pi + \partial_x (\rho \pi + a^2) u &= \frac{\rho}{\epsilon} (p(\tau, e) - \pi), \\ \partial_t \rho Z + \partial_x \rho Z u &= \frac{\rho}{\epsilon} (\Phi - Z). \end{aligned} \tag{2}$$

Here, the gravity Φ is approximated by a new variable Z, the pressure p by the new variable π, and $a > 0$ denotes the relaxation parameter.

Since we also need the density and internal energy to be positive, we require the state vector of the relaxation system $W = (\rho, \rho u, E, \rho\pi, \rho Z)$ to belong to the set $\left\{W \in \mathbb{R}^5 \mid \rho > 0,\ e > 0\right\}$.

For a given gravity function Φ, an equilibrium state for the relaxation model is defined by

$$W^{eq} = (\rho, \rho u, E, \rho p(\tau, e), \rho\Phi)^T. \tag{3}$$

The eigenvalues of the system are $\lambda^{\pm} = u \pm \frac{a}{\rho}$ and $\lambda^u = u$ where the eigenvalue λ^u has multiplicity three. Following [7] one finds the fields associated to the eigenvalues are linearly degenerate and the Riemann invariants with respect to $\lambda^{\pm}$ are

$$I_1^{\pm} = u \pm \frac{a}{\rho},\ I_2^{\pm} = \pi \mp au,\ I_3^{\pm} = e - \frac{\pi^2}{2a^2},\ I_4^{\pm} = Z \tag{4}$$

and with respect to λ^u are

$$I_1^u = u. \tag{5}$$

In the following, let us consider a Riemann problem as initial data with two constant values separated by a discontinuity at $x = 0$

$$W_0(x) = \begin{cases} W_L & x < 0 \\ W_R & x > 0. \end{cases} \tag{6}$$

The solution, consists of four constant states separated by contact discontinuities and has the following structure

$$W_R\left(\frac{x}{t}; W_L, W_R\right) = \begin{cases} W_L & \frac{x}{t} < \lambda^- \\ W_L^* & \lambda^- < \frac{x}{t} < \lambda^u \\ W_R^* & \lambda^u < \frac{x}{t} < \lambda^+ \\ W_R & \lambda^+ < \frac{x}{t} \end{cases}, \tag{7}$$

where W_L^*, W_R^* denote the intermediate states. This leads to 10 unknowns in the Riemann problem, five unknowns each for the intermediate states $W_{L,R}^*$ but one obtains only nine relations from the Riemann invariants (4) and (5), for the computations see [9]. This leaves us with one degree of freedom to choose the 10th relation such that the resulting scheme has the well-balanced property.

How to obtain this 10th relation will be described in the following section.

3 Well-Balanced Property

In the following, we will focus on steady states at rest, which are solutions of

$$\begin{aligned} u &= 0, \\ \partial_x p &= -\rho \partial_x \Phi. \end{aligned} \tag{8}$$

Following [5], we write the hydrostatic solution as $\bar{\rho} = \rho_c \alpha(x)$, $\bar{p} = p_c \beta(x)$, where the constants p_c, ρ_c are reference values at some location $x = x_c$ and $\alpha(x)$, $\beta(x)$ are non-dimensional functions. Since the density and the pressure are strictly positive, we require $\alpha, \beta > 0$. These functions must satisfy the hydrostatic condition (8) which leads to an expression for the derivative of the potential given by

$$\partial_x \Phi(x) = -\frac{p_c}{\rho_c} \frac{\partial_x \beta(x)}{\alpha(x)}. \tag{9}$$

A well-balanced scheme must satisfy the discretized form of the hydrostatic equation. Since the discretized flux derivative must exactly balance the discretized source term, we choose the following symmetrical discretization

$$\pi_R - \pi_L = \frac{\pi_c}{2\rho_c} (\beta_R - \beta_L) \left(\frac{\rho_L}{\alpha_L} + \frac{\rho_R}{\alpha_R} \right). \tag{10}$$

Using this relation in addition to the relations gained from the Riemann invariants, the intermediate states $W^*_{L,R}$ can be determined; for details see [9]. Thus, the Riemann problem of the relaxation system completed by relation (10) has a unique solution which is given by (7).

Using this Riemann solution one obtains an approximate Riemann solver for the original system (1) by projecting the solution of the relaxation system on its first three components

$$w^{eq}\left(\frac{x}{t}; w_L, w_R\right) = W_R^{(\rho, \rho u, E)}\left(\frac{x}{t}; W_L, W_R\right). \tag{11}$$

The following result shows the well-balanced property of the approximative Riemann solver.

Theorem 1. *The approximate Riemann solver stated by* (11) *is well-balanced in the sense that the initial condition on each cell i given by*

$$u_i = 0, \ \frac{\rho_i}{\alpha_i} = const., \ \frac{p_i}{\beta_i} = const., \tag{12}$$

is preserved.

Proof. For the proof, we refer the reader to [9]. □

We want to conclude this section by mentioning some additional properties of the above defined Riemann solver; for detailed proof see [4].

- The approximative Riemann solver ensures the positivity of the density ρ and the pressure p for a sufficiently large relaxation parameter a. That means starting with data belonging to $\Omega := \{w = (\rho, \rho u, E) \in \mathbb{R}^3,\ \rho > 0,\ e > 0\}$, and then the solution $w^{eq}(\frac{x}{t}; w_L, w_R)$ also belongs to Ω.
- If one considers an entropy inequality $\partial_t \rho F(\eta) + \partial_x F(\eta) u \leq 0$ for the Euler equations with gravity where $\eta(\tau, e)$ denotes a specific entropy, then the approximative Riemann solver is consistent with the entropy inequality.

4 Numerical Scheme

In this section, we describe the numerical scheme associated with the approximative Riemann solver developed above.

The computational domain is divided in N cells $C_i = (x_{i-1/2}, x_{i+1/2})$ with fixed step-size Δx. The time discretization on the interval $[0, T]$ is given by $t^{n+1} = t^n + \Delta t$ where $\Delta t > 0$ denotes the time step restricted by a CFL condition. Define the approximative solution at time t^n as $w^n(x, t^n) = w_i^n$ for $x \in \left(x_{i-1/2}, x_{i+1/2}\right)$ and the updated state at time t^{n+1} as

$$w_i^{n+1} = \frac{1}{\Delta x} \int_{C_i} w^n(x, t^n + \Delta t). \tag{13}$$

Thereby $w^n(x, t^n + t)$ is a sequence of the approximative Riemann solver (11) at each interface $x_{i+/2}$ given by

$$w^n(x, t^n + t) = w^{eq}(\frac{x - x_{i+1/2}}{t}, w_i^n, w_{i+1}^n) \tag{14}$$

for $x \in (x_i, x_{i+1})$ and $t \in (0, \Delta t)$.

Following the computations in [6, 7, 10], we obtain for the updated state

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x}(F_{i+1/2} - F_{i-1/2}) + \frac{\Delta t}{2}(S_{i-1/2} + S_{i+1/2}). \tag{15}$$

The approximated source term is given by

$$S_{i+1/2} = \left(0, \frac{p_c}{\rho_c} \frac{(\beta_{i+1} - \beta_i)}{\Delta x} \frac{1}{2} \left(\frac{\rho_i}{\alpha_i} + \frac{\rho_{i+1}}{\alpha_{i+1}}\right), u_{i+1/2}^* \frac{p_c}{\rho_c} \frac{(\beta_{i+1} - \beta_i)}{\Delta x} \frac{1}{2} \left(\frac{\rho_i}{\alpha_i} + \frac{\rho_{i+1}}{\alpha_{i+1}}\right)\right) \tag{16}$$

Defining $s_{LR} = \frac{p_c}{\rho_c}\frac{(\beta_R-\beta_L)}{\Delta x}\frac{1}{2}\left(\frac{\rho_L}{\alpha_L}+\frac{\rho_R}{\alpha_R}\right)$ and using the formulas for the intermediate states, the numerical flux function reads

$$f_{i+1/2} = \begin{cases} (\rho_L u_L, \rho_L u_L^2 + \pi_L + s_{LR}, (E_L + \pi_L)u_L + u^* s_{LR})^T & u_L - \frac{a}{\rho_L} > 0, \\ (\rho_L^* u^*, \rho_L^* u^{*2} + \pi_L^* + s_{LR}, (E_L^* + \pi_L^*)u^* + u^* s_{LR})^T & u_L - \frac{a}{\rho_L} < 0 < u^*, \\ (\rho_R^* u^*, \rho_R^* u^{*2} + \pi_R^* - s_{LR}, (E_R^* + \pi_R^*)u^* - u^* s_{LR})^T & u* < 0 < u_R - \frac{a}{\rho_R}, \\ (\rho_R u_R, \rho_R u_R^2 + \pi_R - s_{LR}, (E_R + \pi_R)u_R - u^* s_{LR})^T & u_R - \frac{a}{\rho_R} < 0. \end{cases} \tag{17}$$

5 Numerical Results

In the following section, two types of test cases are presented. First a well-balanced test is performed, to verify that the initial condition, if satisfying the condition (8), is preserved on machine precision. The second test addresses the evolution of small perturbations of a hydrostatic atmosphere.

Well-Balanced Tests

For the well-balanced tests, we consider stationary solutions for three different potential functions $\Phi(x) = x$, $\Phi(x) = \frac{1}{2}x^2$ and $\Phi(x) = \sin(2\pi x)$ to demonstrate that the scheme can deal with more complex gravitational fields.

For all examples, the computational domain is $[0, 1]$ and the initial velocity is zero. All errors are given in the L^1-norm and computations are performed in double precision.

As a first example, we consider a isothermal hydrostatic atmosphere given by

$$\rho_0(x) = \exp(-\Phi(x)), \;\; p_0(x) = \exp(-\Phi(x)). \tag{18}$$

In Table 1 the error in density, velocity, and pressure with respect to the initial condition are given. The calculations are performed on a grid with 100 and 1000 cells, respectively, up to a final time $T_f = 2.0$. As can be seen from Table 1, the error is of the order of machine precision and thus the hydrostatic atmosphere is preserved.

To show that the scheme can also preserve more general steady states, we consider as a second test the stationary solution from [3]. For the quadratic potential $\Phi(x) = \frac{1}{2}x^2$, a stationary solution is given by

$$\bar{\rho}(x) = \exp(-x), \quad \bar{p}(x) = (1 + x)\exp(-x) \tag{19}$$

which corresponds to a non-uniform temperature profile given by $T(x) = 1 + x$. Thus, the steady state is not isothermal. The number of cells used for the calculations is doubled for each calculation starting with 100 cells. The error in density, velocity,

Table 1 Error in density, velocity, and pressure for isothermal example using different potentials

$\Phi(x)$	Cells	Density	Velocity	Pressure
x	100	1.88738E-017	3.67483E-017	2.05391E-017
	1000	3.47499E-017	8.74191E-017	4.32431E-017
$\frac{1}{2}x^2$	100	3.94129E-016	3.19565E-016	6.11732E-016
	1000	1.10456E-015	4.84117E-016	1.84618E-015
$\sin(2\pi x)$	100	8.60422E-017	7.39687E-017	1.73749E-016
	1000	1.07663E-015	5.38106E-015	1.11399E-015

Table 2 Error in density, velocity, and pressure for a non-hydrostatic steady state

Cells	Density	Velocity	Pressure
100	7.04991E-017	4.84102E-016	7.54951E-017
200	8.21565E-017	3.17104E-016	8.38218E-017
400	2.28983E-016	6.08430E-016	5.95357E-016
800	3.49997E-016	1.40357E-015	5.23331E-016
1600	6.12600E-016	1.22546E-015	5.05290E-016

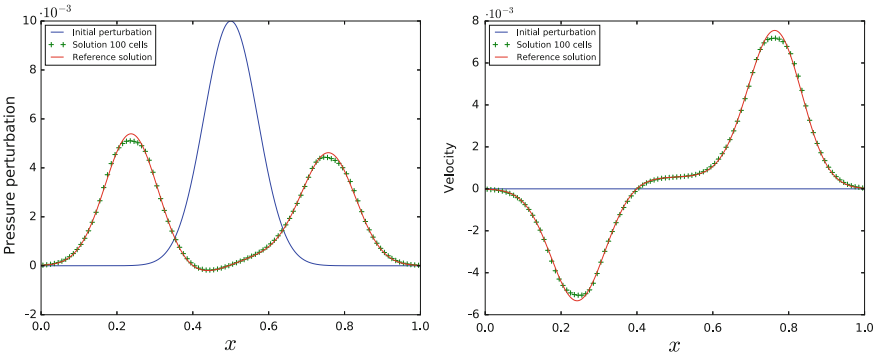


Fig. 1 Perturbation in pressure (left) and velocity (right)

and pressure with respect to the initial condition is reported in Table 2. One can see that the initial steady state is preserved with machine precision.

Evolution of Small Perturbations

As a last example, taken from [8], the evolution of a small perturbation added to an initial isothermal hydrostatic equilibrium is investigated. The initial condition on the domain [0, 1] is given by

$$\begin{aligned}\Phi(x) &= x,\\ \rho(x) &= \exp(-\Phi(x)),\\ p(x) &= \exp(-\Phi(x)) + 0.01\exp(-100(x-0.5)^2),\end{aligned}$$

where the pressure is perturbed by a Gauß function centered in $x = 0.5$. The solution is computed at time $T = 0.2$ with 100 cells and a reference solution using 30000 cells. In Fig. 1, the pressure perturbation $p(x) - p_0(x)$ and the resulting velocity perturbation are plotted in comparison with the initial perturbation.

Acknowledgements The authors want to thank Praveen Chandrashekar for pointing out to us the potential usefulness of reference [5].

References

1. F. Bouchut, Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources, in *Frontiers in Mathematics* (Birkhäuser Verlag, Basel, 2004). https://doi.org/10.1007/b93802
2. W. Barsukow, P. Edelmann, C. Klingenberg, F. Miczek, F. Röpke, A numerical scheme for the compressible low-mach number regime of ideal fluid dynamics. Accepted in J. Sci. Comput
3. P. Chandrashekar, C. Klingenberg, A second order well-balanced finite volume scheme for Euler equations with gravity. SIAM J. Sci. Comput. **37**(3), B382–B402 (2015). https://doi.org/10.1137/140984373
4. V. Desveaux, M. Zenk, C. Berthon, C. Klingenberg, A well-balanced scheme to capture non-explicit steady states in the Euler equations with gravity. Int. J. Numer. Methods Fluids **81**(2), 104–127 (2016). https://doi.org/10.1002/fld.4177
5. D. Ghosh, E.M. Constantinescu, *Well-Balanced Formulation of Gravitational Source Terms for Conservative Finite-Difference Atmospheric Flow Solvers. AIAA Aviation* (American Institute of Aeronautics and Astronautics, 2015). https://doi.org/10.2514/6.2015-2889
6. A. Harten, P.D. Lax, B. van Leer, On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. Publ. Soc. Ind. Appl. Math. **25**(1), 35–61 (1983). https://doi.org/10.1137/1025002
7. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics (Cambridge University Press, Cambridge, 2002). https://doi.org/10.1017/CBO9780511791253
8. R.J. LeVeque, D.S. Bale, Wave propagation methods for conservation laws with source terms, in *Hyperbolic Problems: Theory, Numerics, Applications*, ed. by R. Jeltsch, M. Fey. International Series of Numerical Mathematics, vol. 130 (Birkhäuser, Basel, 1999), pp. 609–618
9. A. Thomann, C. Klingenberg, A second order well-balanced finite volume scheme for Euler equations with gravity for arbitrary hydrostatic equilibrium
10. E.F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 3rd edn. (Springer, Berlin, 2009). https://doi.org/10.1007/b79761 (A practical introduction)

On Well-Posedness for a Multi-particle Fluid Model

Christian Klingenberg, Jens Klotzky and Nicolas Seguin

Abstract In this paper, we study a one-dimensional fluid modelled by the Burgers equation influenced by an arbitrary but finite number of particles $N(t)$ moving inside the fluid, each one acting as a point-wise drag force with a particle-related friction constant λ. For given particle paths $h_i(t)$, we only assume finite speed of particles, allowing for crossing, merging and splitting of particles. This model is an extension of existing models for fluid interactions with a single particle; compare (Andreianov et al., SIAM J Math Anal 46(2):1030–1052, 2014, [3], Lagoutière et al., J Differ Equ 245(11):3503–3544, 2008, [10]):

$$\partial_t u(x,t) + \partial_x \left(\frac{u^2}{2} \right) = \sum_{i=1}^{N} \lambda (h_i'(t) - u(t, h_i(t)) \delta(x - h_i(t))$$

Well-posedness for the Cauchy problem, as well as an L^∞ bound, is proven under the weak assumption that particle paths are Lipschitz continuous. In this context, an entropy admissibility criteria are shown, using the theory of L^1-dissipative germs, compare (Andreianov et al., Arch Ration Mech Anal 201:26–86, 2011, [2]), to deal with the moving interfaces resulting from the point-wise particles and the shock waves from the fluid equation interacting with them.

Keywords Fluid-particle interaction · Burgers equation · Well-posedness · Germ

C. Klingenberg · J. Klotzky (✉)
Universität Würzburg, Campus Hubland Nord, Emil-Fischer-Straße 40,
97074 Würzburg, Germany
e-mail: jens.klotzky@mathematik.uni-wuerzburg.de

C. Klingenberg
e-mail: klingen@mathematik.uni.wuerzburg.de

N. Seguin
Université de Rennes 1, 263 avenue du Général Leclerc, 35042 Rennes, France
e-mail: nicolas.seguin@univ-rennes1.fr

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_13

1 Introduction

We consider an inviscid fluid with velocity $u(t,x)$ and a finite number of particles moving inside. The fluid is modelled by the inviscid Burgers equation, and the particles act as a point-wise drag force on the fluid, namely $\lambda(h_i'(t) - u(t,h_i(t))$, where λ is the friction constant related to the particle and $h_i(t)$ the given path of the ith particle. The Cauchy problem writes

$$\begin{aligned} \partial_t u + \partial_x(u^2/2) &= \sum_{i=1}^{N} \lambda(h_i'(t) - u(t,h_i(t))\delta(x - h_i(t)), \\ u(0,x) &= u_0(x) \end{aligned} \tag{1}$$

with

$u(x,t)$	velocity of the one-dimensional fluid
$h_i(t)$	the given position of the ith particle at time t
λ	the friction constant corresponding to a particle
$N(t)$	arbitrary but finite number of particles at time t
$u_0 \in L^\infty(\mathbb{R})$	the given L^∞initial data for the fluid

Note that this model also bears the difficulty of interpretating the non-conservative product $u(t,h_i(t))\delta(x - h_i(t))$. This problem was tackled in [3] by a regularization of the particle, using sequences of non-negative, compactly supported density functions (see also [7] for a similar approach). However, an analysis of the behaviour of the fluid at the position of the particle allows for a well-posedness proof considering the influence of the particle as a condition on the behaviour of the fluid at a moving interface located at the particle position. The theory extends the analysis of the fluid–solid interaction of [3, 10], where the original model also includes coupling to an ordinary differential equation, to the case of multiple particles. Models of this kind are of increasing interest theoretically, cf. [4], as well as in applications like trajectory tracking in traffic flow, cf. [5, 6].

We proceed in the following way. In Sect. 2, we give an admissibility condition for the selection of physical shock waves and therefore a definition of entropy solutions to the problem. At the end of Sect. 2, we will state the main theorem, which is the well-posedness result for Problem (1) and an L^∞ bound. Sections 3 and 4 give the proof to this theorem, where Sect. 3 contains the existence proof as well as the L^∞ bound and Sect. 4 is devoted to the uniqueness proof using almost classical Kruzkov-type arguments combined with the notion of germs, i.e. sets of admissible states connected by shock waves, first introduced in [2].

2 Definition of Entropy Solutions

The behaviour of a solution across one particle is dictated by the drag of the particle. However, there might also be shock waves originating from the fluid equation. A travelling wave study with respect to the particle speed was done in [10] regarding a single particle and acts as a building block for the analysis of the behaviour in the case of multiple particles. It is proven in [10] that the following definition of sets describes the admissible jumps across the interface of a single particle.

Definition 1. Let $\mathscr{G}_\lambda$ be the set of possible states left and right of a particle with friction λ. A case-by-case study with respect to u_l, u_R, h gives the characterization

$$(u_L, u_R) \in \mathscr{G}_\lambda$$
$$\Leftrightarrow u_R \in \begin{cases} \{u_L - \lambda\} & \text{if } u_L < h', \\ [2h' - u_L - \lambda, h'] & \text{if } h' \le u_L \le h' + \lambda, \\ \{u_L - \lambda\} \cup [2h' - u_L - \lambda, 2h' - u_L + \lambda] & \text{if } u_L > h' + \lambda. \end{cases}$$

In the case of more that one particle, the behaviour of the fluid at each particle is governed by an interface admissibility condition $\mathscr{G}_i = \mathscr{G}_{\lambda_i}$, meaning that the trace of the solution at the left and right of each particle lie in $\mathscr{G}_\lambda$. Thus, we are able to define entropy admissible solutions to the problem as long as the particle paths do not intersect using the notion of admissible particle-related jumps and the notion of Kruzkov entropy η, entropy flux Φ, defined by

$$\eta(a, c) = |a - c|$$
$$\Phi(a, c) = \operatorname{sgn}(a - c)(f(a) - f(c)),$$

which enable comparison to any constant $c \in \mathbb{R}$.

Definition 2. Given $u_0 \in L^\infty, N > 0, h_i(t) \in W^{1,\infty}([0, T]), h_i(t) \neq h_j(t) \forall t \in [0, T], i \neq j$. We call $u \in L^\infty(\mathbb{R}^+ \times \mathbb{R})$ weak entropy solution to the Cauchy problem; if for $N \in \mathbb{N}$ the finite number of particles, $h_i(t)$ the position and $h_i'(t)$ the velocity of particle i, with $i = 1, \ldots, N$, u satisfies for all $c \in \mathbb{R}$ and almost every time t

$$\int_0^T \int_{\mathbb{R}} |u - c| \partial_t \phi + \Phi(u, c) \partial_x \phi) \, dx dt + \int_{\mathbb{R}} |u_0 - c| \phi(0, x) dx \ \geq \ 0 \qquad (2)$$

with $\phi \in C^\infty([0, T] \times \mathbb{R}, \mathbb{R}^+), \phi(t, h_i(t)) = 0$, and additionally

$$(\gamma_u^-(t, h_i(t)), \gamma_u^+(t, h_i(t))) \in \mathscr{G}_\lambda(t), \quad \text{for a.e. } t \in (0, T)$$

where we denoted the left and right traces of $u(t, x)$ at the position of the particles by $\gamma_u^-(t, h_i(t)), \gamma_u^+(t, h_i(t))$, respectively. Due to the nature of the Burgers equation, these traces exist a priori, even for L^∞ initial data, cf [11].

Note that whenever two particles are located at the same position, a careful new definition of the particle-related germs has to be taken into account. This is not a problem for crossing, as the condition is enforced only almost everywhere in time; however, if two or more particles merge, the corresponding germ changes, resulting in the following definition of time-dependent interface condition

$$\mathcal{G}_\lambda(t) = \mathcal{G}_{n_i(t)\times\lambda}(h_i'(t)), \quad \text{with } n_i(t) := \#\{j \in [0, N], h_i(t) = h_j(t)\}.$$

This definition makes sure that the interface condition really applies the drag of both particles and does not impose two (maybe contradictory) conditions at the same position. The fact that the influence of the particles adds up like that uses the specific form of the germ $\mathcal{G}_\lambda$, was proven by the authors and can be found in the upcoming publication [8].

Remark 1. The definition of entropy solution is done using the notion of germs, introduced in [2]. Furthermore, the entropy condition cannot be distinguished from an entropy condition for a discontinuous flux problem with interfaces located at the particle positions $h_i(t)$, emphasizing the point-wise influence of the particles.

At this point, we state our main theorem.

Theorem 1. *Given any finite time T, initial data $u_0(x) \in L^\infty(\mathbb{R})$ and Lipschitz continuous in time particles paths $h_i(t), i \in [1, N]$, then there exists a unique solution $u(t, x) \in L^\infty([0, T] \times \mathbb{R})$, entropy admissible in the sense of Definition 2. Additionally, $u(t, x)$ satisfies for all $t \in [0, T]$*

$$\|u(t, \cdot)\|_{L^\infty} \leq \|u_0(\cdot)\|_{L^\infty} + N\lambda. \tag{3}$$

The proof of this theorem is distributed between the next two sections.

3 Existence

We will prove existence of entropy admissible solutions in the following way. Given initial data $u_0 \in L^\infty(\mathbb{R})$ and any finite time interval $[0, T]$, we divide the problem into several local problems and use the following existence result for the problem with a single particle, which is proven in [3].

Lemma 1. *Given $h \in W^{1,\infty}([0, T])$ and $u_0 \in L^\infty(\mathbb{R})$, then there exists a unique entropy admissible solution u of (1) with $N(t) = 1$.*

Several difficulties arise. Even though the behaviour of the fluid in the presence of a single particle is known, each particle generates waves interfering with the other particles, creating domains of unknown behaviour. Additionally, the possibility of crossing, merging and splitting of particles seems to complicate some of the nice

properties that were holding as long as only one particle was present, e.g. the global in time bound on the total variation.

The proof is done using an explicit construction algorithm based on the existence result in the presence of a single particle, which we will present here for the case of two particles. Note, however, that this can be easily extended to any finite number of particles by simply choosing a good time stepping, creating domains where the following analysis applies locally.

At the same time, we will prove the L^∞ bound (3), justifying the existence of a maximum speed of propagation, denoted L from here on, which, though a very natural property of hyperbolic equations, needs to be checked in the presence of source terms. Both the L^∞ bound and the existence are constructed using a time stepping, which ensures that the cones of influence of two particles do not intersect.

Lemma 2. *Given any time $t_i \in [0, T]$, there exists a time $t_{i+1} > t_i$, such that given problem (1) with two particles with particle paths $h_1, h_2 \in Lip([t_i, t_{i+1}])$ with $h_1(t) \neq h_2(t) \in [t_i, t_{i+1}]$ and initial data $u(t_i) \in L^\infty(\mathbb{R})$, then there exists a solution $u(t, x) \in L^\infty([t_i, t_{i+1}] \times \mathbb{R})$, entropy admissible in the sense of (2).*
Additionally, if $u(t_i, x)$ satisfies for $x \in \mathbb{R}$

$$c_{min}(t_i, x) \leq u(t_i, x) \leq c_{max}(t_i, x),$$

then $u(t, x)$ satisfies for almost every $t \in [t_i, t_{i+1}]$, $x \in \mathbb{R}$

$$c_{min}(t, x) \leq u(t, x) \leq c_{max}(t, x), \tag{4}$$

with piece-wise constant functions

$$c_{min,\, max}(t, x) = \begin{cases} c^1_{min,\, max} & \text{for } x \in \Omega_1(t), t \in [t_i, t_{i+1}] \\ c^2_{min,\, max} & \text{for } x \in \Omega_2(t), t \in [t_i, t_{i+1}] \\ c^3_{min,\, max} & \text{for } x \in \Omega_3(t), t \in [t_i, t_{i+1}] \end{cases}$$

with

$$\begin{aligned} \Omega_1(t) &:= (-\infty, h_1(t)) \\ \Omega_2(t) &:= (h_1(t), h_2(t)) \\ \Omega_3(t) &:= (h_2(t), \infty) \end{aligned}$$

such that for $j = 1, 2$

$$c^j_{min,\, max} = c^{j+1}_{min,\, max} + \lambda$$

and $c^{k_1}_{min} = \inf\limits_{\Omega_{k_1}(t_i)} u(t_i, x)$, $c^{k_2}_{max}(t, x) = \sup\limits_{\Omega_{k_2}(t_i)} u(t_i, x)$ with

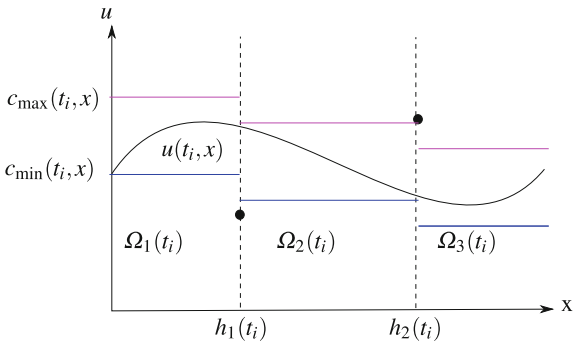


Fig. 1 Boundaries $c_{\min}, c_{\max}$ on the solution in the regions between the particles at time t_i. As time passes, the particles will change their position and the respective bounds will shift along the x-axis

$$k_1 = \arg \min_{j=1,2,3} \left\{ ess \inf_{x\in\Omega_1} u(t_i,x),\ ess \inf_{x\in\Omega_2} \big(u(t_i,x) - \lambda\big),\ ess \inf_{x\in\Omega_3} \big(u(t_i,x) - 2\lambda\big) \right\}$$
$$k_2 = \arg \max_{j=1,2,3} \left\{ ess \sup_{x\in\Omega_1} u(t_i,x),\ ess \sup_{x\in\Omega_2} \big(u(t_i,x) - \lambda\big),\ ess \sup_{x\in\Omega_3} \big(u(t_i,x) - 2\lambda\big) \right\}.$$

The last statement (4) is actually a stronger result than the L^∞ bound, as (3) follows directly from (4) as soon as it is established for all times $t \in [0, T]$. To see this, it is very important to note that the time dependence of c_{min}, c_{max} is only due to the position of the particles and does not change the values of the two functions, cf. Fig. 1.

Proof. To be able to make use of the existing results for the case of a single particle, i.e. Lemma 1, we choose t_{i+1} such that the waves propagating from the two particles cannot intersect in $[t_i, t_{i+1}] \times \mathbb{R}$. This is achieved by defining

$$t_{i+1} = t_i + \frac{h_2(t_i) - h_1(t_i) - 2\varepsilon}{2L}.$$

where $L = L(\|u\|_{L^\infty}, h_1', h_2') = \max_{x\in\Omega} (c_{\max}(0,x), -c_{\min}(0,x))$ denotes the finite speed of propagation and $\varepsilon > 0$ can be chosen arbitrarily small. We define the superposition of $[t_i, t_{i+1}] \times \mathbb{R} = B_1 \cup P_1 \cup B_2 \cup P_2 \cup B_3$ such that P_1, P_2 contain the particles and all waves emanating from them.

$$\begin{aligned}
P_{1,2}(t) &:= [h_{1,2}(t_{i+1}) - L(t_{i+1} - t), h_{1,2}(t_{i+1}) + L(t_{i+1} - t)] \\
B_1(t) &:= (-\infty, h_1(t_{i+1}) - L(t_{i+1} - t)] \\
B_2(t) &:= [h_1(t_{i+1}) + L(t_{i+1} - t), h_2(t_{i+1}) - L(t_{i+1} - t)] \\
B_3(t) &:= [h_2(t_{i+1}) + L(t_{i+1} - t), \infty).
\end{aligned}$$

From the analysis done for a single particle, we know that given $u(t_i, \cdot) \in L^\infty(P_1)$ and given that the solution $u(t, x)$ with $x \in \mathbb{R} \backslash P_1$ in the adjacent regions to P_1 satisfies $c_{\min}(t, x) \leq u(t, x) \leq c_{\max}(t, x)$, the bounds are also true in P_1,[1] namely

$$c_{\min}(t, x) \leq u(t, x) \leq c_{\max}(t, x) \quad \text{for } x \in P_1$$

and the same holds equivalently for P_2. Also, we know for the regions B_j, $j = 1, 2, 3$, given $u(t_i, \cdot) \in L^\infty(B_j)$, $u(t, x)$ on the boundaries of B_j and given that the solution $u(t, x)$ with $x \in \mathbb{R} \setminus B_j$ in the adjacent region to B_j satisfies $c_{\min}(t_{in}, x) \leq u(t, x) \leq c_{\max}(t_{in}, x)$, the bounds are also true in B_j

$$c_{\min}(t, x) \leq u(t, x) \leq c_{\max}(t, x). \quad \text{for } x \in B_j,$$

as the Burgers equation with L^∞ boundary data satisfies an L^∞ bound for any finite time. Piecing together the different regions, given $c_{\min}(t_i, x) \leq u(t_i, x) \leq c_{\max}(t_i, x)$, we obtain (4).

Therefore, we define the partition $[t_i, t_{i+1}] \times \mathbb{R} = \Sigma_1 \cup \Sigma_2$ with

$$\Sigma_1(t) = (-\infty, h_2(t_i) - L(t - t_i)]$$
$$\Sigma_2(t) = (h_1(t_i) + L(t - t_i), \infty)$$

Each of those regions contains only one particle, and therefore, applying Lemma 2 twice, we obtain existence of an entropy solution in $[t_i, t_{i+1}] \times \mathbb{R}$. □

Iterating this by using $t_i = t_{i+1}$ as new starting time for Lemma 3 until reaching time T gives the existence result of Theorem 1, and the L^∞ bound follows from property (4) as long as the particle paths do not intersect.

It remains to investigate the case of particles being located at the same position at some time $t \leq T$. We choose to stop the current timestep, whenever two particles are located at the same position; thus, from the three considered cases of particle interactions, i.e. crossing, merging and splitting of particles, only the following two cases need to be dealt with. Again, we restrict us to two particles for simplicity, as the case of more particles follows using the same mechanism; see Fig. 2.

1. Two particles are located at the same position at the end of a given time interval $[t_0, T]$ (merging).
2. Two particles are located at the same position at the initial time of a given time interval $[t_0, T]$ (splitting).

Case 1: $h_1(T) = h_2(T)$. The difficulty of this case lies in the time stepping, as at first glimpse, it is unclear whether or not the proposed method of construction used in the proof of Lemma 3 can actually reach time T. The reason is that

[1] This was a by-product of constructing the L^∞ bound in [3] and can be found in the proof of the corresponding lemma.

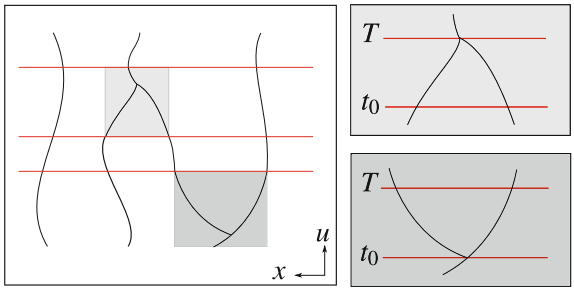


Fig. 2 On the left a random movement of particles, including sections where particles merge, corresponding to Case 1 (upper right), and split, corresponding to Case 2 (lower right). Whenever there is one of the two special cases, the method of construction has to be adapted

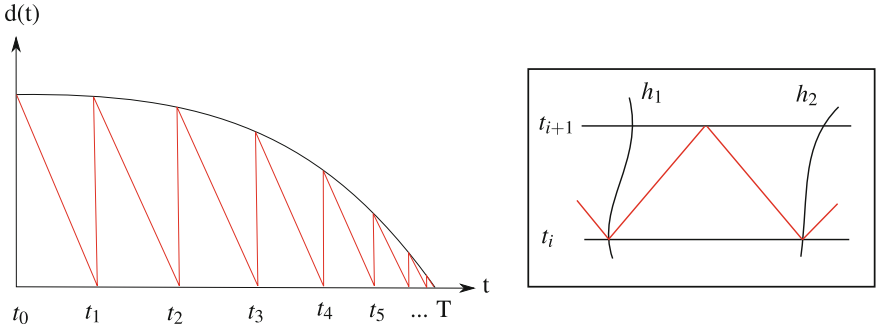


Fig. 3 On the left a visualization of the convergence for the construction method and $t_i \to T$. The method behaves like a simplified Newton method, where the slope corresponding to the maximal speed of propagation remains fixed. On the right the construction of the length of a single timestep, given by the longest possible time, such that waves propagating from the two particles do not intersect

$$t_{i+1} - t_i = \frac{h_2(t_i) - h_1(t_i) - 2\varepsilon}{2L},$$

meaning the length of each timestep depends on the distance between the particles $h_2(t_i) - h_1(t_i)$ which goes to zero as t goes to T.

However, the method of construction is equivalent to finding the root of the distance between the particles, denoted $d(t)$, by means of a simplified Newton method, which can be seen by measuring the distance against time and including the method in the picture, cf. Fig. 3. Therefore, Lemma 3 holds for all timesteps where $t_i \in [t_0, T)$, reaching time T either in a finite number of timesteps or as the limit of $n \to \infty$ if the particles have zero contact angle.

Case 2: $h_1(t_0) = h_2(t_0)$. This case is more delicate, as the method of construction fails to construct any solution between the two particles because all information about this region emanates from the two particles. There is no first timestep, as the choice

of each timestep depends on finding a superposition suitable in the sense that each domain contains only waves coming from one particle (or none).

We solve this problem by shifting the particles apart, defining the particle paths of the approximated problem by

$$h_1^\varepsilon(t) = h_1(t),$$
$$h_2^\varepsilon(t) = h_2(t) + \varepsilon.$$

Therefore, the particle paths do not intersect anymore and we meet the conditions of Lemma 3. Using the method of construction, we obtain existence and the L^∞ bound for u^ε and any given finite time T. It remains to show convergence of the approximate solution u^ε to the solution of the original problem, which is done using Helly's theorem. In order to be able to apply the latter, a bound in the total variation has to be established, which can be proven using a bound on the total variation for the problem with only a single particle, which was proven in [3] using a wavefront tracking method and regularized initial data

$$u_0^\delta(x) = u_0(x) * \rho^\delta$$

with ρ^δ being a regularizing kernel such that $u_0^\delta \in L^\infty \cap BV_{loc}(\mathbb{R})$.

For a more in-depth analysis of the latter cases, which would exceed the purpose of this article, we refer the reader to the upcoming publication [8].

4 Uniqueness of Entropy Solutions

This section is devoted to proving the uniqueness of solutions to problem (1), whenever the admissibility condition (2) is satisfied. Following the ideas of Kruzkov, this is done by using the method of doubling of variables and the framework of germs, introduced by Andreianov, Karlsen and Risebro [2]. The key property of an admissibility germ to allow to conclude uniqueness, as proven in their paper, is dissipativity. We state this property of $\mathcal{G}_\lambda$ in the following lemma.

Lemma 3. *The admissibility germ $\mathcal{G}_\lambda$ corresponding to the particle with velocity h' and friction λ is dissipative in the sense that*

$$(c_l, c_r) \in \mathcal{G}_\lambda \;\Leftrightarrow\; \left[\forall (b_l, b_r) \in \mathcal{G}_\lambda : \overline{\Phi}(h'; c_l, b_l) \geq \overline{\Phi}(h'; c_r, b_r)\right] \tag{5}$$

where

$$\overline{\Phi}(h'; c, b) = \Phi(c, b) - h'|c - b|.$$

Let $\Omega = [0, T] \times \mathbb{R}$, $\phi \in C_c^\infty(\Omega)$ be a classical, compactly supported test function and

$$w_\varepsilon(x) = \begin{cases} 0, & \text{when } |x| \le \frac{\varepsilon}{2}, \\ 1, & \text{when } |x| \ge \varepsilon. \end{cases}$$

a continuous function with $w'_\varepsilon(x) = \operatorname{sgn}(x)\,\frac{2}{\varepsilon}$ for $\frac{\varepsilon}{2} \le |x| \le \varepsilon$.

Given two entropy solutions u, v, with the same initial data $u_0 = v_0$, we apply the method of doubling of variables, cf. [9], and choosing as a test function $\psi(x,t) = \phi(x,t) \times w_\varepsilon(t, x - h_1(t)) \times \ldots \times w_\varepsilon(t, x - h_N(t))$, we obtain

$$\int_\Omega |u-v| \partial_t \big(\phi \times \prod_{1\le i\le N} w_\varepsilon(x-h_i(t))\big) + \int_{\mathbb{R}} |u_0 - v_0| \big(\phi(0,x) \times \prod_{1\le i\le N} w_\varepsilon(x-h_i(0))\big) dx$$
$$+ \int_\Omega \Phi(u,v) \partial_x \big(\phi \times \prod_{1\le i\le N} w_\varepsilon(x-h_i(t))\big) \ge 0.$$

Remark 2. Note that ψ is not C^∞ and one should regularize w_ε using classical mollifiers to this aim, but this would introduce unnecessary heavy notations that we skip for the sake of brevity.

Due to the choice of test function, we cannot see the interfaces and the method of Kruzkov works classically. Using chain rule and recognizing that

$$\partial_t w_\varepsilon(x - h_i(t)) = w'_\varepsilon(x - h_i(t))(-h'_i(t))$$
$$\partial_x w_\varepsilon(x - h_i(t)) = w'_\varepsilon(x - h_i(t))$$

give

$$\int_\Omega |u-v| \left(\partial_t \phi \prod_{1\le i\le N} w_\varepsilon(x-h_i(t)) + \phi \sum_{i=1}^N \prod_{1\le j\ne i\le N} (-h'_i(t)) w'_\varepsilon(x-h_i(t)) w_\varepsilon(x-h_j(t)) \right)$$
$$+ \int_\Omega \Phi(u,v) \left(\partial_x \phi \prod_{1\le i\le N} w_\varepsilon(x-h_i(t)) + \phi \sum_{i=1}^N \prod_{1\le j\ne i\le N} w'_\varepsilon(x-h_i(t)) w_\varepsilon(x-h_j(t)) \right)$$
$$+ \int_{\mathbb{R}} |u_0 - v_0| \big(\phi(0,x) \times \prod_{1\le i\le N} w_\varepsilon(x-h_i(0))\big) dx \ \ge\ 0.$$

Using that we know the form of the derivative of w_ε, namely $w'_\varepsilon(x - h_i(t)) = -\frac{2}{\varepsilon}\mathbf{1}_{[h_i-\varepsilon, h_i-\frac{\varepsilon}{2}]} + \frac{2}{\varepsilon}\mathbf{1}_{[h_i+\frac{\varepsilon}{2}, h_i+\varepsilon]}$, we can pass to the limit $\varepsilon \to 0$ and recognizing that in the sense of distributions

$$\lim_{\varepsilon\to 0} w_\varepsilon(x - h_i(t)) = \mathbf{1}$$

reincorporates the interfaces created by the particles and the related terms. Making use of the traces $\gamma_i^{\pm}(u)$, $\gamma_i^{\pm}(v)$, respectively, at the position of the interfaces $h_i(t)$, we obtain

$$\int_\Omega |u-v|\partial_t\phi + \Phi(u,v)\partial_x\phi\,dx\,dt + \int_{\mathbb{R}} |u_0 - v_0|\phi(0,x)\,dx$$
$$\geq \sum_{i=1}^{N} \int_0^T \left(\overline{\Phi}(h_i', \gamma_i^-(u), \gamma_i^-(v))\phi(h_i(s),s) - \overline{\Phi}(h_i', \gamma_i^+(u), \gamma_i^+(v))\phi(h_i(s),s)\right) ds.$$

Using the dissipativity of the germs for each particle, given by Lemma 3, we get the good signs of the right-side terms of the last inequality, which we then can drop to obtain the Kato inequality,

$$\int_\Omega |u-v|\partial_t\phi + \Phi(u,v)\partial_x\phi\,dx\,dt + \int_{\mathbb{R}} |u_0 - v_0|\phi(0,x)\,dx \geq 0,$$

which classically gives uniqueness of entropy solutions. Furthermore, integrating along the cone $C := \{(x,t), |x| = R + L(T-t), t \in [0,T]\}$ gives the L^1-contraction property.

References

1. B. Andreianov, N. Seguin, Well-posedness of a singular balance law. Dyn. Syst. Ser. A **32**, 1939–1964 (2012)
2. B. Andreianov, K.H. Karlsen, N.H. Risebro, A theory of L^1-dissipative solvers for scalar conservation laws with discontinuous flux. Arch. Ration. Mech. Anal. **201**, 26–86 (2011)
3. B. Andreianov, F. Lagoutière, N. Seguin, T. Takahashi, Well-posedness for a one-dimensional fluid-particle interaction model. SIAM J. Math. Anal. **46**(2), 1030–1052 (2014)
4. R. Borsche, M. Colombo, M. Garavello, On the coupling of systems of hyperbolic conservation laws with ordinary differential equations. Nonlinearity **23**(11), 2749–2770 (2010)
5. M.L. Delle Monache, P. Goatin, A front tracking method for a strongly coupled PDE-ODE system with moving density constraints in traffic flow. Discret. Contin. Dyn. Syst. Ser. S **7**(3), 435–447 (2014)
6. M.L. Delle Monache, P. Goatin, Scalar conservation laws with moving constraints arising in traffic flow modeling: an existence result. J. Differ. Equ. **257**, 4015–4029 (2014)
7. E. Isaacson, B. Temple, Convergence of the 2×2 Godunov method for a general resonant nonlinear balance law. SIAM J. Appl. Math. **55**(3), 625–640 (1955)
8. C. Klingenberg, J. Klotzky, N. Seguin, Well-posedness for a multi-particle fluid model (2017)
9. S.N. Kruzkov, First order quasilinear equations with several independent variables. Mat. Sb. **81**, 228–255 (1970)
10. F. Lagoutière, N. Seguin, T. Takahashi, A simple 1D model of inviscid fluid-solid interaction. J. Differ. Equ. **245**(11), 3503–3544 (2008)
11. A. Vasseur, Strong traces for solutions of multidimensional scalar conservation laws. A. Arch. Ration. Mech. Anal. **160**(3), 181–193 (2001)

On Quantifying Uncertainties for the Linearized BGK Kinetic Equation

Christian Klingenberg, Qin Li and Marlies Pirner

Abstract We consider the linearized BGK equation and want to quantify uncertainties in the case of modeling errors. More specifically, we want to quantify the error produced if the predetermined equilibrium function is chosen inaccurately. In this paper, we consider perturbations in the velocity and in the temperature of the equilibrium function and consider how much the error is amplified in the solution.

Keywords Linearized BGK equation · Uncertainty quantification
Perturbations in equilibrium distribution

1 Introduction

Kinetic equation is a set of integro-differential equations that describe the collective behavior of many-particle systems. The to-be-solved unknown function is a probability distribution of particles defined on the phase space, and kinetic equation characterizes its evolution in time and space. The equation typically has one transport term representing the movement of particles and one collision operator that describes the interactions between particles. The specific form of the transport and the collision operators depends on the system one is looking at. Typically people use radiative transfer equation for photon particles, the Boltzmann equation for rarified

C. Klingenberg (✉) · M. Pirner
Department of Mathematics at Würzburg University, Emil Fischer Str. 40, 97074 Würzburg, Germany
e-mail: klingen@mathematik.uni-wuerzburg.de
e-mail: marlies.pirner@mathematik.uni-wuerzburg.de

Q. Li
University of Wisconsin Madison, 480 Lincoln Dr., Madison, WI 53705, USA
e-mail: qinli@math.wisc.edu

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_14

gas particles, the Fokker–Planck equation for plasma, and run-and-tumble models for bacteria. There are many more other examples.

Uncertainty is a nature of kinetic theory. It has various of origins. The forms of terms in the equation are usually unjustified due to the modeling error, the blurred measurements are typically not enough to sufficiently determine the coefficients, and the initial and boundary conditions are never provided as accurate as they are supposed to be. They all contribute the inaccuracy of the system description. It is not realistic to look for the most accurate description of systems, nor expect the exact true solution, and thus, we instead look for possibilities of quantifying the uncertainties, and ask if the error is controllable even if the models and measurements are not accurate. As presented above, there are many origins of error, and in this paper, we focus on the modeling error. More specifically, a typical way of simplifying kinetic equations is to perform linearization around a predetermined equilibrium function and compute the linearized kinetic equation, and we would like to understand the error produced if the predetermined equilibrium function is chosen inaccurately. We plan to answer this question from both analytical point of view and numerical point of view. In particular, we would like to understand that given certain perturbation on the predetermined equilibrium where we perform the linearization, by how much the error is amplified in the solution, and how to characterize the perturbation numerically.

There have been many numerical techniques that were developed to address uncertainties. One very popular category of methods is termed generalized polynomial types. These include generalized polynomial chaos method (gPC) [12, 15, 16, 35], and stochastic collocation method [5, 34]. These methods assume the uncertainties in the parameters of the equations are reflected as a polynomial type in the solution. And based on this assumption, one applies either the spectral method, or the pseudo-spectral method, and expand the solution in the random direction using polynomials. Another popular or even classical method is the Monte Carlo type method, which also has many variations [6, 7, 13, 14]. With these methods, one simply samples the random variable many times, and for each sample, the parameters are fixed and the equation is considered deterministic, and one computes the equation. In the end, one ensembles the solutions for the mean and the variance. Sometimes, mathematicians categorize these methods based on if new implementations are needed. Since the Monte Carlo type method and stochastic collocation method simply call the deterministic solver many times, the old algorithms are therefore recycled and they are categorized as non-intrusive methods, while on the other hand, the traditional generalized polynomial chaos method is intrusive, wherein a completely new implementation is needed. In terms of the convergence rate, it is well known that the Monte Carlo method converges slowly, while the gPC type methods are spectral types along the random directions, and automatically inherits the so-called spectral convergence: Depending on the regularity of the solution in the random space, the method could be either algebraically fast or exponentially fast.

We would like to adopt the gPC framework for its possible fast convergence. To do that, in our setting, we mainly need to prove that the perturbation in the solution continuously depends on the perturbation in the equilibrium function where

we choose to perform linearization. According to the standard spectral method theory, the higher degree of continuity means the faster convergence. Traditionally, this framework has been successfully applied in treating elliptic type equation [3, 4, 9, 10, 36], and the analysis sometimes even suggests new algorithms that better explore the solution structure [1, 8, 11, 18, 19, 30–33], but when applied onto hyperbolic type equations, this framework sees limited success due to the intrinsic difficulties [2, 11]: The solution develops non-smooth structure, breaking the assumptions the spectral methods rely on.

The standard kinetic equation does not belong to either of the category mentioned above but could produce both. Depending on the regime one is interested in, kinetic equation would either converge to a hyperbolic type (such as BGK equation converging to the Euler equation) or a parabolic type (such as radiative transfer equation converging to the heat equation). On one hand, its transport term represents hyperbolic type and shows a traveling wave behavior; in the meantime, the collision term in kinetic equations is all coercive terms and thus provides some dissipative behavior and represents the parabolic type. This unique feature presents mathematicians a new world to explore and it indeed triggers many studies recently. Some recent results on the topic can be found in [17, 20–25]. We have to mention, however, most of the proofs are accomplished on a case-by-case basis, and not necessarily in their sharpest estimates, especially in the big space long time regime, except in [25] where the authors started with an abstract form and were able to employ the hypocoercivity for a uniform bound across regimes.

Follow the previous work, in this paper we explore the perturbation on the linearization point. We take the BGK equation as a starting point and perturb u, the bulk velocity, and T, the temperature in the equilibrium function, by z, a random variable. The domain of z indicates the strength of the perturbation. And we would like to study how f, the solution to the linearized equation, responds to the variations in z.

We lay out the equation and its basic assumptions in Sect. 2, together with detailed studies of the convergence rate in time in the deterministic setting. Sections 3 and 4 are respectively devoted to the study extended to equations in various of regimes, to equations involving randomness, and to scenarios when both present.

2 Setup

The BGK equation, known as a simplified model of the Boltzmann equation, writes as:

$$\partial_t F + v \cdot \nabla_x F = \frac{1}{\mathsf{Kn}} (M[F] - F) \tag{1}$$

where $F(t, x, v)$ is the distribution function living on phase space indicating the distribution of rarified gas. $M[F]$, the so-called Maxwellian function, is a Gaussian distribution function:

$$M[F] = \frac{\rho}{(2\pi T)^{d/2}} \exp^{-\frac{|v-u|^2}{2T}}, \tag{2}$$

with its macroscopic quantities defined implicitly by F such that the first $d+2$ moments are the same:

$$\int \phi(M[F] - F)\mathrm{d}v = 0\,, \tag{3}$$

with $\phi = [1, v, v^2]^T$. This property is typically called conservation property, since it immediately leads to density, momentum, and energy conservation:

$$\partial_t \int \phi F \mathrm{d}v + \nabla_x \int v \otimes v F \mathrm{d}v = 0\,. \tag{4}$$

If we use the definition:

$$\int F\mathrm{d}v = \rho(x)\,, \quad \int vF\mathrm{d}v = \rho(x)u(x)\,, \quad \text{and} \quad \int \frac{|v|^2}{2} F\mathrm{d}v = E = \frac{1}{2}\rho u^2 + \rho T\,. \tag{5}$$

then the first two equations express the conservation law of the density and momentum. Note that second term in the last equation cannot be presented using any macroscopic quantities, and thus, the system is not closed.

Kn is termed the Knudsen number. It comes from rescaling the system by setting $t \to \frac{t}{\mathsf{Kn}}$ and $x \to \frac{x}{\mathsf{Kn}}$. When Kn is small, the system is seen in large domain and long time scale and falls in the hyperbolic regime. More specifically, as $\mathsf{Kn} \to 0$, the leading term in the equation reads:

$$\frac{1}{\mathsf{Kn}}\left(M[F] - F\right) = 0 \quad \Rightarrow \quad F = M[F]\,, \tag{6}$$

and thus, $\int v|v|^2 F\mathrm{d}v$ could be explicitly expressed and we rewrite equation as:

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0 \\ \partial_t \rho u + \nabla_x (\rho u \otimes u + \rho T) = 0 \\ \partial_t E + \nabla_x \left((E + \rho T)u\right) = 0 \end{cases} \tag{7}$$

For linearization, we typically assume the solution is close enough to a particular Maxwellian, meaning there exists f and M_* such that:

$$F = (1 + f)M_*\,, \quad \text{with} \quad |f| \ll 1\,. \tag{8}$$

Plug this ansatz back into the full BGK equation and ignore the higher order expansion terms, we have:

$$\partial_t f + v \cdot \nabla_x f = \frac{1}{\mathsf{Kn}} \mathscr{L}_* f = \frac{1}{\mathsf{Kn}}\left(m[f] - f\right)\,,$$

where m is a quadratic function that shares the same moments with f, meaning:

$$\langle \phi \, , m - f \rangle_* = \int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} (m[f] - f) M_* \mathrm{d}v = 0 \, . \tag{9}$$

Here, we used the definition of the inner product:

$$\langle f, g \rangle_* = \int f g M_* \mathrm{d}v \, . \tag{10}$$

This is the counterpart of the conservation law in linearized system since:

$$\partial_t \int \phi f M_* \mathrm{d}v + \nabla_x \int v \otimes \phi f M_* \mathrm{d}v = 0 \, . \tag{11}$$

Once again if Kn is small, then in the leading order $f = m$ which leads to a closed Euler system, termed acoustic limit:

$$\partial_t U + A \cdot \partial_x U = 0 \, . \tag{12}$$

Here

$$A = \begin{pmatrix} u_* & \rho_* & 0 \\ \frac{T_*}{\rho_*} & u_* & 1 \\ 0 & 2T_* & u_* \end{pmatrix} \, , \quad \text{and} \quad U = [\tilde{\rho}, \tilde{u}, \tilde{T}]^T \, , \tag{13}$$

and the macroscopic quantities are defined by:

$$\int f \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} \mathrm{d}v = \begin{pmatrix} \tilde{\rho} \\ \tilde{\rho} u_* + \rho_* \tilde{u} \\ \tilde{\rho}(u_*^2 + T_*) + 2\rho_* u_* \tilde{u} + \rho_* \tilde{T} \end{pmatrix} . \tag{14}$$

There are several very well-known properties of the linear operator:

1 Coercive: $\langle \mathscr{L}_* f \, , f \rangle_* \leq 0$,
2 Explicit null space: $\mathscr{L}_* f = 0 \quad f \in \mathrm{Span}\{1, v, v^2\}$,
3 Self-adjoint: $\langle \mathscr{L}_* f \, , g \rangle_* = \langle f \, , \mathscr{L}_* g \rangle_*$.

Combining item 2 and 3, it is easy to see $\langle \mathscr{L}_* f \, , \phi \rangle_* = 0$. If we consider $f \in L_2(M_* \mathrm{d}v)$, one could express $\mathscr{L}_*$ more explicitly. By the definition of $m[f]$, it is easy to see that it is in fact a projection of f weighted by M_* on the quadratic function space:

$$\mathscr{L}_* f = m - f = \Pi_* f - f \, , \quad \text{with} \quad \Pi_* f = \sum_{i=0}^{d+1} \langle \chi_i \, , f \rangle_* \chi_i \, , \tag{15}$$

where χ_i are basis functions satisfying:

1 Expand the space $\text{Span}\{\chi_m, m = 0, \cdots d+1\} = \text{Span}\{1, v, v^2\}$,
2 Orthogonality $\langle \chi_m , \chi_n \rangle_* = \delta_{mn}$.

With the Maxwellian function M_* predetermined, they are simply the first $d+2$ Hermite polynomials associated with the Maxwellian. Even more if we set χ_m the mth Hermite polynomial for all m, then

$$\mathscr{L}_* f = -\sum_{m=d+2}^{\infty} \langle \chi_m , f \rangle_* \chi_m . \tag{16}$$

This expression also explicitly suggests the coercivity of the operator.

The linearized BGK operator has been studied by many researchers. Serving as the simplified version of the linearized Boltzmann equation. Its negative spectrum provides dissipative behavior, which helps us in getting existence and uniqueness of the solution at ease. In the boundary layer analysis, the nonlinear collision operator is far from being understood, the linearized equation is the stepping stone for connecting the Dirichlet data for the kinetic and the Dirichlet data for the interior Euler equation. We mention several recent work on boundary layer analysis for the linearized BGK equation here [26–29].

However, all these studies are based on the assumption that the Maxwellian M_*, the function we linearize upon, is given a priori, which is typically not the case. Taking numerical algorithm provided in [29] for example, we choose to perform linearization upon the Maxwellian function provided from the previous time step as an approximation to the true Maxwellian, which is in fact at least $\mathscr{O}(\Delta t)$ away from the real Maxwellian. A natural question one needs to address there is: Is such approximation a good approximation, or rather, if the Maxwellian chosen is off from the accurate one by $\mathscr{O}(\Delta t)$, how much error does f contain.

Since M_*'s dependence on ρ_* is linear, thus its reflection in f is of less interest. We in this paper only study the possible deviation of the solution f when M_* has a uncertain u_* and a uncertain T_*.

3 Variation in u

In this section, we study the solution's response to deviations in u_*. We firstly repeat the equation in 1D:

$$\begin{cases} \partial_t f + v\partial_x f = \mathscr{L}_* f , \quad (t, x, v) \in [0, \infty) \times \mathbb{R} \times \mathbb{R} \\ f(t = 0, x, v) = f_{\text{i}}(x, v) \end{cases} ,$$

with $\mathscr{L}_* f = m - f$ such that $\langle \phi , m - f \rangle_* = 0$, and f_{i} is the initial data. Assume the Maxwellian:

$$M_* = \frac{\rho_*}{\sqrt{2\pi T_*}} \exp\left(-\frac{|v - u_*|^2}{2T_*} \right) \tag{17}$$

and assume that f decays fast enough to zero as $x \to \infty$ such that $\langle \partial_x f, f \rangle_x = 0$.

With $u_*(z)$ depending on a random parameter z.[1] We would like to understand the regularity of the solution f on z direction; namely, we need to find a good bound for $\partial_z f$ in certain norm.

The standard way of pursuing such analysis is simply to take the derivative of z on the entire equation for an equation for $\partial_z f$, and then study the bound of $\partial_z f$. The bound could serve as a Lipschitz constant, and if small, numerical solvers that require certain regularities could be applied. Sometimes, people go beyond the first derivative and seek for high differentiation, and they are all bounded in a reasonable way, spectral method could be proved to be a effective method.

If we follow that procedure, however, the difficulty would be immediate: The random variable's dependence is hidden in the operator through $\mathscr{L}_*$ in a very subtle way. That means taking z derivative of the whole equation will produce very complicated formulation on the right-hand side. We thus choose a easy way that overcomes it by shifting the coordinates. Define

$$g(t, x, v) = f(t, x, v - u_*)\,, \tag{18}$$

then the equation for g will have a trivial collision but a shifted transport term:

$$\begin{cases} \partial_t g + (v + u_*)\partial_x g = \mathscr{L}_0 g \\ g(t = 0, x, v) = g_{\mathrm{i}}(x, v) = f_{\mathrm{i}}(x, v - u_*) \end{cases}, \tag{19}$$

with $\mathscr{L}_0$ being associated with the Maxwellian with zero velocity. The z dependence of the two functions could be easily written down:

$$\partial_z g = \partial_z f - \partial_v f \partial_z u_*\,, \quad \text{or} \quad \partial_z g + \partial_v f \partial_z u_* = \partial_z f\,. \tag{20}$$

Since $\partial_v f$ is more understood, for now we focus on studying $\partial_z g$. We take the derivative of the entire equation to get:

$$\partial_t \partial_z g + (v + u_*)\partial_x \partial_z g + \partial_z u_* \partial_x g = \mathscr{L}_0 \partial_z g\,,$$

or by defining $h = \partial_z g$ and reorganize the equation:

$$\partial_t h + (v + u_*)\partial_x h = \mathscr{L}_0 h - \partial_z u_* \partial_x g\,. \tag{21}$$

Immediately, we see that h satisfies also the linearized BGK equation but has one more negative source term $-\partial_z u_* \partial_x g$ compared with (19). To have a certain bound of h, we mainly need to go through two steps:

[1] for practical purpose, the range of z is controlled by Δt but we study the general case here

1 bound the source term: one needs to prove that the source term $\partial_z u_* \partial_x g$ is bounded;
2 bound h itself: here we need to show that a bounded $\partial_x g$ will produce a bounded h.

These two statements are summarized in the following two theorems.

Theorem 3.1. *$\|\partial_x g\|_2$ is bounded. More specifically:*

$$\|\partial_x g\|_{L^2(\mathrm{d}x\mathrm{d}v)}(t) \leq \|\partial_x g_\mathrm{i}\|_{L^2(\mathrm{d}x\mathrm{d}v)}$$

.

Proof. To show this, we first write down the equation for $\partial_x g$. Take the derivative of Eq. (19) with respect to x one gets:

$$\begin{cases} \partial_t \partial_x g + (v + u_*)\partial_x^2 g = \mathscr{L}_0 \partial_x g \\ \partial_x g(t = 0, x, v) = \partial_x g_\mathrm{i}(x, v) \end{cases} . \tag{22}$$

Here, we note that $\mathscr{L}_0$ is an operator on $\mathrm{d}v$ and commute with ∂_x. It immediately suggests that $\partial_x g$ satisfies the same equation as g in (19). Considering that the linearized BGK equation is a dissipative system and the L_2 norm decays in time, we cite the following lemma:

Lemma 3.1. *Suppose g satisfies equation (19), then*

$$\|g\|_{L^2(\mathrm{d}x\mathrm{d}v)}(t) \leq \|g_\mathrm{i}\|_{L^2(\mathrm{d}x\mathrm{d}v)} \tag{23}$$

where g_i is the initial condition.

Proof. The proof is based on energy estimate. We multiply the equation by g and integrate with respect to x and v, then:

$$\langle \partial_t g \, , \, g \rangle_{x,v} + \langle v \partial_x g \, , \, g \rangle_{x,v} = \langle \mathscr{L}_0 g \, , \, g \rangle_{x,v} \, . \tag{24}$$

Since we are considering the Cauchy problem, we throw the second term away. The term on the right-hand side is negative considering the coercivity of the collision operator. We then immediately get $\partial_t \langle g \, , \, g \rangle_{x,v} \leq 0$, meaning the L_2 norm of g decays in time and thus:

$$\|g\|_{L^2(\mathrm{d}x\mathrm{d}v)}(t) \leq \|g_\mathrm{i}\|_{L^2(\mathrm{d}x\mathrm{d}v)} \, . \tag{25}$$

□

Apply this lemma on (22), and we conclude with Theorem 3.1. □

With the boundedness of the source term $\partial_z u_* \partial_x g$, we could start analyzing the bound for h.

Theorem 3.2. *Suppose $h = \partial_z g$ satisfies (21), then $\|h\|_{L^2(\mathrm{d}x\mathrm{d}v)}$ grows at most linearly:*

$$\|h\|_{L^2(\mathrm{d}x\mathrm{d}v)} \lesssim C\|\partial_x g_{\mathrm{i}}\|_{L_2(\mathrm{d}x\mathrm{d}v)} t\,. \tag{26}$$

Here, $f \lesssim g$ means $\frac{f}{g}$ is bounded by a constant in large time. We care only about the long-time behavior of the solution. The reason is that after order one time, the highest order polynomial in time dominates the lower orders, and thus, one only needs to specify the highest order coefficient.

Proof. It is once again energy method. We multiply (21) on both sides with h and take the inner product in (x, v):

$$\langle \partial_t h\,, h\rangle_{x,v} = \langle \mathscr{L}_0 h\,, h\rangle_{x,v} - \partial_z u_* \langle \partial_x g\,, h\rangle_{x,v}\,. \tag{27}$$

Considering the coercivity of $\mathscr{L}_0$ the first term on the right disappear. And we use Cauchy–Schwartz inequality to control the second term to get:

$$\frac{1}{2}\frac{d}{dt}\|h\|^2_{L^2(\mathrm{d}x\mathrm{d}v)} \le \|\partial_z u_* \partial_x g\|_{L^2(\mathrm{d}x\mathrm{d}v)}\|h\|_{L^2(\mathrm{d}x\mathrm{d}v)}\,. \tag{28}$$

Assume $|\partial_z u_*| < C$, and it is known from Theorem 3.1 that

$$\|\partial_x g\|_{L^2(\mathrm{d}x\mathrm{d}v)} \le \|\partial_x g_{\mathrm{i}}\|_{L^2(\mathrm{d}x\mathrm{d}v)},$$

then

$$\frac{d}{dt}\|h\|_{L^2(\mathrm{d}x\mathrm{d}v)} \le C\|\partial_x g_{\mathrm{i}}\|_{L^2(\mathrm{d}x\mathrm{d}v)} \tag{29}$$

which leads to a linear growth of h: $\|h\|_{L^2(\mathrm{d}x\mathrm{d}v)} \lesssim C\|\partial_x g_{\mathrm{i}}\|_{L^2(\mathrm{d}x\mathrm{d}v)} t$. □

The theorem above states the bounded of the first derivative of g in z. One could extend it to treat higher order derivatives.

Theorem 3.3. *Denote $h^{(n)} = \partial_z^n g$, then $\|h^{(n)}\|_{L^2(\mathrm{d}x\mathrm{d}v)}$ is bounded by t^n:*

$$\|h^{(n)}\|_{L^2(\mathrm{d}v\mathrm{d}x)} \lesssim C_n t^n\,. \tag{30}$$

Again we are mainly interested in the long-time behavior of the solution so it suffices to consider only the highest order in time.

Proof. The proof is based on induction. According to the definition, $h^{(0)} = g$ and Lemma 3.1 guarantees that $h^{(0)}$ is bounded by a constant, and $h^{(1)}$ is the h in Theorem 3.2, and we have seen it is bounded by a linear growth. We thus perform math induction, assuming $h^{(k-1)}$ is bounded by t^{k-1} we show that $h^{(k)}$ is bounded by t^k.

We first take the kth-order derivative of Eq. (19):

$$\partial_t \partial_z^k g + \sum_{n=0}^{k} \binom{k}{n} \partial_z^n (v + u_*) \partial_x \partial_z^{k-n} g = \mathscr{L}_0 \partial_z^k g \,,$$

or moving the source term to the right:

$$\partial_t h^{(k)} + (v + u_*) \partial_x h^{(k)} = \mathscr{L}_0 h^{(k)} - \sum_{n=1}^{k} \binom{k}{n} \partial_z^n u_* \partial_x h^{(k-n)} \,.$$

According to our assumption, $\partial_z^n u_*$ is bounded by a constant, one has:

$$\begin{aligned} \langle \partial_t h^{(k)} \,, h^{(k)} \rangle_{x,v} + \langle (v + u_*) \partial_x h^{(k)} \,, h^{(k)} \rangle_{x,v} &= \langle \mathscr{L}_0 h^{(k)} \,, h^{(k)} \rangle_{x,v} \\ &\quad - \sum_{n=1}^{k} \binom{k}{n} \langle \partial_z^n u_* \partial_x h^{(k-n)} \,, h^{(k)} \rangle_{x,v} \,. \end{aligned}$$

which means:

$$\frac{1}{2} \frac{d}{dt} \| h^{(k)} \|_{L_2(\mathrm{d}x\mathrm{d}v)}^2 \leq C_k \| \partial_x h^{(k-n)} \|_{L_2(\mathrm{d}x\mathrm{d}v)} \| h^{(k)} \|_{L_2(\mathrm{d}x\mathrm{d}v)} \,, \tag{31}$$

where we used the Cauchy boundary condition, the coercivity of $\mathscr{L}_0$, and Cauchy–Schwartz inequality. By our assumption $h^{(k-1)}$ is bounded by t^{k-1}, since $\partial_x h$ and h satisfies the same equation, it can be extrapolated as $\partial_x h$ being bounded by the same order, and then putting it back into (31), we have:

$$\| h^{(k)} \|_{L_2(\mathrm{d}x\mathrm{d}v)} \lesssim t^k \,, \tag{32}$$

which finishes the math induction loop, and complete the proof. □

4 Variation in T

In this section, we want to study the solution's response to the deviations in T_*. Namely, we assume the Maxwellian defined in (17) has its $T_*(z)$ depending on a random parameter z. Once again, in order to get rid of the complicated dependence of $\mathscr{L}_*$ on z, we perform change of variable and define

$$p(t, x, v) = f\left(t, x, \frac{v}{\sqrt{T_*}}\right) \,. \tag{33}$$

Then p satisfies the equation

$$\begin{cases} \partial_t p + \sqrt{T_*} v \partial_x p = \mathscr{L}_1 p \\ p(t=0,x,v) = p_{\mathrm{i}}(x,v) = f_{\mathrm{i}}\left(x, \frac{v}{\sqrt{T_*}}\right) \end{cases}, \tag{34}$$

where $\mathscr{L}_1$ is the collision operator associated with the Maxwellian with temperature one, and p_{i} is the initial data. Again we focus on studying $\partial_z p$ instead of $\partial_z f$. Denote $q = \partial_z p$, we obtain its governing equation by taking the derivative in z of Eq. (34). Rearranging the terms, we have:

$$\partial_t q + \sqrt{T_*} v \partial_x q = \mathscr{L}_1 q - \partial_z(\sqrt{T_*}) v \partial_x p\,. \tag{35}$$

This equation has the same structure as equation (21): It is a linearized kinetic equation with a source term, and for the boundedness of q, we simply need to show the boundedness of $v\partial_x p$. In the previous section, we showed that the source term $\partial_x g$ satisfies the same equation as g does and thereby was able to give the bound. This is no longer the case here. Instead of writing the equation, we write:

$$v\partial_x p = \frac{\mathscr{L}_1 p - \partial_t p}{\sqrt{T_*}}\,, \tag{36}$$

and are able to prove the following:

Theorem 4.1. *Suppose $q = \partial_z p$ satisfies (35), then $||q||_{L^2(\mathrm{d}x\mathrm{d}v)}$ grows at most linearly:*

$$||q||_{L^2(\mathrm{d}x\mathrm{d}v)} \lesssim C\left(||p_i||_{L^2(\mathrm{d}x\mathrm{d}v)} + ||\partial_t p_i||_{L^2(\mathrm{d}x\mathrm{d}v)}\right) t$$

Proof. We once again use the energy method. We insert (36) into (35) and multiply the obtained equation with q and take the inner product in (x,v):

$$\langle \partial_t q\,, q\rangle_{x,v} = \langle \mathscr{L}_1 q\,, q\rangle_{x,v} - \frac{\partial_z\sqrt{T_*}}{\sqrt{T_*}}\langle (\mathscr{L}_1 p - \partial_t p)\,, q\rangle_{x,v}. \tag{37}$$

Due to the coercivity of $\mathscr{L}_1$, the first term on the right disappears. For the second term on the right, we use Cauchy–Schwarz and the triangle inequality

$$\frac{1}{2}\frac{d}{dt}||q||^2_{L^2(\mathrm{d}x\mathrm{d}v)} \leq \left|\frac{\partial_z\sqrt{T_*}}{\sqrt{T_*}}\right| \left(||\mathscr{L}_1 p||_{L^2(\mathrm{d}x\mathrm{d}v)} + ||\partial_t p||_{L^2(\mathrm{d}x\mathrm{d}v)}\right) ||q||_{L^2(\mathrm{d}x\mathrm{d}v)} \tag{38}$$

We assume $\left|\frac{\partial_z\sqrt{T_*}}{\sqrt{T_*}}\right| < C$. Similar to $\Pi_x f$ in (15), $\Pi_1 g$ can be represented as

$$\Pi_1 f = \sum_{i=0}^{d+1} \langle \chi_i^0, g\rangle_1 \chi_i^0$$

with orthonormal basis fuctions χ_i^0, where $\langle\cdot\rangle_1$ denotes integration with respect to v with the weight M_1. Then, $||\mathscr{L}_1 p||_{L^2(M_1 \mathrm{d}x\mathrm{d}v)}$ can be estimated by above using the explicit expression of $\mathscr{L}_1$ and Cauchy–Schwartz inequality by

$$\begin{aligned}\langle \mathscr{L}_1 p, \mathscr{L}_1 p\rangle_{x,v,1} &= \langle \sum_{i=1}^{d+1} \langle \chi_i, p\rangle_{x,v,1}\ \chi_i - p, \sum_{j=1}^{d+1} \langle \chi_j, p\rangle_{x,v,1}\ \chi_j - p\rangle_{x,v,1} \\ &= \sum_{i=1}^{d+1} (\langle \chi_i, p\rangle_{x,v,1})^2 - \langle p, p\rangle_{x,v,1} \leq (d+1)\langle p, p\rangle_{x,v,1} - \langle p, p\rangle_{x,v,1} \\ &= d\langle p, p\rangle_{x,v,1} \end{aligned} \tag{39}$$

Since the norm $||\cdot||_{L^2(M_1 \mathrm{d}x\mathrm{d}v)}$ is equivalent to $||\cdot||_{L^2(\mathrm{d}x\mathrm{d}v)}$, the term $||\mathscr{L}_1 p||_{L^2(\mathrm{d}x\mathrm{d}v)}$ is also bounded by $C||p||_{L^2(\mathrm{d}x\mathrm{d}v)}$.

Realizing that $\partial_t p$ satisfies the same equation as p does, according to Lemma (4.1), their L_2 norm decrease in time, meaning:

$$\frac{d}{dt}||q||_{L^2(\mathrm{d}x\mathrm{d}v)} \leq C\left(||p||_{L^2(\mathrm{d}x\mathrm{d}v)}(t) + ||\partial_t p||_{L^2(\mathrm{d}x\mathrm{d}v)}(t)\right) \tag{40}$$

$$\leq C\left(||p_i||_{L^2(\mathrm{d}x\mathrm{d}v)} + ||\partial_t p_i||_{L^2(\mathrm{d}x\mathrm{d}v)}\right), \tag{41}$$

which leads to a linear growth of q:

$$||q||_{L^2(\mathrm{d}x\mathrm{d}v)} \lesssim C\left(||p_i||_{L^2(\mathrm{d}x\mathrm{d}v)} + ||\partial_t p_i||_{L^2(\mathrm{d}x\mathrm{d}v)}\right) t\,. \tag{42}$$

which concludes the proof. □

The lemma used in the theorem is stated in the following:

Lemma 4.1. *Suppose p satisfies equation (34), then*

$$\|p\|_{L^2(\mathrm{d}x\mathrm{d}v)}(t) \leq \|p_\mathrm{i}\|_{L^2(\mathrm{d}x\mathrm{d}v)} \tag{43}$$

where p_i is the initial condition.

Proof. The proof is analogous to the proof of Lemma 3.1. □

We can also extend the result of Theorem 4.1 to derivatives of higher orders. This is done in the following theorem

Theorem 4.2. *Suppose $q^{(n)} := \partial_z^n p$ satisfies*

$$\partial_t q^{(n)} + \sum_{k=0}^{n} \binom{n}{k} \partial_z^{(n-k)} \left(\sqrt{T_*}\right) v \partial_x q^{(k)} = \mathscr{L}_1 q^{(n)} \tag{44}$$

for all $n \in \mathbb{N}_0$. Then

$$||q^{(N)}||_{L^2(\mathrm{d}x\mathrm{d}v)} \lesssim C_N\, t^N \quad \textit{for all } N \leq n$$

where C_N *depends on* $||\mathscr{L}_1^k \partial_t^l q^{(0)}||_{L^2(\mathrm{d}x\mathrm{d}v)},\ k, l \leq N$.

Proof. We proof the statement via induction. For $n = 0$ and $n = 1$, we proved it in Theorem 4.1 and Lemma 4.1. We have shown in Lemma 4.1 that if $q^{(0)}$ satisfies (34), then $||q^{(0)}||_{L^2(\mathrm{d}x\mathrm{d}v)}$ is bounded by $||q_i^0||_{L^2(\mathrm{d}x\mathrm{d}v)}$, and in Theorem 4.1 if $q^{(1)}$ satisfies (35), we can replace $v\partial_x q^{(0)}$ by (36). We can show that $\mathscr{L}_1 q^{(0)}$ is bounded in $L^2(\mathrm{d}x\mathrm{d}v)$ by $||p||_{L^2(\mathrm{d}x\mathrm{d}v)}$ and $\partial_t q^{(0)}$ also satisfy (34) and can deduce that $||q^{(1)}||_{L^2(\mathrm{d}x\mathrm{d}v)}$ is bounded by $C(||q_i^{(0)}||_{L^2(\mathrm{d}x\mathrm{d}v)} + ||\partial_t q_i^{(0)}||_{L^2(\mathrm{d}x\mathrm{d}v)})t$, see the proof of Theorem 4.1. Assume now that the statement is true for a fixed $n \in \mathbb{N}$. We want to deduce that it is true for $n + 1$. If $q^{(n+1)}$ satisfies

$$\partial_t q^{(n+1)} + \sum_{k=0}^{n+1} \binom{n+1}{k} \partial_z^{(n+1-k)} \left(\sqrt{T_*}\right) v\partial_x q^{(k)} = \mathscr{L}_1 q^{(n+1)} \tag{45}$$

we can replace $v\partial_x q^{(n)}$ in terms of $\partial_t q^{(n)}$, $\mathscr{L}_1 q^{(n)}$, $v\partial_x q^{(N)}$, $N < n$ from the equation for $q^{(n)}$ given by (44). In the resulting equation, we can replace $v\partial_x q^{(n-1)}$ in terms of $\partial_t q^{(n-1)}$, $\mathscr{L}_1 q^{(n-1)}$, $v\partial_x q^{(N)}$, $N < n - 1$ from the equation for $q^{(n-1)}$. Next, we can replace $v\partial_x q^{(n-2)}$ from the equation for $q^{(n-2)}$ and so on until we do not have terms with $v\partial_x q^{(k)}$ for some $k < n + 1$ any more. So all in all, we obtain an equation of the form

$$\partial_t q^{(n+1)} + A(\partial_t q^{(0)}, \mathscr{L}_1 q^{(0)}, \ldots, \partial_t q^{(n)}, \mathscr{L}_1 q^{(n)}, T_*) + v\partial_x q^{(n+1)} = \mathscr{L}_1 q^{(n+1)} \tag{46}$$

where A is a linear combination of $\partial_t q^{(0)}, \mathscr{L}_1 q^{(0)}, \ldots, \partial_t q^{(n)}, \mathscr{L}_1 q^{(n)}$ with coefficients depending on T_* of the form

$$\frac{(\partial_z^a(\sqrt{T_*}))^b}{\sqrt{T_*}^c} \quad \text{for } a, b, c \leq n + 1 \tag{47}$$

We can show that $\partial_t q^{(N)}$, $N \leq n$ satisfy the same equation as $q^{(N)}$ similar as it is done in Sect. 3 for $\partial_x g$ and g and $\mathscr{L}_1 q^{(N)}$, $N \leq n$ is bounded in $L^2(\mathrm{d}x\mathrm{d}v)$ by $||q^{(N)}||_{L^2(\mathrm{d}x\mathrm{d}v)}$, and that they are bounded in $L^2(\mathrm{d}x\mathrm{d}v)$ by $C_N t^N$ where C_N depends on $||\mathscr{L}_1^k \partial_t^l q^{(0)}||_{L^2(\mathrm{d}x\mathrm{d}v)}$, $k, l \leq N$ due to the induction assumption. Finally, by the energy method we can deduce from (46) that $q^{(n+1)}$ is bounded in $L^2(\mathrm{d}x\mathrm{d}v)$ by $C_{n+1}t^{n+1}$. □

References

1. G. Albi, L. Pareschi, M. Zanella, Uncertainty quantification in control problems for flocking models. Math. Probl. Eng. **850124**, 14 (2015)

2. M. Branicki, A.J. Majda, Fundamental limitations of polynomial chaos for uncertainty quantification in systems with intermittent instabilities. Commun. Math. Sci. **11**(1), 55–103 (2013)
3. I. Babuška, F. Nobile, R. Tempone, Galerkin finite element approximations of stochastic elliptic partial differential equations. SIAM J. Numer. Anal. **42**(2), 800–825 (2004)
4. I. Babuška, F. Nobile, R. Tempone, A stochastic collocation method for elliptic partial differential equation with random input data. SIAM J. Numer. Anal. **45**(3), 1005–1034 (2007)
5. I. Babuška, F. Nobile, R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal. **45**(3), 1005–1034 (2007)
6. A. Barth, C. Schwab, N. Zollinger, Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. Numer. Math. **119**(1), 123–161 (2011)
7. J. Charrier, R. Scheichl, A.L. Teckentrup, Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods. SIAM J. Numer. Anal. **51**(1), 322–352 (2013)
8. A. Chkifa, A. Cohen, C. Schwab, Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. Journal de Mathématiques Pures et Appliquées (2014)
9. A. Cohen, R. DeVore, C. Schwab, Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. Found. Comput. Math. **10**(6), 615–646 (2010)
10. A. Cohen, R. Devore, C. Schwab, Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. Anal. Appl. **9**(01), 11–47 (2011)
11. B. Despres, B. Perthame, Uncertainty propagation; intrusive kinetic formulations of scalar conservation laws. SIAM/ASA J. Uncertain. Quantif. **4**(1), 980–1013 (2016)
12. D. Xiu, G. Karniadakis, The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002)
13. G. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications* (Springer, New York, 2013)
14. M.B. Giles, Multilevel Monte Carlo path simulation. Op. Res. **56**(3), 607–617 (2008)
15. R.G. Ghanem, A. Doostan, On the construction and analysis of stochastic models: characterization and propagation of the errors associated with limited data. J. Comput. Phys. **217**(1), 63–81 (2006)
16. R.G. Ghanem, R.M. Kruger, Numerical solution of spectral stochastic finite element systems. Comput. Methods Appl. Mech. Eng. **129**(3), 289–303 (1996)
17. J. Hu, S. Jin, A stochastic Galerkin method for the Boltzmann equation with uncertainty. J. Comput. Phys. **315**, 150–168 (2016)
18. Y.T. Hou, Q. Li, P. Zhang, Exploring the locally low dimensional structure in solving random elliptic PDEs. SIAM Multiscale Model. Simul. (2016)
19. Y.T. Hou, Q. Li, P. Zhang, A sparse decomposition of low rank symmetric positive semi-definite matrices. SIAM Multiscale Model. Simul. (2016)
20. S. Jin, L. Liu, An asymptotic-preserving stochastic Galerkin method for the semiconductor Boltzmann equation with random inputs and diffusive scalings. *SIAM Multiscale Model. Simul.* (2016)
21. S. Jin, H. Lu, An Asymptotic-Preserving stochastic Galerkin method for the radiative heat transfer equations with random inputs and diffusive scalings (2016)
22. S. Jin, Y. Zhu, The Vlasov-Poisson-Fokker-Planck system with uncertainty and a one-dimensional asymptotic-preserving method (2016)
23. S. Jin, D. Xiu, X. Zhu, Asymptotic-preserving methods for hyperbolic and transport equations with random input and diffusive scalings. J. Comput. Phys. **289**, 35–52 (2015)
24. S. Jin, J.G. Liu, Z. Ma, Uniform spectral convergence of the stochastic galerkin method for the linear transport equations with random inputs in diffusive regime and a micro-macro decomposition based asymptotic preserving method (2016)
25. Q. Li, L. Wang, Uniform regularity for linear kinetic equations with random input based on hypocoercivity (2016). arXiv:1612.01219
26. Q. Li, J. Lu, W. Sun, A convergent method for linear half-space kinetic equation. Math. Model. Numer. Anal. (in press)
27. Q. Li, J. Lu, W. Sun, Half-space kinetic equations with general boundary conditions. Math. Comput. (in press)

28. Q. Li, J. Lu, W. Sun, Validity and regularization of classical half-space equations. J. Stat. Phys. (in press)
29. Q. Li, J. Lu, W. Sun, Diffusion approximations of linear transport equations: asymptotics and numerics. J. Comput. Phys. **292**, 141–167 (2015)
30. F. Nobile, R. Tempone, C. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(3), 2309–2345 (2008)
31. F. Nobile, R. Tempone, C.G. Webster, An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2411–2442 (2008)
32. F. Nobile, R. Tempone, C.G. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2309–2345 (2008)
33. C. Schwab, R.-A. Todor, Sparse finite elements for elliptic problems with stochastic loading. Numer. Math. **95**(4), 707–734 (2003)
34. D. Xiu, J.S. Hesthaven, High-order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. **27**(3), 1118–1139 (2005)
35. D. Xiu, G.E. Karniadakis, Modeling uncertainty in flow simulations via generalized polynomial chaos. J. Comput. Phys. **187**(1), 137–167 (2003)
36. G. Zhang, M. Gunzburger, Error analysis of a stochastic collocation method for parabolic partial differential equations with random input data. SIAM J. Numer. Anal. **50**(4), 1922–1940 (2012)

Kinetic ES-BGK Models for a Multi-component Gas Mixture

Christian Klingenberg, Marlies Pirner and Gabriella Puppo

Abstract We consider a multi-component mixture of inert gas in the kinetic regime by assuming that the total number of particles of each species remains constant. In this article, we shall illustrate our model for the case of two species. To account for thermal effects, we extend a BGK model based on the presence of a collision term for each possible interaction (Klingenberg et al., A consistent kinetic model for a two-component mixture with an application to plasma. Kinet Relat Models 10:444–465, 2017, [19]) by including ES-BGK effects. We prove consistency of the extended model like conservation properties, positivity of all temperatures, H-theorem, and convergence to a global equilibrium in the shape of a global Maxwell distribution.

Keywords Multi-fluid mixture · Kinetic model · ES-BGK equation · H-theorem

Introduction

In this paper, we shall concern ourselves with a kinetic description of gases. This is traditionally done via the Boltzmann equation for the density distributions f_1 and f_2. Under certain assumptions, the complicated interaction terms of the Boltzmann equation can be simplified by a so-called BGK approximation, consisting of a collision frequency multiplied by the deviation of the distributions from local Maxwellians. This approximation should be constructed in a way such that it has the same main properties of the Boltzmann equation, namely conservation of mass, momentum, and energy, further it should have an H-theorem with its entropy inequality and the

C. Klingenberg (✉) · M. Pirner
Department of Mathematics at Würzburg University, Emil Fischer Str. 40, 97074 Würzburg, Germany
e-mail: klingen@mathematik.uni-wuerzburg.de

M. Pirner
e-mail: marlies.pirner@mathematik.uni-wuerzburg.de

G. Puppo
Università degli Studi dell'Insubria, Via Valleggio 11, 22100 Como, Italy
e-mail: gabriella.puppo@uninsubria.it

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_15

equilibrium must still be Maxwellian. BGK models give rise to efficient numerical computations, which are asymptotic preserving, that is they remain efficient even approaching the hydrodynamic regime [6, 7, 10–12, 20]. However, the drawback of the BGK approximation is its incapability of reproducing the correct Boltzmann hydrodynamic regime in the asymptotic continuum limit. Therefore, a modified version called ES-BGK approximation was suggested by Holway in the case of one species [16]. The H-Theorem of this model then was proven in [3] and existence and uniqueness of solutions in [21].

Here, we shall focus on gas mixtures modeled via an ES-BGK approach. In the literature, there is a BGK model for gas mixtures suggested by Andries, Aoki, and Perthame in [4] which contains only one collision term on the right-hand side. Extensions of this model to an ES-BGK model for gas mixtures are given by Groppi in [14] or the model by Brull [5] with an extension leading to a correct Prandtl number in the Navier–Stokes equation, adapting the ES-BGK model for mixtures.

In this paper, we are interested in an extension to an ES-BGK model of a BGK model for gas mixtures [19] which just like the Boltzmann equation for gas mixtures contains a sum of collision terms on the right-hand side. Other examples of ES-BGK models for gas mixtures are the models of Gross and Krook [13], Hamel [15], Asinari [1]. The advantage of this extended model is that we have free parameters to possibly being able to determine macroscopic physical constants like viscosity or heat conductivity when taking the limit to the Navier–Stokes equations.

The outline of the paper is as follows: in Sect. 1, we will present the BGK model for two species developed in [19]. In Sect. 2, we suggest extensions to an ES-BGK model for mixtures and prove the corresponding H-Theorem.

1 The BGK Approximation

In this section, we will present the BGK model for a mixture of two species and mention its fundamental properties like the conservation properties and the H-theorem.

For simplicity in the following, we consider a mixture composed of two different species, but the discussion can be generalized to multi-species mixtures. Thus, our kinetic model has two distribution functions $f_1(x, v, t) > 0$ and $f_2(x, v, t) > 0$ where $x \in \Lambda \subset \mathbb{R}^3$ and $v \in \mathbb{R}^3$ are the phase space variables and $t \geq 0$ the time. The distribution functions are determined by two equations to describe their time evolution. Furthermore, we only consider binary interactions. So the particles of one species can interact with either themselves or with particles of the other species. In the model, this is accounted for introducing two interaction terms in both equations. These considerations allow us to write formally the system of equations for the evolution of the mixture. The following structure containing a sum of the collision operator is also given in [8, 9].

Furthermore, for any $f_1, f_2 : \Lambda \subset \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}$ with $(1 + |v|^2) f_1, (1 + |v|^2) f_2 \in L^1(\mathbb{R}^3)$, $f_1, f_2 \geq 0$ we relate the distribution functions to macroscopic quantities by mean-values of $f_k, k = 1, 2$

$$\int f_k(v) \begin{pmatrix} 1 \\ v \\ m_k|v-u_k|^2 \\ m_k(v-u_k(x,t)) \otimes (v-u_k(x,t)) \end{pmatrix} dv =: \begin{pmatrix} n_k \\ n_k u_k \\ 3n_k T_k \\ \mathbb{P}_k \end{pmatrix}, \quad k = 1, 2, \quad (1)$$

where n_k is the number density, u_k the mean velocity, and T_k the mean temperature of species k, $k = 1, 2$. Note that in this paper, we shall write T_k instead of $k_B T_k$, where k_B is Boltzmann's constant.

We are interested in a BGK approximation of the interaction terms. This leads us to define equilibrium distributions not only for each species itself but also for the two interspecies equilibrium distributions. We choose the collision terms as BGK operators and denote them for future references by Q_{11}, Q_{12}, Q_{21}, and Q_{22}. Then the model can be written as:

$$\begin{aligned} \partial_t f_1 + \nabla_x \cdot (v f_1) &= \nu_{11} n_1 (M_1 - f_1) + \nu_{12} n_2 (M_{12} - f_1), \\ \partial_t f_2 + \nabla_x \cdot (v f_2) &= \nu_{22} n_2 (M_2 - f_2) + \nu_{21} n_1 (M_{21} - f_2), \end{aligned} \quad (2)$$

with the Maxwell distributions

$$\begin{aligned} M_k(x,v,t) &= \frac{n_k}{\sqrt{2\pi \frac{T_k}{m_k}}^3} \exp\left(-\frac{|v-u_k|^2}{2\frac{T_k}{m_k}}\right), \quad k = 1, 2, \\ M_{kj}(x,v,t) &= \frac{n_{kj}}{\sqrt{2\pi \frac{T_{kj}}{m_k}}^3} \exp\left(-\frac{|v-u_{kj}|^2}{2\frac{T_{kj}}{m_k}}\right), \quad k, j = 1, 2,\ k \neq j, \end{aligned} \quad (3)$$

where $\nu_{11} n_1$ and $\nu_{22} n_2$ are the collision frequencies of the particles of each species with itself, while ν_{12} and ν_{21} are related to interspecies collisions. To be flexible in choosing the relationship between the collision frequencies, we now assume the relationship

$$\nu_{12} = \varepsilon \nu_{21}, \qquad 0 < \varepsilon \leq 1, \quad (4)$$

$$\nu_{11} = \beta_1 \nu_{12}, \quad \nu_{22} = \beta_2 \nu_{21}, \qquad \beta_1, \beta_2 > 0. \quad (5)$$

The restriction $\varepsilon \leq 1$ is without loss of generality. If $\varepsilon > 1$, exchange the notation 1 and 2 and choose $\frac{1}{\varepsilon}$. In addition, we assume that all collision frequencies are positive.

The structure of the collision terms ensures that if one collision frequency $\nu_{kl} \to \infty$, the corresponding distribution function becomes a Maxwell distribution. In addition at global equilibrium, the distribution functions become Maxwell distributions with the same velocity and temperature (see Sect. 2.8 in [19]). The Maxwell distributions M_1 and M_2 in (3) have the same moments as f_1 and f_2, respectively. With this choice, we guarantee the conservation of mass, momentum and energy in interactions of one species with itself (see Sect. 2.2 in [19]). The remaining parameters $n_{12}, n_{21}, u_{12}, u_{21}, T_{12}$, and T_{21} will be determined using conservation of total momentum and energy, together with some symmetry considerations.

If we assume that

$$n_{12} = n_1 \quad \text{and} \quad n_{21} = n_2, \tag{6}$$

$$u_{12} = \delta u_1 + (1-\delta)u_2, \quad \delta \in \mathbb{R}, \tag{7}$$

and

$$T_{12} = \alpha T_1 + (1-\alpha)T_2 + \gamma |u_1 - u_2|^2, \quad 0 \le \alpha \le 1, \gamma \ge 0, \tag{8}$$

we have conservation of the number of particles, total momentum, and total energy provided that

$$u_{21} = u_2 - \frac{m_1}{m_2}\varepsilon(1-\delta)(u_2 - u_1), \tag{9}$$

and

$$\begin{aligned} T_{21} = \left[\frac{1}{3}\varepsilon m_1(1-\delta)\left(\frac{m_1}{m_2}\varepsilon(\delta-1)+\delta+1\right) - \varepsilon\gamma\right]|u_1-u_2|^2 \\ +\varepsilon(1-\alpha)T_1 + (1-\varepsilon(1-\alpha))T_2, \end{aligned} \tag{10}$$

see Theorems 2.1, 2.2, and 2.3 in [19].

We see that without using an ES-BGK extension, we already have three free parameters in (7) and (8) in order to match coefficients like the Fick's constant or the heat conductivity in the Navier–Stokes equations. But when we derive the Navier–Stokes equations by a Chapman–Enskog expansion $f_k = f_k^0 + \tilde{\varepsilon} f_k^1 + \tilde{\varepsilon}^2 f_k^2 + \cdots$, one can show that $|u_1 - u_2|^2$ is of order $\tilde{\varepsilon}^2$, so γ from (8) does not appear in the first order Navier–Stokes equations and therefore cannot be used to match parameters there.

In order to ensure the positivity of all temperatures, we need to impose restrictions on δ and γ,

$$0 \le \gamma \le \frac{m_1}{3}(1-\delta)\left[(1+\frac{m_1}{m_2}\varepsilon)\delta + 1 - \frac{m_1}{m_2}\varepsilon\right], \tag{11}$$

and

$$\frac{\frac{m_1}{m_2}\varepsilon - 1}{1+\frac{m_1}{m_2}\varepsilon} \le \delta \le 1, \tag{12}$$

see Theorem 2.5 in [19].

This summarizes our kinetic model (2) in of two species that contains three free parameters. More details can be found in [19].

2 Extensions to an ES-BGK Approximation

2.1 Extension of the Single Relaxation Terms

Motivated by the need to find a two species kinetic model that allows us to model physical parameters better we extend the above model by generalizing the Maxwellians. The simplest choice is to only replace the collision operators which represent the collisions of a species with itself by the ES-BGK collision operator for one species suggested in [2]. Then the model can be written as:

$$\partial_t f_k + \nabla_x \cdot (v f_k) = \nu_{kk} n_k (G_k - f_k) + \nu_{kj} n_j (M_{kj} - f_k), \quad k, j = 1, 2, \ j \neq k, \tag{13}$$

with the modified Maxwell distributions

$$G_k(x, v, t) = \frac{n_k}{\sqrt{det(2\pi \frac{\mathscr{T}_k}{m_k})}} \exp(-\frac{1}{2}(v - u_k) \cdot (\frac{\mathscr{T}_k}{m_k})^{-1} \cdot (v - u_k)), \quad k = 1, 2, \tag{14}$$

and M_{12}, M_{21} the Maxwellians described in the previous section. G_1 and G_2 have the same densities, velocities, and pressure tensors as f_1 respective f_2, so we still guarantee the conservation of mass, momentum, and energy in interactions of one species with itself. Since the first term describes the interactions of a species with itself, it should correspond to the single ES-BGK collision operator suggested in [2]. So we choose $\mathscr{T}_1$ and $\mathscr{T}_2$ as

$$\mathscr{T}_k = (1 - \mu_k) T_k \mathbf{1} + \mu_k \frac{\mathbb{P}_k}{n_k}, \tag{15}$$

with $\mu_k \in \mathbb{R}$, $k = 1, 2$ being free parameters which we can choose in a way to fix physical parameters in the Navier–Stokes equations. So, all in all, together with the parameters in the mixture Maxwellians (7) and (8), we now have five free parameters.

Since we wrote $\mathscr{T}_k^{-1}$ we have to check if $\mathscr{T}_k$ is invertible. Otherwise, the model is not well-posed. For the one species tensor, this is done by the following theorem proven in [2].

Theorem 1. *Assume that $f_k > 0$. Then $\frac{\mathbb{P}_k}{n_k}$ has strictly positive eigenvalues. If we further assume that $-\frac{1}{2} \leq \mu_k \leq 1$, then $\mathscr{T}_k$ has strictly positive eigenvalues and therefore $\mathscr{T}_k$ is invertible.*

2.1.1 Equilibrium and Entropy Inequality

In global equilibrium when f_1 and f_2 are independent of x and t, the right- hand side of (13) has to be zero. In this case, we get

$$f_1 = \frac{1}{\nu_{11}n_1 + \nu_{12}n_2}(\nu_{11}n_1G_1 + \nu_{12}n_2M_{12}).$$

If we compute the velocities of this expression, we can deduce $u_1 = u_2$ for $\delta \neq 1$. If we compute the temperatures of this expression using $u_1 = u_2$, we get

$$T_1 = \frac{1}{\nu_{11}n_1 + \nu_{12}n_2}(\nu_{11}n_1T_1 + \nu_{12}n_2(\alpha T_1 + (1-\alpha)T_2)),$$

which is equivalent to $T_1 = T_2$ for $\alpha \neq 1$. So let $T := T_1 = T_2$ and use $u_1 = u_2$. If we compute pressure tensors, we get

$$\begin{aligned}(\nu_{11}n_1 + \nu_{12}n_2)\mathbb{P}_1 &= \nu_{11}n_1\mathscr{T}_1 + \nu_{12}n_2T_{12}\\ &= \nu_{11}n_1(1-\mu_1)T\mathbf{1} + \nu_{11}n_1\mu_1\mathbb{P}_1 + \nu_{12}n_2T\mathbf{1},\end{aligned}$$

which is equivalent to

$$(\nu_{11}n_1 + \nu_{12}n_2 - \nu_{11}n_1\mu_1)\mathbb{P}_1 = (\nu_{11}n_1 + \nu_{12}n_2 - \nu_{11}n_1\mu_1)T\mathbf{1},$$

which is $\mathbb{P}_1 = T\mathbf{1}$ for $\delta, \alpha \neq 1$, $\mu_1 \leq 1$. This means that the pressure tensor of f_1 and f_2 is diagonal and f_1, f_2 are Maxwellian distributions with equal mean velocity and temperature. $\delta = 1$ or $\alpha = 1$ are cases in which the mixture Maxwellians do not contain the velocity or the temperature of the other species, see (7) and (8). In this case, the two gases do not exchange information and a global equilibrium cannot be reached.

Theorem 2 (H-theorem for the mixture). *Assume that f_1, $f_2 > 0$ are solutions to (2). Assume the relationship between the collision frequencies (5), the conditions for the interspecies Maxwellians (7), (9), (8), and (10) and the positivity of the temperatures (11), then*

$$\int (\ln f_1)\, Q_{11}(f_1, f_1) + (\ln f_1)\, Q_{12}(f_1, f_2)dv + \int (\ln f_2)\, Q_{22}(f_2, f_2) + (\ln f_2)\, Q_{21}(f_2, f_1)dv \leq 0,$$

with equality if and only if f_1 and f_2 are Maxwell distributions with equal velocity and temperature.

Proof. The fact that $\int \ln f_k Q_{kk}(f_k, f_k)dv \leq 0, \quad k = 1, 2$ with a criteria for equality follows from the H-Theorem of the ES-BGK model for one species, see [2]. The fact

that $\int \ln f_1 Q_{12}(f_1, f_2)dv + \int \ln f_2 Q_{21}(f_1, f_2)dv \leq 0$ with a corresponding criteria for equality follows from the H-Theorem of the BGK model for mixtures, see Theorem 2.7 in [19].

2.2 Alternative Extensions to an ES-BGK Model

In this subsection, we also want to replace the scalar temperatures in the mixture Maxwellians by a tensor. In the first model, the terms $(v_j - u_{kj}) f_k(v_i - u_{ki})$ for $i \neq j$ do not appear in the relaxation operator. To obtain a more detailed description of the viscous effects in the mixture, we take into account these cross terms during the relaxation process. Then the model can be written as:

$$\partial_t f_k + \nabla_x \cdot (v f_k) = \nu_{kk} n_k (G_k - f_k) + \nu_{kj} n_j (G_{kj} - f_k), \quad k = 1, 2, k \neq j, \tag{16}$$

with the modified Maxwell distributions

$$\begin{aligned} G_k(x, v, t) &= \frac{n_k}{\sqrt{\det(2\pi \frac{\mathscr{T}_k}{m_k})}} \exp(-\frac{1}{2}(v - u_k) \cdot (\frac{\mathscr{T}_k}{m_k})^{-1} \cdot (v - u_k)) \quad k = 1, 2, \\ G_{kj}(x, v, t) &= \frac{n_k}{\sqrt{\det(2\pi \frac{\mathscr{T}_{kj}}{m_k})}} \exp(-\frac{1}{2}(v - u_{kj}) \cdot (\frac{\mathscr{T}_{kj}}{m_k})^{-1} \cdot (v - u_{kj})) \quad k = 1, 2, k \neq j. \end{aligned} \tag{17}$$

Again, the conservation of mass, momentum, and energy in interactions of one species with itself is ensured by this choice of the modified Maxwell distributions G_1 and G_2 which have the same densities, velocities, and pressure tensor as f_1 and f_2, respectively. In addition, the choice of the densities in G_{12} and G_{21}, we also guarantee conservation of mass in interactions of one species with the other one.

If we extend T_{12} and T_{21} in the same fashion to a tensor as in the case of one species, we obtain

$$\mathscr{T}_{12} = (1 - \mu_{12})(\alpha T_1 + (1 - \alpha) T_2)\mathbf{1} + \mu_{12} \frac{\alpha \mathbb{P}_1 + (1 - \alpha)\mathbb{P}_2}{n_1} + \gamma |u_1 - u_2|^2 \mathbf{1}, \tag{18}$$

$$\begin{aligned} \mathscr{T}_{21} &= (1 - \mu_{21})((1 - \varepsilon(1 - \alpha)) T_2 + \varepsilon(1 - \alpha) T_1)\mathbf{1} \\ &\quad + \mu_{21} \frac{(1 - \varepsilon(1 - \alpha))\mathbb{P}_2 + \varepsilon(1 - \alpha)\mathbb{P}_1}{n_2} + (\frac{1}{3}\varepsilon m_1 (1 - \delta)(\frac{m_1}{m_2}\varepsilon(\delta - 1) + \delta + 1) - \varepsilon\gamma)|u_1 - u_2|^2 \mathbf{1}. \end{aligned} \tag{19}$$

If we check the equilibrium distributions as in Sect. 2.1.1, we obtain the following restrictions on μ_{12} and μ_{21} given by

$$\mu_{12} = 1 + (1 - \mu_1)\frac{n_1}{n_2}\frac{\nu_{11}}{\nu_{12}}, \tag{20}$$

and

$$\begin{aligned}\frac{1}{n_1^2}[-(\alpha-1)^2\mu_{12}^2 n_2^2 \nu_{12}^2 + \frac{n_1}{n_2^2}((\frac{\mu_{21}}{\varepsilon} - \mu_{21} + \alpha\mu_{21})n_1\nu_{12} + (\mu_2 - 1)n_2\nu_{22}) \\ \cdot(n_1((\alpha-1)\mu_{21}n_1 + \frac{1}{\varepsilon}(\mu_{21}-1)n_2)\nu_{12} + (\mu_2-1)n_2^2\nu_{22})] = 0,\end{aligned} \tag{21}$$

An alternative choice to (18), (19), which is less complicated, is given by

$$\mathscr{T}_{12} = \alpha\frac{\mathbb{P}_1}{n_1} + (1-\alpha)T_2\mathbf{1} + \gamma|u_1 - u_2|^2\mathbf{1}, \tag{22}$$

$$\begin{aligned}\mathscr{T}_{21} = (1 - \varepsilon(1-\alpha))\frac{\mathbb{P}_2}{n_2} + \varepsilon(1-\alpha)T_1\mathbf{1} \\ + (\frac{1}{3}\varepsilon m_1(1-\delta)(\frac{m_1}{m_2}\varepsilon(\delta-1) + \delta + 1) - \varepsilon\gamma)|u_1 - u_2|^2\mathbf{1}.\end{aligned} \tag{23}$$

This choice still contains the temperature of gas 1, since the trace of the pressure tensor is the temperature.

In (22) compared to (18), we replace only the temperature T_1 of species 1 by the pressure tensor $\mathbb{P}_1$ while we keep the temperature T_2. This asymmetric choice can be motivated by the theory of "persistence of velocity" described by Jeans in [17, 18]. He argues that in the post-collisional speed of particle 1 there is a memory of the pre-collisional speed of particle 1. In the single species, BGK equation this yields to the choice of

$$\mathscr{T} = (1-\mu)T\mathbf{1} + \mu\mathbb{P}, \quad -\frac{1}{2} \le \mu \le 1,$$

the tensor chosen in the well-known ES-BGK model, where $\mu\mathbb{P}$ preserves the memory of the off-equilibrium content of the pre-collisional velocity. This can be rewritten as

$$\mathscr{T} = T\mathbf{1} + \mu\text{traceless}[\mathbb{P}],$$

where traceless$[\mathbb{P}]$ denotes the traceless part of $\mathbb{P}$. So the off-equilibrium part is contained in μtraceless$[\mathbb{P}]$. Doing this analogously for two species, we arrive at

$$\mathscr{T}_{12} = T_{12}\mathbf{1} + \frac{\alpha}{n_1}\text{traceless}[\mathbb{P}_1].$$

If we plug in the definition of T_{12} given by (8), we end up with (22).

With the second choice, the model is well-defined, because $\mathscr{T}_{12}$ and $\mathscr{T}_{21}$ are invertible as a combination of strictly positive matrices as soon as all coefficients in front of these matrices are positive, which is the case due to (11) and (12). The first choice needs additional conditions coming from the restrictions on μ_{12} and μ_{21}

given by (20) and (21). The first one leads to

$$\mu_1 \leq \frac{n_2}{n_1}\frac{\nu_{12}}{\nu_{11}} + 1,$$

such that μ_{12} given by (20) is positive. The requirement of positivity of μ_{21} leads to a corresponding restriction on μ_2 using (21).

2.2.1 Equilibrium and Entropy Inequality

The aim of this subsection is to discuss the property of equilibrium and the entropy inequality for the alternative extensions described in Sect. 2.2 with the tensors (18), (19) respective (22), (23). For the tensors (18), (19), we proved the property of equilibrium and the H-Theorem in Sect. 2.1.1 in the particular case for $\mu_{12} = \mu_{21} = 0$ for simplicity, but we can also prove it in the general case. In this section, we will prove an entropy inequality for the alternative model (22), (23). First we will check that the equilibrium distributions are Maxwellians. In global equilibrium, when f_1 and f_2 are independent of x and t, the right-hand side of (16) has to be zero. In this case, we get

$$f_1 = \frac{1}{1 + \frac{1}{\beta_1^2}\frac{n_2}{n_1}}(G_1 + \frac{1}{\beta_1^2}\frac{n_2}{n_1}G_{12}).$$

If we compute the temperatures of this expression, we get

$$T_1 = \frac{1}{1 + \frac{1}{\beta_1^2}\frac{n_2}{n_1}}(T_1 + \frac{1}{\beta_1^2}\frac{n_2}{n_1}(\alpha T_1 + (1-\alpha)T_2)),$$

which is equivalent to $T_1 = T_2$ for $\alpha \neq 1$. So denote $T := T_1 = T_2$. If we compute pressure tensors, we get

$$\begin{aligned}(1 + \frac{1}{\beta_1^2}\frac{n_2}{n_1})\mathbb{P}_1 &= \mathscr{T}_1 + \frac{1}{\beta_1^2}\frac{n_2}{n_1}\mathscr{T}_{12} \\ &= (1-\nu_1)T + \nu_1\mathbb{P}_1 + \frac{1}{\beta_1^2}\frac{n_2}{n_1}\alpha\mathbb{P}_1 + \frac{1}{\beta_1^2}\frac{n_2}{n_1}(1-\alpha)T\mathbf{1}\end{aligned}$$

which is equivalent to

$$((1-\nu_1) + \frac{1}{\beta_1^2}\frac{n_2}{n_1}(1-\alpha))\mathbb{P}_1 = ((1-\nu_1) + \frac{1}{\beta_1^2}\frac{n_2}{n_1}(1-\alpha))T\mathbf{1},$$

which is $\mathbb{P}_1 = T\mathbf{1}$ for $\nu_1, \alpha \neq 1$. That means that the pressure tensors of f_1 and f_2 are diagonal and they are Maxwellian distributions with equal mean velocity and temperature.

Next, we want to prove the H-Theorem of the simpler model (22) and (23). For this proof, we need the following lemmas.

Lemma 1. (Brunn–Minkowski inequality). *Let* $0 \leq a \leq 1$ *and* A, B *positive symmetric matrices, then*

$$\det(aA + (1-a)B) \geq (\det A)^a (\det B)^{1-a}.$$

Proof. The proof is given in [2].

Lemma 2. *Assuming* (22) *and* (23) *and the positivity of all temperatures and pressure tensors* (11)*, we have the following inequality*

$$S := (\det \mathscr{T}_{12})^{\varepsilon} (\det \mathscr{T}_{21}) \geq (\det \frac{\mathbb{P}_1}{n_1})^{\varepsilon} \det \frac{\mathbb{P}_2}{n_2}.$$

Proof. Using the definition of $\mathscr{T}_{12}$, we get

$$\det \mathscr{T}_{12} = \det(\alpha \frac{\mathbb{P}_1}{n_1} + (1-\alpha) T_2 \mathbf{1} + \gamma |u_1 - u_2|^2 \mathbf{1}).$$

Since γ is non-negative, we can estimate the expression by dropping the positive term on the diagonal $\gamma |u_1 - u_2|^2 \mathbf{1}$

$$\det \mathscr{T}_{12} \geq \det(\alpha \frac{\mathbb{P}_1}{n_1} + (1-\alpha) T_2 \mathbf{1}).$$

With the Brunn–Minkowski inequality, we obtain

$$\det \mathscr{T}_{12} \geq (\det \frac{\mathbb{P}_1}{n_1})^{\alpha} (\det T_2 \mathbf{1})^{1-\alpha}.$$

In a similar way, we can show it for $\mathscr{T}_{21}$, so all in all we get

$$S \geq (\det \frac{\mathbb{P}_1}{n_1})^{\alpha\varepsilon} (\det T_2 \mathbf{1})^{\varepsilon(1-\alpha)} (\det \frac{\mathbb{P}_2}{n_2})^{1-\varepsilon(1-\alpha)} (\det T_1 \mathbf{1})^{\varepsilon(1-\alpha)}.$$

Consider the logarithm of this equation

$$\begin{aligned} \ln S \geq \varepsilon\alpha \ln \left(\det \left(\frac{\mathbb{P}_1}{n_1} \right) \right) + \varepsilon(1-\alpha) \ln (\det (T_2 \mathbf{1})) \\ + (1 - \varepsilon(1-\alpha)) \ln \left(\det \left(\frac{\mathbb{P}_2}{n_2} \right) \right) + \varepsilon(1-\alpha) \ln (\det (T_1 \mathbf{1})) . \end{aligned}$$

We use that $\ln (\det (T_i \mathbf{1})) = \mathrm{Tr}(\ln (T_i \mathbf{1}))$, $T_i = \mathrm{Tr} \frac{\mathbb{P}_i}{3n_i}$ and denote the eigenvalues of $\frac{\mathbb{P}_i}{n_i}$ by $\lambda_{i,1}$, $\lambda_{i,2}$ and $\lambda_{i,3}$. Since the pressure tensors are symmetric, we can diagonalize

them and use that $T_i = \mathrm{Tr}\frac{\mathbb{P}}{3n_i} = 1/3(\lambda_{i,1} + \lambda_{i,2} + \lambda_{i,3})$.

$$\ln S \geq \varepsilon\alpha(\ln\lambda_{1,1} + \ln\lambda_{1,2} + \ln\lambda_{1,3}) + \varepsilon(1-\alpha)3\ln\frac{1}{3}(\lambda_{1,1} + \lambda_{1,2} + \lambda_{1,3})$$
$$+(1-\varepsilon(1-\alpha))(\ln\lambda_{2,1} + \ln\lambda_{2,2} + \ln\lambda_{2,3}) + \varepsilon(1-\alpha)3\ln\frac{1}{3}(\lambda_{2,1} + \lambda_{2,2} + \lambda_{2,3}).$$

Since *ln* is concave, we can estimate $\ln\frac{1}{3}(\lambda_{1,1} + \lambda_{1,2} + \lambda_{1,3})$ from below by $\frac{1}{3}(\ln\lambda_{1,1} + \ln\lambda_{1,2} + \ln\lambda_{1,3})$ and obtain

$$\ln S \geq \varepsilon\ln\left(\det\left(\frac{\mathbb{P}_1}{n_1}\right)\right) + \varepsilon(1-\alpha)\ln\left(\det\left(\frac{\mathbb{P}_2}{n_2}\right)\right).$$

This is equivalent to the required inequality.

Remark 1. From the case of one species ES-BGK model, we know that

$$\int G_k \ln G_k dv \leq \int G_{k,\mu_k=1}\ln G_{k,\mu_k=1}dv \leq \int f_k \ln f_k dv,$$

for $k = 1, 2$, see [2], where $G_{k,\mu_k=1}$ denotes the modified Maxwellian where $\mu_k = 1$ in the tensor (15).

Theorem 3. (H-theorem for mixture). *Assume* $\alpha, \delta \neq 1$. *Assume* $f_1, f_2 > 0$. *Assume the relationship between the collision frequencies* (5)*, the conditions for the inter-species Maxwellians* (7)*,* (9)*,* (22) *and* (23) *and the positivity of the temperatures* (11)*, then*

$$\int(\ln f_1)\,Q_{11}(f_1, f_1) + (\ln f_1)\,Q_{12}(f_1, f_2)dv + \int(\ln f_2)\,Q_{22}(f_2, f_2) + (\ln f_2)\,Q_{21}(f_2, f_1)dv \leq 0,$$

with equality if and only if f_1 *and* f_2 *are Maxwell distributions with equal mean velocity and temperature.*

Proof. The fact that $\int \ln f_k Q_{kk}(f_k, f_k)dv \leq 0$, $k = 1, 2$ is shown in proofs of the H-theorem of the single ES-BGK-model, for example, in [2]. In both cases, we have equality if and only if $f_1 = M_1$ and $f_2 = M_2$.
Let us define

$$S(f_1, f_2) := \nu_{12}n_2\int \ln f_1(G_{12} - f_1)dv + \nu_{21}n_1\int \ln f_2(G_{21} - f_2)dv.$$

The task is to prove that $S(f_1, f_2) \leq 0$. Since the function $H(x) = x\ln x - x$ is strictly convex for $x > 0$, we have $H'(f)(g - f) \leq H(g) - H(f)$ with equality if and only if $g = f$. So

$$(g-f)\ln f \le g\ln g - f\ln f + f - g. \tag{24}$$

Consider now $S(f_1, f_2)$ and apply the inequality (24) to each of the two terms in S.

$$S(f_1,f_2) \le \nu_{12}n_2\left[\int G_{12}\ln G_{12}dv - \int f_1\ln f_1 dv - \int G_{12}dv + \int f_1 dv\right]$$

$$+\nu_{21}n_1\left[\int G_{21}\ln G_{21}dv - \int f_2\ln f_2 dv - \int G_{21}dv + \int f_2 dv\right],$$

with equality if and only if $f_1 = G_{12}$ and $f_2 = G_{21}$. If we compute the velocities of $f_1 = G_{12}$ and $f_2 = G_{21}$, we can deduce $u_1 = u_{12}$ and $u_2 = u_{21}$ which lead to $u_1 = u_2$ using the definitions of u_{12}, u_{21} given by (7) and (9). Analogously, computing the temperatures, we get $T_{12} = T_{21} = T_1 = T_2 =: T$. Finally, computing the pressure tensors, we obtain $\frac{\mathbb{P}_1}{n_1} = \frac{\mathbb{P}_2}{n_2} = T\mathbf{1}$, which means that we have equality if and only if f_1 and f_2 are Maxwellians with equal temperatures and velocities.

Since G_{12} and f_1 have the same density and G_{21} and f_2 have the same density too, the right-hand side reduces to

$$\nu_{12}n_2\Big(\int G_{12}\ln G_{12}dv - \int f_1\ln f_1 dv\Big) + \nu_{21}n_1\Big(\int G_{21}\ln G_{21}dv - \int f_2\ln f_2 dv\Big).$$

Since $\int G\ln G dv = n\ln\Big(\frac{n}{\sqrt{\det(\frac{2\pi\mathscr{T}}{m})}}\Big) - \frac{3}{2}n$ for $G = \frac{n}{\sqrt{\det(\frac{2\pi\mathscr{T}}{m})^3}}e^{-(v-u)\cdot(\frac{\mathscr{T}}{m})^{-1}\cdot(v-u)}$, we will have that

$$\nu_{12}n_2\int G_{12}\ln G_{12}dv + \nu_{21}n_1\int G_{21}\ln G_{21}dv$$

$$\le \nu_{21}n_1\int G_{2,\mu_2=1}\ln M_{2,\mu_2=1}dv + \nu_{12}n_2\int G_{1,\mu_1=1}\ln G_{1,\mu_1=1}dv,$$

provided that

$$\nu_{12}n_2n_1\ln\frac{n_1}{\sqrt{\det(2\pi\frac{\mathscr{T}_{12}}{m_1})}} + \nu_{21}n_2n_1\ln\frac{n_2}{\sqrt{\det(2\pi\frac{\mathscr{T}_{21})}{m_2}}}$$

$$\le \nu_{12}n_2n_1\ln\frac{n_1}{\sqrt{\det(2\pi\frac{\mathbb{P}_1}{m_1})}} + \nu_{21}n_2n_1\ln\frac{n_2}{\sqrt{\det(2\pi\frac{\mathbb{P}_2}{m_2})}},$$

which is equivalent to the condition

$$(\det\mathscr{T}_{12})^{\varepsilon}(\det\mathscr{T}_{21}) \ge (\det\frac{\mathbb{P}_1}{n_1})^{\varepsilon}\det\frac{\mathbb{P}_2}{n_2},$$

proven in Lemma 2.

With this inequality, we get

$$S(f_1, f_2) \leq \nu_{12} n_2 [\int G_{1,\mu_1=1} \ln G_{1,\mu_1=1} dv - \int f_1 \ln f_1 dv] + \nu_{21} n_1 [G_{2,\mu_2=1} \ln G_{2,\mu_2=1} dv - \int f_2 \ln f_2 dv] \leq 0.$$

The last inequality follows from Remark 1. Here, we also have equality if and only if $f_1 = M_1$ and $f_2 = M_2$, but since we already noticed that equality also implies $f_1 = G_{12}$ and $f_2 = G_{21}$.

Define the total entropy $H(f_1, f_2) = \int (f_1 \ln f_1 + f_2 \ln f_2) dv$. We can compute

$$\partial_t H(f_1, f_2) + \nabla_x \cdot \int (f_1 \ln f_1 + f_2 \ln f_2) v dv = S(f_1, f_2),$$

by multiplying the BGK equation for the species 1 by $\ln f_1$, the BGK equation for the species 2 by $\ln f_2$ and integrating the sum with respect to v.

Corollary 1. (Entropy inequality for mixtures). *Assume $f_1, f_2 > 0$. Assume a fast enough decay of f to zero for $v \to \infty$. Assume relationship (5), the conditions (7), (9), (22) and (23) and the positivity of the temperatures (11), then we have the following entropy inequality*

$$\partial_t \left(\int f_1 \ln f_1 dv + \int f_2 \ln f_2 dv \right) + \nabla_x \cdot \left(\int v f_1 \ln f_1 dv + \int v f_2 \ln f_2 dv \right) \leq 0,$$

with equality if and only if f_1 and f_2 are Maxwell distributions with equal bulk velocity and temperature.

In summary, the ES-BGK models (13), (16) have five free parameters. We expect this will aid in determining macroscopic physical constants, analogously to how it is done in [14].

References

1. P. Asinari, Asymptotic analysis of multiple-relaxation-time lattice Boltzmann schemes for mixture modeling. Comput. Math. Appl. **55**, 1392–1407 (2008)
2. P. Andries, B. Perthame, *The ES-BGK Model Equation With Correct Prandtl Number, AIP Conference Proceedings*, vol. 30 (2001)
3. P. Andries, P. Le Tallec, J. Perlat, B. Perthame, The Gaussian -BGK model of Boltzmann equation with small Prandtl number. Eur. J. Mech. B - Fluids **19**, 813–830 (2000)
4. P. Andries, K. Aoki, B. Perthame, A consistent BGK-type model for gas mixtures. J. Stat. Phys. **106**, 993–1018 (2002)
5. S. Brull, An ellipsoidal statistical model for gas mixtures. Commun. Math. Sci. **8**, 1–13 (2015)

6. M. Bennoune, M. Lemou, L. Mieussens, Uniformly stable numerical schemes for the Boltzmann equation preserving the compressible Navier-Stokes asymptotics. J. Comput. Phys. **227**, 3781–3803 (2008)
7. F. Bernard, A. Iollo, G. Puppo, Accurate asymptotic preserving boundary conditions for kinetic equations on Cartesian grids. J. Sci. Comput. **65**, 735–766 (2015)
8. C. Cercignani, *The Boltzmann Equation and its Applications* (Springer, Berlin, 1975)
9. C. Cercignani, *Rarefied Gas Dynamics, From Basic Concepts to Actual Calculations* (Cambridge University Press, Cambridge, 2000)
10. A. Crestetto, N. Crouseilles, M. Lemou, Kinetic/fluid micro-macro numerical schemes for Vlasov-Poisson-BGK equation using particles. Kinet. Relat. Models **5**, 787–816 (2012)
11. G. Dimarco, L. Pareschi, Numerical methods for kinetic equations. Acta Numer. **23**, 369–520 (2014)
12. F. Filbet, S. Jin, A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. J. Comput. Phys. **20**, 7625–7648 (2010)
13. E.P. Gross, M. Krook, Model for collision processes in gases: small-amplitude oscillations of charged two-component systems. Phys. Rev. **3**, 593 (1956)
14. M. Groppi, S. Monica, G. Spiga, A kinetic ellipsoidal BGK model for a binary gas mixture. EPL: Eur. Lett. **96**, 64002 (2011)
15. B. Hamel, Kinetic model for binary gas mixtures. Phys. Fluids **8**, 418–425 (1965)
16. L. Holway, New statistical models for kinetic theory: methods of construction. Phys. Fluids **9**, 1658–1673 (1966)
17. J.H. Jeans, The persistence of molecular velocities in the kinetic theory of gases. Philos. Mag. 6 **8**(48), 700–703 (1904)
18. J.H. Jeans, *The Dynamical Theory of Gases* (Cambridge University Press, Cambridge, 1916)
19. C. Klingenberg, M. Pirner, G. Puppo, A consistent kinetic model for a two-component mixture with an application to plasma. Kinet. Relat. Models **10**, 444–465 (2017)
20. S. Pieraccini, G. Puppo, Implicit-explicit schemes for BGK kinetic equations. J. Sci. Comput. **32**, 1–28 (2007)
21. S.-B. Yun, Classical solutions for the ellipsoidal BGK model with fixed collision frequency. J. Differ. Equ. **259**, P6009–6037 (2015)

An Arbitrary Lagrangian–Eulerian Discontinuous Galerkin Method for Conservation Laws: Entropy Stability

Christian Klingenberg, Gero Schnücke and Yinhua Xia

Abstract In Klingenberg, Schnücke and Xia (Math. Comp. Available via https://doi.org/10.1090/mcom/3126) an arbitrary Lagrangian–Eulerian Discontinuous Galerkin (ALE-DG) method to solve conservation laws has been developed and analyzed. In this paper, the ALE-DG method will be briefly presented. Furthermore, the semi-discrete method will be discretized by the so-called ϑ-method. The ϑ-method is a generalization of the forward or backward Euler step. In particular, the method degenerates to the forward Euler step for $\vartheta = 0$ and to the backward Euler step for $\vartheta = 1$. The corresponding fully discrete ϑ-P^k-ALE-DG method for scalar conservation laws will be analyzed with respect to entropy stability, where P^k denotes the space of polynomials of degree k which is used on a reference cell. The main results are a cell entropy inequality for the fully discrete ϑ-P^k-ALE-DG method with respect to the square entropy function, when ϑ has a lower bound given by a mesh parameter depending constant, and a cell entropy inequality for the fully discrete ϑ-P^0-ALE-DG method with respect to the Kružkov entropy functions.

Keywords Arbitrary Lagrangian–Eulerian discontinuous Galerkin method Conservation laws · Entropy stability

C. Klingenberg (✉) · G. Schnücke
University of Würzburg, Emil-Fischer-Str. 40, 97074 Würzburg, Germany
e-mail: klingen@mathematik.uni-wuerzburg.de

G. Schnücke
e-mail: gero.schnuecke@mathematik.web.de

Y. Xia
University of Science and Technology of China Hefei,
Anhui 230026, People's Republic of China
e-mail: yhxia@ustc.edu.cn

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_16

1 Introduction

The present paper investigates an arbitrary Lagrangian–Eulerian discontinuous Galerkin (ALE-DG) method to solve one-dimensional conservation laws

$$\partial_t u + \partial_x f(u) = 0 \; in \; \Omega \times (0,T), \qquad u(x,0) = u_0(x) \; in \; \Omega \tag{1}$$

with periodic boundary conditions. The function $u_0 : \Omega \to \mathbb{R}$ is sufficiently smooth and compactly supported, and the flux function $f : \mathbb{R} \to \mathbb{R}$ is at least one times continuously differentiable. This method was introduced and analyzed by Klingenberg et al. in [9].

The Arbitrary Lagrangian–Eulerian (ALE) approach has been rigorously described by Donea et al. in [5]. It is a kind of compromise between the Lagrangian and Eulerian approach. These two approaches are the two commonly used descriptions of motions in computational fluid dynamics. In the Lagrangian approach the mesh points are moving with the fluid velocity. This approach could produce distortions in the mesh. The distortions lead to numerical artifacts. This has been discussed by Donea et al. in [5]. In the Eulerian approach the mesh is static. Hence, numerical artifacts by geometric distortions are avoided in this approach. Nevertheless, a drawback of the Eulerian approach is the loss of specific properties of the physical model. Springel [11] compared the Lagrangian and Eulerian approach in cosmological hydrodynamical simulations using the finite volume method and observed a lack of the Galilean invariance when the Eulerian approach is used. Furthermore, in the same paper, Springel showed by numerical simulations with a second-order finite volume moving mesh method that the Galilei invariance is preserved when the mesh moves almost with the fluid.

The main idea of the ALE approach is to described the fluid motion almost as in the Lagrangian approach, and if distortions with a destabilizing effect occur, the description of motion moves closer to the Eulerian approach. The implementation and mathematical description of the ALE approach ensure by a mapping which connects the physical domain with a suitable reference configuration. The mapping provides a description of the grid velocity field. In addition, the test function space is defined by the mapping, in the context of Galerkin methods. In general, the mapping is globally defined. This is quite unattractive for discontinuous Galerkin methods, since these methods lose their local structure, when a global defined ALE mapping is used. Furthermore, if the ALE approach is combined with numerical schemes, which are derived by the method of lines approach, and the Jacobi matrix of the mapping depends on spatial variables, a geometric error could appear by an unsuitable choice of the time integration method. This geometric error destabilizes the numerical scheme. The geometric error does not appear, if the ALE method satisfies the geometric conservation law (GCL). The error and the GCL have been analyzed by Guillard and Farhat in [7].

The ALE-DG method in [9] is derived by local affine linear ALE mappings. Hence, the method has a local structure like the DG methods for static grids, and it

has been proven that the method to solve one-dimensional conservation laws satisfies the GCL for any first-order time discretization method or high-order single-step method in which the stage order is equal or higher than first order. Moreover, for the semi-discrete method, a cell entropy inequality with respect to the square entropy function and a priori error estimates have been proven. For the time integration, the total variation diminishing (TVD) Runge–Kutta methods, which were introduced by Shu in [10], are adopt. Hence, the ALE-DG method degenerates to the Runge–Kutta discontinuous Galerkin (RK-DG) method on a static non-moving mesh. The RK-DG method was developed by Cockburn, Shu et al. in a series of papers [2–4] and is designed for the Eulerian description of fluid motion. Over the last decades, this method has become quite popular in computational fluid dynamics. The TVD Runge–Kutta methods are convex combinations of the forward Euler step. Hence, a stability result for the forward Euler step could be extent by an adequate time step regulation. This feature of the TVD Runge–Kutta methods has been proven by Gottlieb and Shu in [6]. According to this property of the TVD Runge–Kutta methods, it has been proven that the full discrete ALE-DG method satisfies a local maximum principle and the average values of the ALE-DG solution are total variation stable.

The next step is the analysis of the fully discrete ALE-DG method with respect to entropy stability. Unfortunately, Chavent and Cockburn proved in [1] that the P^1-DG method to solve scalar conservation laws with a linear flux function on static grids is unconditionally $\mathrm{L}^\infty\left(0, T; \mathrm{L}^2\,(0, 1)\right)$-unstable for any CFL restriction, when the forward Euler step is used. Hence, we cannot expect entropy stability for the forward Euler P^k-ALE-DG method, if $k \geq 1$. In particular, the entropy stability for the P^k-ALE-DG method with a TVD Runge–Kutta cannot be proven by Gottlieb and Shu's theorem and needs to be investigated separated from the forward Euler P^k-ALE-DG method.

Jiang and Shu analyzed in [8] fully discrete DG methods with respect to entropy stability. They applied the ϑ-method for the time integration of the semi-discrete DG method and proved for $\frac{1}{2} \leq \vartheta \leq 1$ and polynomials of arbitrary degree a cell entropy inequality with respect to the square entropy function. The ϑ-method for the ordinary differential equation $\partial_t u = \mathscr{L}\,(u, t)$ is given by

$$u^{n+1} = u^n + \Delta t \mathscr{L}\left(u^{n+\vartheta}, t_{n+\vartheta}\right), \tag{2a}$$

$$u^{n+\vartheta} := (1-\vartheta)\,u^n + \vartheta u^{n+1}, \quad t_{n+\vartheta} := (1-\vartheta)\,t_n + \vartheta t_{n+1}. \tag{2b}$$

In this paper, the ϑ-method is applied for the time integration of the semi-discrete ALE-DG method and the corresponding ϑ-P^k-ALE-DG method is analyzed with respect to entropy stability in the sense of the square entropy and the Kružkov entropy functions.

This paper is organized as follows: It starts with a briefly presentation of the ALE-DG method in Sect. 2. Afterward, in the same section, two entropy inequalities are proven for the fully discrete method. It will be completed with some concluding remarks in Sect. 3.

2 The ALE-DG Method

This section is started with a summary of the main ingredients to describe the ALE-DG method. Let $\Omega \subseteq \mathbb{R}$ be an open interval. It need to be assumed that it exists for any time level $t = t_n$ a partition of the domain Ω with

$$\overline{\Omega} = \bigcup_{j=1}^{N} \overline{K_j^n}, \quad K_j^n := \left(x_{j-\frac{1}{2}}^n, x_{j+\frac{1}{2}}^n \right), \quad \Delta_j^n := x_{j+\frac{1}{2}}^n - x_{j-\frac{1}{2}}^n.$$

This assumption enables to define time-dependent straight lines for all $j = 1, ..., N$

$$x_{j-\frac{1}{2}}(t) := x_{j-\frac{1}{2}}^n + \omega_{j-\frac{1}{2}}^n (t - t_n), \quad \omega_{j-\frac{1}{2}}^n := \frac{1}{\Delta t} \left(x_{j-\frac{1}{2}}^{n+1} - x_{j-\frac{1}{2}}^n \right),$$

where Δt is specified by the partition of the time interval $(0, T)$. The straight lines provide for any $t \in \left[t_n, t_{n+1} \right]$ and all $j = 1, ..., N$ time-dependent cells

$$K_j(t) := \left(x_{j-\frac{1}{2}}(t), x_{j+\frac{1}{2}}(t) \right), \quad \Delta_j(t) := x_{j+\frac{1}{2}}(t) - x_{j-\frac{1}{2}}(t).$$

The local grid velocity of the ALE-DG method is for all $t \in \left[t_n, t_{n+1} \right)$ and $x \in K_j(t)$ given by

$$\omega(x, t) = \frac{1}{\Delta_j(t)} \left(\omega_{j+\frac{1}{2}}^n - \omega_{j-\frac{1}{2}}^n \right) \left(x - x_{j-\frac{1}{2}}(t) \right) + \omega_{j-\frac{1}{2}}^n. \tag{3}$$

The time-dependent cells can be connected with a reference cell [0, 1] by an affine linear mapping

$$\chi_j : [0, 1] \to \overline{K_j(t)}, \quad \xi \mapsto \chi_j(\xi, t) := \Delta_j(t) \xi + x_{j-\frac{1}{2}}(t).$$

This mapping enables to define the following time-dependent finite-dimensional test function space

$$\mathcal{V}_h(t) := \left\{ v \in \mathrm{L}^2(\Omega) : \left(v \circ \chi_j \right) \in P^k([0, 1]) \right\},$$

where $P^k([0, 1])$ denotes the space of polynomials in [0, 1] of degree at most k. The test functions $v \in \mathcal{V}_h(t)$ are discontinuous in the points $x_{j-\frac{1}{2}}(t)$. Hence, the limits in these points are defined by

$$v_{j-\frac{1}{2}}^{\pm} := \lim_{\varepsilon \to 0} v \left(x_{j-\frac{1}{2}}(t) \pm \varepsilon, t \right).$$

Finally, it should be mentioned that in [9] for sufficiently smooth functions $u : \Omega \times (0, T) \to \mathbb{R}$ the following ALE transport equation has been proven

$$\frac{d}{dt}\int_{K_j(t)} uv\,dx = \int_{K_j(t)} (\partial_t u)\,v\,dx + \int_{K_j(t)} (\partial_x(\omega u))\,v\,dx, \quad \forall v \in \mathscr{V}_h(t). \tag{4}$$

2.1 The Semi-discrete ALE-DG Discretization

At the beginning, the solution u of the problem (1) is approximated by the function

$$u_h(x,t) = \sum_{\ell=0}^{k} u_\ell^j(t)\,\phi_\ell^j(x,t) \in \mathscr{V}_h(t), \quad \forall t \in \left[t_n, t_{n+1}\right) \text{ and } x \in K_j(t),$$

where $\left\{\phi_0^j(x,t), ..., \phi_k^j(x,t)\right\}$ is a basis of the space $\mathscr{V}_h(t)$ in the cell $K_j(t)$. The coefficients $u_0^j(t), ..., u_k^j(t)$ are the unknowns of the ALE-DG method. In order to determine these coefficients, the Eq. (1) is multiplied by a test function $v \in \mathscr{V}_h(t)$ and the transport equation (4) as well as the integration by parts formula are applied. In general, the function u_h is discontinuous in the cell interface points $x_{j-\frac{1}{2}}(t)$. Hence, in these points, the following Lax–Friedrichs flux is applied

$$\widehat{g}\left(\omega, u^-, u^+\right) := \widehat{g}_+\left(\omega, u^-\right) - \widehat{g}_-\left(\omega, u^+\right), \quad \widehat{g}_\pm(\omega,u) := \frac{1}{2}\left(\lambda_j(t)\,u \pm g(\omega,u)\right)$$

where $g(\omega,u) := f(u) - \omega u$ and

$$\lambda_j(t) := \max\left\{\left|\partial_u g\left(\omega(x,t),u\right)\right| :\ u \in [m,M],\ x \in K_j(t)\right\} \tag{5}$$

with $m := \min_{x\in\Omega} u_0(x)$ and $M := \max_{x\in\Omega} u_0(x)$. Finally, the semi-discrete ALE-DG method can be summarized as follows:

Problem 1 (Semi-discrete ALE-DG method). Seek a function $u_h \in \mathscr{V}_h(t)$, such that for all $v \in \mathscr{V}_h(t)$ and $j = 1, ..., N$ holds

$$\begin{aligned} 0 = \frac{d}{dt}\int_{K_j(t)} u_h v\,dx - \int_{K_j(t)} g(\omega,u_h)\,(\partial_x v)\,dx \\ + \widehat{g}\left(\omega^n_{j+\frac{1}{2}}, u^-_{h,j+\frac{1}{2}}, u^+_{h,j+\frac{1}{2}}\right) v^-_{j+\frac{1}{2}} - \widehat{g}\left(\omega^n_{j-\frac{1}{2}}, u^-_{h,j-\frac{1}{2}}, u^+_{h,j-\frac{1}{2}}\right) v^+_{j-\frac{1}{2}}. \end{aligned} \tag{6}$$

The time discretization method for the problem (6) needs to be chosen carefully, since according to Guillard and Farhat [7] the geometric conservation needs to be respected. However, in [9], it has been proven that the ALE-DG method satisfies the geometric conservation law for any single-step method with stage order equal or higher than first order. Hence, there is a lot of freedom in the choice of a time discretization method for the ALE-DG method.

The capability of the ALE-DG method with a third-order TVD Runge–Kutta method for problems with a compressible flow has been shown by numerical experiments for the inviscid Burgers' equation and Euler equations in [9]. In particular, it has been shown numerically that the method is able to reach the optimal rate of convergence and can handle strong singularities like shock waves.

2.2 Cell Entropy Inequalities

In this section, cell entropy inequalities for the fully discrete ϑ-P^k-ALE-DG method are discussed, where the ϑ-P^k-ALE-DG results from a discretization of the semi-discrete formulation (6) with the ϑ-method (2). The corresponding method can be written on the reference cell $(0, 1)$ as follows:

Problem 2 (The ϑ-P^k-ALE-DG method). For a given function $\widehat{u}_h^n \in \mathscr{V}_h(t_n)$ seek a function $\widehat{u}_h^{n+1} \in \mathscr{V}_h(t_{n+1})$, such that for all $\widehat{v} \in P^k([0,1])$ and $j = 1, ..., N$ holds

$$
\begin{aligned}
0 = &\int_0^1 \Delta_j^{n+1} \widehat{u}_h^{n+1} \widehat{v} \, d\xi - \int_0^1 \Delta_j^n \widehat{u}_h^n v \, d\xi - \int_0^1 g\left(\widehat{\omega}\left(t_{n+\vartheta}\right), \widehat{u}_h^{n+\vartheta}\right)\left(\partial_\xi \widehat{v}\right) d\xi \\
&+ \widehat{g}\left(\omega_{j+\frac{1}{2}}^n, \widehat{u}_{h,j+\frac{1}{2}}^{n+\vartheta,-}, \widehat{u}_{h,j+\frac{1}{2}}^{n+\vartheta,+}\right) \widehat{v}_{j+\frac{1}{2}}^- - \widehat{g}\left(\omega_{j-\frac{1}{2}}^n, \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,+}, \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,+}\right) \widehat{v}_{j-\frac{1}{2}}^+,
\end{aligned}
$$

where $\widehat{u}_h := u_h \circ \chi_j, \widehat{v} := v \circ \chi_j, \widehat{\omega} = \omega \circ \chi_j$ and $t_{n+\vartheta}$ is defined as in (2b).

At first, a cell entropy inequality with respect to the square entropy function $\eta(u) = \frac{1}{2}u^2$ is proven. The proof based on the upcoming discrete variation on the ALE transport equation (4).

Lemma 1. *Let $u : [0,1] \times [0,T] \to \mathbb{R}$ be a sufficiently smooth function and $\eta(u) = \frac{1}{2}u^2$. Then holds*

$$
\begin{aligned}
&\int_0^1 \Delta_j^{n+1} u^{n+1} u^{n+\vartheta} \, d\xi - \int_0^1 \Delta_j^n u^n u^{n+\vartheta} \, d\xi \\
&= \int_0^1 \Delta_j^{n+1} \eta\left(u^{n+1}\right) d\xi - \int_0^1 \Delta_j^n \eta\left(u^n\right) d\xi \\
&\quad + \Delta t \int_0^1 \left(\partial_\xi \widehat{\omega}\left(t_{n+\vartheta}\right)\right) \eta\left(u^{n+\vartheta}\right) d\xi \\
&\quad + \int_0^1 \left[\vartheta^2 \Delta_j^n - (1-\vartheta)^2 \Delta_j^{n+1}\right] \eta\left(u^{n+1} - u^n\right) d\xi, \qquad (7)
\end{aligned}
$$

where $u^{n+\vartheta}$ and $t_{n+\vartheta}$ are defined as in (2b).

Proof. First of all, by a simple algebraic manipulation follows

$$\begin{aligned}
&\int_0^1 \Delta_j^{n+1} u^{n+1} u^{n+\vartheta}\, d\xi - \int_0^1 \Delta_j^n u^n u^{n+\vartheta}\, d\xi \\
&= \int_0^1 \Delta_j^{n+1} \eta\left(u^{n+1}\right)\, d\xi - \int_0^1 \Delta_j^n \eta\left(u^n\right)\, d\xi \\
&\quad + \int_0^1 \left(\Delta_j^{n+1} - \Delta_j^n\right)\left((2\vartheta - 1)\,\eta\left(u^{n+1}\right) + (1-\vartheta)\, u^{n+1} u^n\right)\, d\xi \\
&\quad + (2\vartheta - 1) \int_0^1 \Delta_j^n \eta\left(u^{n+1} - u^n\right)\, d\xi. \qquad (8)
\end{aligned}$$

Next, it should be noted that $\partial_x \omega\left(t_{n+\vartheta}\right) \Delta_j^{n+\vartheta} = \partial_\xi \widehat{\omega}\left(t_{n+\vartheta}\right)$. Hence, by formula (3) follows

$$\Delta t\, \partial_\xi \widehat{\omega}\left(t_{n+\vartheta}\right) = \Delta t \left(\omega^n_{j+\frac{1}{2}} - \omega^n_{j-\frac{1}{2}}\right) = \left(\Delta_j^{n+1} - \Delta_j^n\right). \qquad (9)$$

Moreover, the identity (9) and the integration by substitution formula provide

$$\begin{aligned}
&\int_0^1 \left(\Delta_j^{n+1} - \Delta_j^n\right)\left((2\vartheta - 1)\,\eta\left(u^{n+1}\right) + (1-\vartheta)\, u^{n+1} u^n\right)\, d\xi \\
&\quad - (1-\vartheta)^2 \int_0^1 \left(\Delta_j^{n+1} - \Delta_j^n\right) \eta\left(u^{n+1} - u^n\right)\, d\xi \\
&= \Delta t \int_0^1 \left(\partial_\xi \omega\left(t_{n+\vartheta}\right)\right) \eta\left(u^{n+\vartheta}\right)\, d\xi. \qquad (10)
\end{aligned}$$

Finally, the discrete transport equation (7) follows by combining the equations (8) and (10). □

The discrete transport equation (7) provides only a cell entropy inequality, if it can be ensured that

$$\int_0^1 \left[\vartheta^2 \Delta_j^n - (1-\vartheta)^2\, \Delta_j^{n+1}\right] \eta\left(u^{n+1} - u^n\right)\, d\xi \geq 0. \qquad (11)$$

In fact it follows from a simple calculation that ϑ needs to satisfy

$$0 < \frac{\sqrt{\Delta_j^{n+1}}}{\sqrt{\Delta_j^n} + \sqrt{\Delta_j^{n+1}}} \leq \vartheta \leq 1. \qquad (12)$$

It should be noted that on a static mesh the equation $\Delta_j^n = \Delta_j^{n+1}$ is satisfied. Hence, in this case, (12) yields the restriction $\frac{1}{2} \leq \vartheta \leq 1$. This is the same restriction as in Jiang and Shu's paper [8]. However, the restriction (12) ensures the upcoming entropy inequality with respect to the square entropy function for the ϑ-P^k-ALE-DG method.

Proposition 1. *Suppose ϑ satisfies the restriction (12). Then the solution of the ϑ-P^k-ALE-DG method satisfies with respect to the square entropy function $\eta(u) = \frac{1}{2}u^2$ the cell entropy inequality*

$$0 \geq \int_{K_j^{n+1}} \eta\left(u_h^{n+1}\right) dx - \int_{K_j^n} \eta\left(u_h^n\right) dx + \Delta t \left(H\left(\omega_{j+\frac{1}{2}}^n, u_{h,j+\frac{1}{2}}^{n+\vartheta,-}, u_{h,j+\frac{1}{2}}^{n+\vartheta,+}\right) - H\left(\omega_{j-\frac{1}{2}}^n, u_{h,j-\frac{1}{2}}^{n+\vartheta,-}, u_{h,j-\frac{1}{2}}^{n+\vartheta,+}\right)\right),$$

where $H\left(\omega, u^-, u^+\right) := -\int^{u^-} f(u)\, du + \omega\eta\left(u^-\right) + \widehat{g}\left(\omega, u^-, u^+\right) u^-$. *Furthermore, holds* $\left\|u_h^n\right\|_{\mathrm{L}^2(\Omega)} \leq \left\|u_h^0\right\|_{\mathrm{L}^2(\Omega)}$.

Proof. The ϑ-P^k-ALE-DG can be written as follows

$$\begin{aligned}
0 = &\int_0^1 \Delta_j^{n+1} \eta\left(\widehat{u}_h^{n+1}\right) d\xi - \int_0^1 \Delta_j^n \eta\left(\widehat{u}_h^n\right) d\xi \\
&+ \int_0^1 \left[\vartheta^2 \Delta_j^n - (1-\vartheta)^2 \Delta_j^{n+1}\right] \eta\left(\widehat{u}_h^{n+1} - \widehat{u}_h^n\right) d\xi \\
&- \Delta t \int_0^1 f\left(\widehat{u}_h^{n+\vartheta}\right)\left(\partial_\xi \widehat{u}_h^{n+\vartheta}\right) d\xi + \Delta t \int_0^1 \partial_\xi \left(\widehat{\omega}\left(t_{n+\vartheta}\right) \eta\left(\widehat{u}_h^{n+\vartheta}\right)\right) d\xi \\
&+ \Delta t \left(\widehat{g}\left(\omega_{j+\frac{1}{2}}^n, \widehat{u}_{h,j+\frac{1}{2}}^{n+\vartheta,-}, \widehat{u}_{h,j+\frac{1}{2}}^{n+\vartheta,+}\right) \widehat{u}_{h,j+\frac{1}{2}}^{n+\vartheta,-} - \widehat{g}\left(\omega_{j-\frac{1}{2}}^n, \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,-}, \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,+}\right) \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,+} \right),
\end{aligned}$$

when the test function $\widehat{v} = \widehat{u}_h^{n+\vartheta}$ and the discrete transport equation (7) are applied. The next steps ensues similar as in the proof of the entropy inequality for the semi-discrete ALE-DG method in [9]. First of all, the integration by substitution formula and the function $H\left(\omega, u^-, u^+\right)$ are applied to write the method as

$$\begin{aligned}
0 \geq &\int_0^1 \Delta_j^{n+1} \eta\left(\widehat{u}_h^{n+1}\right) d\xi - \int_0^1 \Delta_j^n \eta\left(\widehat{u}_h^n\right) d\xi + \Theta_{j-\frac{1}{2}}^{n+\vartheta} \\
&+ \Delta t \left(H\left(\omega_{j+\frac{1}{2}}^n, \widehat{u}_{h,j+\frac{1}{2}}^{n+\vartheta,-}, \widehat{u}_{h,j+\frac{1}{2}}^{n+\vartheta,+}\right) - H\left(\omega_{j-\frac{1}{2}}^n, \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,-}, \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,+}\right)\right), \qquad (13)
\end{aligned}$$

where

$$\Theta_{j-\frac{1}{2}}^{n+\vartheta} := \Delta t \left(g\left(\omega_{j-\frac{1}{2}}^n, \theta_j^{n+\vartheta}\right) - \widehat{g}\left(\omega_{j-\frac{1}{2}}^n, \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,-}, \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,+}\right)\right) [\![\widehat{u}_h^{n+\vartheta}]\!]_{j-\frac{1}{2}}$$

with a value $\theta_j^{n+\vartheta}$ between $\widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,-}$ and $\widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,+}$ and $[\![\widehat{u}_h^{n+1}]\!]_{j-\frac{1}{2}} := \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,+} - \widehat{u}_{h,j-\frac{1}{2}}^{n+\vartheta,-}$. It should be noted that (13) is an inequality, since it has been assumed that ϑ satisfies the restriction (12) and thus the inequality (11) is satisfied, too. Moreover, the term $\Theta_{j-\frac{1}{2}}^{n+\vartheta}$ is nonnegative, since the method is considered with a monotone and consistent numerical flux. Next, the inequality (13) is transformed to the physical domain by the integration by substitution formula. The inequality on physical domain provides the

desired cell entropy inequality. Finally, a summation of the cell entropy inequality over all cells yields the L^2-stability, since we consider the problem (1) with periodic boundary conditions. □

The result in Proposition 1 holds merely for the square entropy function. Nevertheless, for the piecewise constant ϑ-P^0-ALE-DG method an entropy inequality with respect to the Kružkov entropy functions can be proven. Henceforth, the upcoming notation is used

$$\overline{u}_j(t) := \frac{1}{\Delta_j(t)} \int_{K_j(t)} u_h(t)\, dx, \quad \Delta_j^{n+1-\vartheta} := \vartheta \Delta_j^n + (1-\vartheta)\, \Delta_j^{n+1}. \tag{14}$$

The following identity follows from a simple calculation

$$\Delta_j^{n+1}\overline{u}_j^{n+1} - \Delta_j^n \overline{u}_j^n = \Delta_j^{n+1-\vartheta}\left(\overline{u}_j^{n+1} - \overline{u}_j^n\right) + \Delta t\left(\omega_{j+\frac{1}{2}}^n - \omega_{j-\frac{1}{2}}^n\right)\overline{u}_j^{n-\vartheta}, \tag{15}$$

since $\Delta_j^{n+1} - \Delta_j^n = \Delta t\left(\omega_{j+\frac{1}{2}}^n - \omega_{j-\frac{1}{2}}^n\right)$. The Eq. (15) provides the upcoming formulation of the ϑ-P^0-ALE-DG method

$$\begin{aligned} 0 = \overline{u}_j^{n+1} - \overline{u}_j^n &+ \frac{\Delta t}{\Delta_j^{n+1-\vartheta}}\left(\widehat{g}_+\left(\omega_{j-\frac{1}{2}}^n, \overline{u}_j^{n+\vartheta}\right) - \widehat{g}_+\left(\omega_{j-\frac{1}{2}}^n, \overline{u}_{j-1}^{n+\vartheta}\right)\right) \\ &- \frac{\Delta t}{\Delta_j^{n+1-\vartheta}}\left(\widehat{g}_-\left(\omega_{j+\frac{1}{2}}^n, \overline{u}_{j+1}^{n+\vartheta}\right) - \widehat{g}_-\left(\omega_{j+\frac{1}{2}}^n, \overline{u}_j^{n+\vartheta}\right)\right). \end{aligned} \tag{16}$$

In the following, an entropy inequality with respect to the Kružkov entropy functions $\eta_\ell(u) := |u-\ell|$, $\ell \in \mathbb{R}$, is presented for the method (16).

Proposition 2. *Suppose the CFL condition*

$$\left(\lambda_j^{n+\vartheta} + \frac{1}{2}\max_{t\in[t_n,t_{n+1}]}\left\{\left|\partial_x \omega(x,t)\, \Delta_j(t)\right| : x \in K_j(t)\right\}\right)\frac{\Delta t}{\Delta_j^{n+1-\vartheta}} \le 1, \tag{17}$$

where the parameters $\lambda_j^{n+\vartheta}$ and $\Delta_j^{n+1-\vartheta}$ are given by (5) and (14), respectively. Then the solution of the scheme (16) satisfies the cell entropy inequality

$$\eta_\ell\left(\overline{u}_j^{n+1}\right) - \eta_\ell\left(\overline{u}_j^n\right) + \frac{\Delta t}{\Delta_j^{n+1-\vartheta}}\left(H_\ell\left(\omega, \overline{u}_j^{n+\vartheta}, \overline{u}_{j+1}^{n+\vartheta}\right) - H_\ell\left(\omega, \overline{u}_{j-1}^{n+\vartheta}, \overline{u}_j^{n+\vartheta}\right)\right) \le 0,$$

where $\eta_\ell(u) := |u-\ell|$, $\ell \in \mathbb{R}$, are the Kružkov entropy functions and for all $a, b \in [m, M]$, $H_\ell(\omega, a, b)$ is given by

$$H_\ell(\omega,a,b) := \frac{1}{2}\int_\ell^a \eta_\ell'(v)\left(\lambda_j^n + f'(v) - \omega_{j-\frac{1}{2}}^n\right) dv$$
$$-\frac{1}{2}\int_\ell^b \eta_\ell'(v)\left(\lambda_j^n - f'(v) + \omega_{j+\frac{1}{2}}^n\right) dv.$$

Proof. The integration by parts formula and the convexity of the functions η_ℓ provide

$$\left(\overline{u}_j^{n+1} - \overline{u}_j^n\right)\eta_\ell'\left(\overline{u}_j^{n+1}\right) \geq \eta_\ell\left(\overline{u}_j^{n+1}\right) - \eta_\ell\left(\overline{u}_j^n\right) + \int_{\overline{u}_j^{n+\vartheta}}^{\overline{u}_j^{n+1}} \left(v - \overline{u}_j^{n+\vartheta}\right)\eta_\ell''(v)\, dv. \tag{18}$$

Next, the scheme (16) is multiplied by $\eta_\ell\left(\overline{u}_j^{n+1}\right)$ and the integration by parts formula, the functions $H_\ell(\omega,a,b)$, (9) and (18) are applied. This results in

$$0 \geq \eta_\ell\left(\overline{u}_j^{n+1}\right) - \eta_\ell\left(\overline{u}_j^n\right) + \Theta_j^{n+\vartheta} + \frac{\Delta t}{\Delta_j^{n+1}}\left(H_\ell\left(\omega,\overline{u}_j^{n+\vartheta},\overline{u}_{j+1}^{n+\vartheta}\right) - H_\ell\left(\omega,\overline{u}_{j-1}^{n+\vartheta},\overline{u}_j^{n+\vartheta}\right)\right), \tag{19}$$

where

$$\Theta_j^{n+\vartheta} := \left(1 - \frac{\Delta t}{\Delta_j^{n+1-\vartheta}} C\left(\lambda_j^{n+\vartheta}, \omega(t_{n+\vartheta})\right)\right)\int_{\overline{u}_j^{n+\vartheta}}^{\overline{u}_j^{n+1}} \left(v - \overline{u}_j^{n+\vartheta}\right)\eta_\ell''(v)\, dv$$
$$+ \int_{\overline{u}_{j-1}^{n+\vartheta}}^{\overline{u}_j^{n+1}} \left(\widehat{g}_+\left(\omega_{j+\frac{1}{2}}^n, v\right) - \widehat{g}_+\left(\omega_{j+\frac{1}{2}}^n, \overline{u}_{j-1}^{n+\vartheta}\right)\right)\eta_\ell''(v)\, dv$$
$$\int_{\overline{u}_{j+1}^{n+\vartheta}}^{\overline{u}_j^{n+1}} \left(\widehat{g}_-\left(\omega_{j+\frac{1}{2}}^n, v\right) - \widehat{g}_-\left(\omega_{j+\frac{1}{2}}^n, \overline{u}_{j+1}^{n+\vartheta}\right)\right)\eta_\ell''(v)\, dv, \tag{20}$$

$$C\left(\lambda_j^{n+\vartheta}, \omega(t_{n+\vartheta})\right) := \lambda_j^{n+\vartheta} + \frac{1}{2}\partial_x\omega(t_{n+\vartheta})\,\Delta_j(t_{n+\vartheta}).$$

The inequality (19) is almost the desired cell entropy inequality. Nevertheless, it need to be ensured that the term $\Theta_j^{n+\vartheta}$ is non-negative. In fact, the integrals in equation (20) are nonnegative, since the functions η_ℓ are convex and the functions $\widehat{g}_\pm(\omega,u)$ are monotone increasing. It should be noted that η_ℓ'' are Dirac delta distributions. However, the products in all the integrals are well defined, since the delta distributions are multiplied with continuous functions. Furthermore, the term in front of the first integral in equation (20) is nonnegative by the CFL condition (17). Hence, the term $\Theta_j^{n+\vartheta}$ is nonnegative and the inequality (19) yields the desired cell entropy inequality. □

3 Conclusions

In this paper, an ALE-DG method for solving scalar conservation laws has been presented. A cell entropy inequality with respect to the Kružkov entropy functions has been proven for the fully discrete ϑ-P^0-ALE-DG method. Likewise, a cell entropy inequality with respect to the square entropy function has been proven for the fully discrete ϑ-P^k-ALE-DG method, when ϑ satisfies the restriction (12). Cell entropy inequalities are very useful in the analysis of numerical methods. Besides the convergence to the physical relevant solution, cell entropy inequalities provide certain stability properties and statements about the qualitative behavior of a numerical method. For instance, a cell entropy inequality with respect to the square entropy function provides the L^2-stability of the method and is the key to a priori error estimates. Hence, in future works, it is of particular interest to prove cell entropy inequalities or at least the L^2-stability for the ALE-DG method when other time integration methods like the explicit third-order TVD-RK methods are adopted.

Acknowledgements The research of Yinhua Xia is supported by the NSFC grants No.11371342, No. 11471306.

References

1. G. Chavent, B. Cockburn, The local projection P^0-P^1-discontinuous-Galerkin finite element method for scalar conservation laws. Modélisation mathématique et analyse numérique (M^2AN) **23**, 565–592 (1989)
2. B. Cockburn, C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework. Math. Comput. **52**, 411–435 (1989)
3. B. Cockburn, S.-Y. Lin, C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one dimensional systems. J. Comput. Phys. **84**, 90–113 (1989)
4. B. Cockburn, C.-W. Shu, Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. J. Sci. Comput. **16**, 173–261 (2001)
5. J. Donea, A. Huerta, J.-P. Ponthot, A. Rodríguez-Ferran, Arbitrary Lagrangian-Eulerian methods, in *Encyclopedia of Computational Mechanics Volume 1: Fundamentals*, ed. by E. Stein, R. De Borst, T.J.R. Hughes (Wiley, New York, 2004), pp. 413–437
6. S. Gottlieb, C.-W. Shu, Total variation diminishing Runge-Kutta schemes. Math. Comput. **67**, 73–85 (1998)
7. H. Guillard, C. Farhat, On the significance of the geometric conservation law for flow computations on moving meshes. Comput. Method. Appl. Mech. Eng. **190**, 1467–1482 (2000)
8. G.S. Jiang, C.-W. Shu, On a cell entropy inequality for discontinuous Galerkin methods. Math. Comput. **62**, 531–538 (1994)
9. C. Klingenberg, G. Schnücke, Y. Xia, Arbitrary Lagrangian-Eulerian discontinuous Galerkin method for conservation laws: analysis and application in one dimension. Math. Comput. **86**, 1203–1232 (2017)
10. C.-W. Shu, Total-variation-diminishing time discretizations. SIAM J. Sci. Stat. Comput. **9**, 1073–1084 (1988)
11. V. Springel, E pur si muove: Galilean-invariant cosmological hydrodynamical simulations on a moving mesh. Oxf. J. Sci. Math. MNRAS **401**, 791–851 (2010)

Simplified Hyperbolic Moment Equations

Julian Koellermeier and Manuel Torrilhon

Abstract Hyperbolicity is a necessary property of model equations for the solution of the BGK equation to achieve stable and physical solutions. However, the standard approach for velocity space discretization developed by Grad only yields locally hyperbolic equations. The method has recently been improved, and several new globally hyperbolic model systems have been derived such as the Hyperbolic Moment Equations (HME) and the Quadrature-Based Moment Equations (QBME). We will describe the derivation and properties of a new model system called Simplified Hyperbolic Moment Equations (SHME) which inherits hyperbolicity from the other models but is simpler to implement and to solve. First simulation results show a good accuracy of the new SHME model in comparison with the existing models.

Keywords Moment method · Hyperbolicity · Boltzmann equation

1 Introduction

There are several possible solution methods for the BGK equation with varying success. Among those are direct solvers like discrete velocity methods [12], particle methods like DSMC [1], and moment methods as for example [13]. A relatively old moment approach was developed by Grad in [6], but due to problems regarding the loss of hyperbolicity of the equations, there has not been much research on this approach for a long time. Hyperbolicity is an important property because otherwise there will be imaginary eigenvalues creating instabilities and non-physical solutions. Grad's equations have been shown to be only hyperbolic in a small region around equilibrium so that the solution can break down for strong non-equilibrium; see [4].

J. Koellermeier (✉) · M. Torrilhon
MathCCES, RWTH Aachen University, Schinkelstrasse 2, 52062 Aachen, Germany
e-mail: koellermeier@mathcces.rwth-aachen.de

M. Torrilhon
e-mail: mt@mathcces.rwth-aachen.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_17

There has been a lot of work regarding different hyperbolic approaches like the maximum entropy method by Levermore [11], and the method gives very accurate results, but is also extremely complex because it requires the solution of a nonlinear optimization problem in every step. Other methods like the Pearson IV model developed by Torrilhon [14] seem to be difficult to generalize to the multi-dimensional case.

Recently, several new hyperbolic moment models have been developed that are based on Grad's approach but modify the system of equations in different ways to achieve global hyperbolicity of the equations. Two examples are the Hyperbolic Moment Equations (HME) by Cai [4] and the Quadrature-Based Moment Equations (QBME) by Koellermeier [10]. As both new models only approximate the original system, it is still necessary to investigate these models with respect to accuracy in various situations. Furthermore, the development of other models is possible, for example using the operator projection approach as explained in [5].

In this paper, we present a new hyperbolic moment model called Simplified Hyperbolic Moment Equations (SHME) that we derive using a special approximation during Grad's method. The new model equations can be explicitly derived, and we also show that the model is globally hyperbolic as a special linearization of Grad's equations around equilibrium.

The paper is organized as follows: We first recall the BGK equation, the derivation of the moment method, and some existing hyperbolic models in Sect. 2 before we derive the new SHME model in the main Sect. 3. We show some shock tube results in Sect. 4, and the paper ends with a conclusion.

2 Moment Method for the BGK Equation

In one spatial dimension, the BGK equation describing the change of the particle distribution function $f(t,x,c)$ reads as follows

$$\frac{\partial}{\partial t} f(t,x,c) + c\frac{\partial}{\partial x} f(t,x,c) = S(f), \tag{1}$$

where we will consider the BGK collision operator [2] with relaxation time τ on the right-hand side

$$S(f) = -\frac{1}{\tau}(f - f_M) \tag{2}$$

and the equilibrium Maxwellian is given by

$$f_M(t,x,c) = \frac{\rho(t,x)}{\sqrt{2\pi\theta(t,x)}} \exp\left(-\frac{(c-u(t,x))^2}{2\theta(t,x)}\right). \tag{3}$$

The macroscopic quantities density ρ, velocity u, and temperature θ can be computed via integration of the distribution function over velocity space

$$\rho(t,x) = \int_{\mathbb{R}} f(t,x,c)\, dc, \tag{4}$$

$$\rho(t,x)u(t,x) = \int_{\mathbb{R}} cf(t,x,c)\, dc, \tag{5}$$

$$\rho(t,x)\theta(t,x) = \int_{\mathbb{R}} |c-u|^2 f(t,x,c)\, dc. \tag{6}$$

The solution of (1) is particularly difficult as it requires an additional discretization of the velocity space. As an efficient way to perform this, we will use the moment method. This method requires different steps that have been outlined previously, e.g., in [4], and we will recall these steps here to derive a new simplified model later.

1. **Expansion of the distribution function**
   The distribution function $f(t,x,c)$ is first expanded in velocity space in a series of basis functions $\phi_\alpha^{[u(t,x),\theta(t,x)]}$ as follows

$$f(t,x,c) = \sum_{\alpha\in\mathbb{N}} f_\alpha(t,x)\phi_\alpha^{[u(t,x),\theta(t,x)]}\left(\frac{c-u}{\sqrt{\theta}}\right) \tag{7}$$

$$= \sum_{\alpha\in\mathbb{N}} f_\alpha(t,x)\phi_\alpha^{[u,\theta]}(\xi) \tag{8}$$

   with new velocity variable

$$\xi = \frac{c-u}{\sqrt{\theta}}, \tag{9}$$

   see remark below. Furthermore, the superscripts mean that the basis function will depend on the macroscopic quantities $u(t,x)$, $\theta(t,x)$. From now on, we will omit the arguments t, x in u and θ to shorten notation. Additionally, we use Einstein's summation notation wherever possible to abbreviate the sum in (8).
2. **Definition and properties of the basis functions**
   The basis functions are defined as weighted Hermite polynomials according to

$$\phi_\alpha^{[u,\theta]}(\xi) = \frac{1}{\sqrt{2\pi}}\theta^{-\frac{\alpha+1}{2}}\mathscr{H}_\alpha(\xi)\exp\left(-\frac{\xi^2}{2}\right), \tag{10}$$

   where $\mathscr{H}_\alpha$ is the Hermite polynomial of degree α.

Remark 1. The argument $\xi = \frac{c-u}{\sqrt{\theta}}$ can be seen as a transformed velocity variable in which the expansion is performed. The microscopic velocity c is shifted by its mean u and scaled by its variance $\sqrt{\theta}$ such that the new velocity variable ξ is normalized with variance 1 and mean 0. This requires less basis functions for the velocity discretization later but leads to a more complicated PDE to discretize.

In order to derive the moment equations, we will need to compute derivatives of the basis functions which in turn need the computation of the Hermite derivatives. We note that the Hermite polynomials fulfill the following formulas:

- Recursion relation:

$$\mathscr{H}_{\alpha+1}(x) = x\mathscr{H}_\alpha(x) - \alpha\mathscr{H}_{\alpha-1}(x) \tag{11}$$

- Derivative:

$$\mathscr{H}'_\alpha(x) = \alpha\mathscr{H}_{\alpha-1}(x), \tag{12}$$

- Weighted derivative:

$$[\mathscr{H}_\alpha(x)\exp(-x^2/2)]' = -\alpha\mathscr{H}_{\alpha+1}(x)\exp(-x^2/2) \tag{13}$$

With some basic calculations, it is now possible to verify that the basis functions $\phi_\alpha^{[u,\theta]}(\xi)$ satisfy

- Recursion relation:

$$\sqrt{\theta}\xi\phi_\alpha^{[u,\theta]}(\xi) = \theta\phi_{\alpha+1}^{[u,\theta]}(\xi) + \alpha\phi_{\alpha-1}^{[u,\theta]}(\xi), \tag{14}$$

- Derivative:

$$\frac{\partial}{\partial\xi}\phi_\alpha^{[u,\theta]}(\xi) = -\sqrt{\theta}\phi_{\alpha+1}^{[u,\theta]}(\xi), \tag{15}$$

- θ derivative:

$$\frac{\partial}{\partial\theta}\phi_\alpha^{[u,\theta]}(\xi) = -\frac{\alpha+1}{2\theta}\phi_\alpha^{[u,\theta]}(\xi). \tag{16}$$

3. **Compatibility constraints**
   According to the definition of the macroscopic quantities, the following conditions can be derived by inserting the ansatz (7) into (4)–(6)

$$f_0 = \rho,\ f_1 = f_2 = 0. \tag{17}$$

   This constrains the unknown coefficients f_α to the subspace where the macroscopic quantities are recovered.
4. **Derivation of the moment equations**
   The derivation of the moment equations in general form now only needs the computation of the terms in the BGK equation (1). The terms $\partial_t f$ and $\partial_x f$ are computed in the following way for $s = x, t$

$$\frac{\partial}{\partial s} f(t,x,c) = \frac{\partial}{\partial s}\left(f_\alpha(t,x)\phi_\alpha^{[u,\theta]}(\xi)\right) = \frac{\partial f_\alpha(t,x)}{\partial s}\phi_\alpha^{[u,\theta]}(\xi) + f_\alpha(t,x)\frac{\partial \phi_\alpha^{[u,\theta]}(\xi)}{\partial s} \tag{18}$$

with

$$\frac{\partial \phi_\alpha^{[u,\theta]}(\xi)}{\partial s} = \frac{\partial \phi_\alpha^{[u,\theta]}(\xi)}{\partial \theta}\frac{\partial \theta}{\partial s} + \frac{\partial \phi_\alpha^{[u,\theta]}(\xi)}{\partial \xi}\frac{\partial \xi}{\partial s}, \tag{19}$$

where the remaining unknown term is given by

$$\frac{\partial \xi}{\partial s} = -\frac{1}{\sqrt{\theta}}\frac{\partial u}{\partial x} - \frac{\xi}{2\theta}\frac{\partial \theta}{\partial s}. \tag{20}$$

Using the previously computed formulas, we get

$$\frac{\partial}{\partial s} f(t,x,c) = \frac{\partial f_\alpha}{\partial s}\phi_\alpha^{[u,\theta]}(\xi) + \frac{\partial u}{\partial s} f_\alpha \phi_{\alpha+1}^{[u,\theta]}(\xi) + \frac{1}{2}\frac{\partial \theta}{\partial s} f_\alpha \phi_{\alpha+2}^{[u,\theta]}(\xi) \tag{21}$$

or equivalently using an index shift in the Einstein notation of the infinite sum $f_\alpha \phi_\alpha^{[u,\theta]}$

$$\frac{\partial}{\partial s} f(t,x,c) = \left(\frac{\partial f_\alpha}{\partial s} + \frac{\partial u}{\partial s} f_{\alpha-1} + \frac{1}{2}\frac{\partial \theta}{\partial s} f_{\alpha-2}\right)\phi_\alpha^{[u,\theta]}(\xi). \tag{22}$$

The convective term $c\frac{\partial}{\partial x} f(t,x,c)$ can now be computed using (9), (14), and (22)

$$\begin{aligned} c\frac{\partial}{\partial x} f(t,x,c) &= \left(u + \sqrt{\theta}\xi\right)\frac{\partial}{\partial x} f(t,x,c) \\ &= \phi_\alpha^{[u,\theta]}(\xi)\left(\theta\frac{\partial f_{\alpha-1}}{\partial x} + u\frac{\partial f_\alpha}{\partial x} + (\alpha+1)\frac{\partial f_{\alpha+1}}{\partial x}\right. \\ &\quad + \frac{\partial u}{\partial x}\left(\theta f_{\alpha-2} + u f_{\alpha-1} + (\alpha+1) f_\alpha\right) \\ &\quad \left. + \frac{1}{2}\frac{\partial \theta}{\partial x}\left(\theta f_{\alpha-3} + u f_{\alpha-2} + (\alpha+1) f_{\alpha-1}\right)\right) \end{aligned} \tag{23}$$

5. **Right-hand side collision term**
The right-hand side collision term is simply computed by inserting expansion (7) into (2). Due to the definition of the equilibrium Maxwellian (3), we get

$$S(f) = -\frac{1-\delta_{0\alpha}}{\tau} f_\alpha \phi_\alpha^{[u,\theta]}, \text{ with } \delta_{0\alpha} = \begin{cases} 1, & \alpha = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

6. **Matrix form of the moment system**
Cutting off the expansion (7) at $M \in \mathbb{N}$, we get $M+1$ unknowns that we write as $\mathbf{w}_M = (\rho, u, \theta, f_3, \ldots f_M)$. The moment system for the unknown vector $\mathbf{w}_M$ can be directly obtained by matching coefficients of the basis functions in (22), (23), and (24). We can write the system in the following form

$$\frac{\partial \mathbf{w}_M}{\partial t} + \mathbf{A}_{\text{Grad}} \frac{\partial \mathbf{w}_M}{\partial x} = \mathbf{S}_M, \tag{25}$$

where the right-hand side reads

$$\mathbf{S}_M = -\frac{1}{\tau} \mathbf{P} \mathbf{w}_M \tag{26}$$

for diagonal matrix $\mathbf{P} = \text{diag}\,(0, 0, 0, 1, \ldots, 1)$.
The explicit expressions for (25) can be found in [4]. For the famous five-moment case, the system matrix reads

$$\mathbf{A}_{\text{Grad}} = \begin{pmatrix} u & \rho & 0 & 0 & 0 \\ \frac{\theta}{\rho} & u & 1 & 0 & 0 \\ 0 & 2\theta & u & \frac{6}{\rho} & 0 \\ 0 & 4f_3 & \frac{\rho\theta}{2} & u & 4 \\ -\frac{f_3\theta}{\rho} & 5f_4 & \frac{3f_3}{2} & \theta & u \end{pmatrix}. \tag{27}$$

Unfortunately, it has been shown that the system (25) loses hyperbolicity for already moderate non-equilibrium values; see, e.g., [4]. This can lead to non-physical values and a breakdown of the solution as exemplified in [9]. It is thus of major importance to derive models with unbounded hyperbolicity regions.

2.1 Existing Hyperbolic Moment Models

Two existing moment models that are globally hyperbolic have been derived in [4, 7]. The Hyperbolic Moment Equations (HME) and the Quadrature-Based Moment Equations (QBME) will be used as comparison for later computations in Sect. 4. For the five-moment test case, the system matrices are given by

$$\mathbf{A}_{\text{HME}} = \begin{pmatrix} u & \rho & 0 & 0 & 0 \\ \frac{\theta}{\rho} & u & 1 & 0 & 0 \\ 0 & 2\theta & u & \frac{6}{\rho} & 0 \\ 0 & 4f_3 & \frac{\rho\theta}{2} & u & 4 \\ -\frac{f_3\theta}{\rho} & 0 & -f_3 & \theta & u \end{pmatrix}, \quad \mathbf{A}_{\text{QBME}} = \begin{pmatrix} v & \rho & 0 & 0 & 0 \\ \frac{\theta}{\rho} & v & 1 & 0 & 0 \\ 0 & 2\theta & v & \frac{6}{\rho} & 0 \\ 0 & 4f_3 & \frac{\rho\theta}{2} - \frac{10f_4}{\theta} & v & 4 \\ -\frac{f_3\theta}{\rho} & 5f_4 & -f_3 & \theta + \frac{15f_4}{\rho\theta} & v \end{pmatrix}. \tag{28}$$

Both models are globally hyperbolic due to the marked changes in the system matrix with respect to Grad's model, and they have recently been summarized in a framework for the derivation of hyperbolic moment models in [5].

3 A New Simplified Hyperbolic Moment Model SHME

In order to derive efficient but simple moment models to accurately capture flow phenomena beyond the standard fluid dynamics equations, we aim to derive new models that overcome difficulties of the standard Grad model. As a result of the derivation of Grad's equations, the problematic loss of hyperbolicity is caused by the effects of the higher-order coefficients in the equations. These coefficients enter the equation system because the basis functions depend on the shifted velocity variable ξ and thus on u and θ. This dependence is reasonable as it yields an efficient approximation, but it effectively spoils the hyperbolicity of the moment equations.

In order to reduce the nonlinearity introduced by the complicated choice of the basis functions, we will reduce the model complexity using the following approximation to Eq. (19), where $s = x, t$

$$\frac{\partial \phi_\alpha^{[u,\theta]}(\xi)}{\partial s} = \frac{\partial \phi_\alpha^{[u,\theta]}(\xi)}{\partial \theta}\frac{\partial \theta}{\partial s} + \frac{\partial \phi_\alpha^{[u,\theta]}(\xi)}{\partial \xi}\frac{\partial \xi}{\partial s} \approx 0, \tag{29}$$

meaning that the derivative of the basis function with respect to time and space is set to zero, which effectively means that the basis functions are treated as if they only depended on a fixed velocity space. This leads to large simplifications as it will cancel most terms containing derivatives with respect to the coefficients f_α. However, we must make sure not to change the conservation laws of mass, momentum, and energy. This is ensured by applying the approximation only to the last $M - 2$ equations, while keeping the first three equations as before.

From a physical point of view, the model can be seen as a linearization of the full model (27), which is extremely nonlinear due to the expansion in the transformed variable (9). By neglecting the respective terms in (29), this nonlinearity is reduced; see also Sects. 3.1 and 3.2 for further interpretations of the simplification in (29).

The new model is called *Simplified Hyperbolic Moment Equations* (SHME), due to the simplification made in (29). The model results in the following system

$$\frac{\partial \mathbf{w}_M}{\partial t} + \mathbf{A}_{\text{SHME}}\frac{\partial \mathbf{w}_M}{\partial x} = \mathbf{S}_M, \tag{30}$$

and in the five-moment case we obtain the system matrix

$$\mathbf{A}_{\text{SHME}} = \begin{pmatrix} u & \rho & 0 & 0 & 0 \\ \frac{\theta}{\rho} & u & 1 & 0 & 0 \\ 0 & 2\theta & u & \frac{6}{\rho} & 0 \\ 0 & 0 & \frac{\rho\theta}{2} & u & 4 \\ 0 & 0 & 0 & \theta & u \end{pmatrix} \tag{31}$$

and for $M > 4$ the matrix is a consistent extension of the tridiagonal matrix in (31).

Note that the system matrix does not depend on the higher-order coefficients any more and is tridiagonal which leads to a reduction of complexity when it comes to

implementing numerical schemes. However, it is important to analyze the effect of this simplification with respect to model accuracy and hyperbolicity as well.

3.1 Discussion of the SHME Model

Comparing the matrix (31) with the original system matrix (27), we see that the SHME system is exactly the Grad system evaluated at equilibrium; i.e. all higher-order coefficients in the matrix are set to zero. SHME can therefore be seen as the linearization of the original Grad system around equilibrium. The characteristic polynomial of the SHME system matrix (31) can thus be computed analogously to Grad's system at equilibrium from which we can directly conclude that the SHME model is globally hyperbolic, as all eigenvalues of (31) are real.

The linearization of Grad's system around equilibrium might sound too simple, but it is actually very similar to the approach by Cai et al. in [4]. In [8], the HME system was written in convective variables and it was shown that the HME system matrix in these variables is nothing else than the convective Grad matrix at equilibrium. In the same way, the QBME model can be written as a linear deviation from the convective Grad system. In that sense, the SHME model is just another reasonable approximation of Grad's system in the original set of variables.

3.2 Relation to Discrete Velocity Model

A different approach to achieve a very simple model for the solution of the BGK equation (1) is the discrete velocity method (DVM) [12]. It uses point evaluations of the BGK equation at fixed velocity points $c_i \in \mathbb{R}$, $i = 0, \dots, M$ to discretize the equation in velocity space as follows

$$\frac{\partial}{\partial t} f_i + c_i \frac{\partial}{\partial x} f_i = -\frac{1}{\tau} \left(f_i - f_M(c_i) \right) \tag{32}$$

This is computationally highly efficient as there is no velocity transformation that results in more complicated equations. The system matrix is just a diagonal matrix $\mathbf{A}_{\text{DVM}} = \text{diag}\,(c_0, \dots, c_M)$ with the discrete velocities on the diagonal.

However, far more velocity points are needed to accurately capture the flow, especially for varying mean velocities in the flow field or over time. As one example, more than 400 velocities were used to compute the DVM reference solution Sect. 4 in comparison with HME and QBME that used only between five and ten variables.

Without the approximation in (29), the basis functions enable a very efficient approach for the discretization in velocity space as the basis effect of the transformed velocity space yields physical adaptivity and also the effect of the derivative is exactly taken into account by (19).

The new SHME model on the other hand simplifies the method in that the derivative of the basis function is neglected; see (29). This reduces the adaptivity of the method. However, the transformation of the velocity variable (9) is still used in all other steps of the derivation such as in the expansion (7) and during the computation of the term $c\frac{\partial}{\partial x} f(t, x, c)$, where the velocity c is substituted by the transformation rule as $c = u + \sqrt{\theta}\xi$. SHME is thus still adaptive but simply neglects some of the nonlinear effects of the adaptivity. We can say that the new SHME method is in the middle between the standard Grad approach and the DVM method.

4 Simulation Results

We test the accuracy of the new SHME model using a shock tube test case as also done for the HME and QBME models in previous papers; see, e.g., [4]. The model equation reads

$$\partial_t \mathbf{w} + \mathbf{A}\partial_x \mathbf{w} = -\frac{1}{\tau}\mathbf{P}\mathbf{w}, \tag{33}$$

where the system matrix varies depending on the model used. Using $M \geq 4$, we solve for variables $\mathbf{w} = (\rho, u, \theta, f_3, f_4, \ldots, f_M)$. The collisions are modeled using a BGK operator with nonlinear relaxation time $\tau = \frac{\mathrm{Kn}}{\rho}$ that leads to the following form of the matrix $\mathbf{P}$

$$\mathbf{P} = \mathrm{diag}\,(0, 0, 0, 1, 1, \ldots) \in \mathbb{R}^{(M+1)\times(M+1)}. \tag{34}$$

We will consider two Knudsen numbers $\mathrm{Kn}_1 = 0.05$ and $\mathrm{Kn}_2 = 0.5$.

The initial condition is given by

$$\mathbf{w}(0, x) = \begin{cases} \mathbf{w}^L, & \text{if } x < 0 \\ \mathbf{w}^R, & \text{if } x > 0 \end{cases} \tag{35}$$

and according to the tests by Cai et al. [4] the left and right states are chosen as

$$\mathbf{w}^L = (7, 0, 1, 0, 0, \ldots, 0)^T, \qquad \mathbf{w}^R = (1, 0, 1, 0, 0, \ldots, 0)^T, \tag{36}$$

corresponding to a jump in density at the initial discontinuity at $x = 0$.

For the spatial discretization, we used 4000 cells in the domain $[-2, 2]$ and the results show the solution at $t_{\mathrm{END}} = 0.3$ using a constant $\Delta t = 0.0001$. The numerical scheme to solve the non-conservative PDE system is the PRICE scheme of Canestrelli [3] that was also used in [9].

The numerical results for the SHME method in comparison with the other moment models HME, QBME, and a DVM reference solution are shown in Figs. 1 and 2. For $\mathrm{Kn} = 0.05$, the results are almost identical to the HME and QBME results. There are

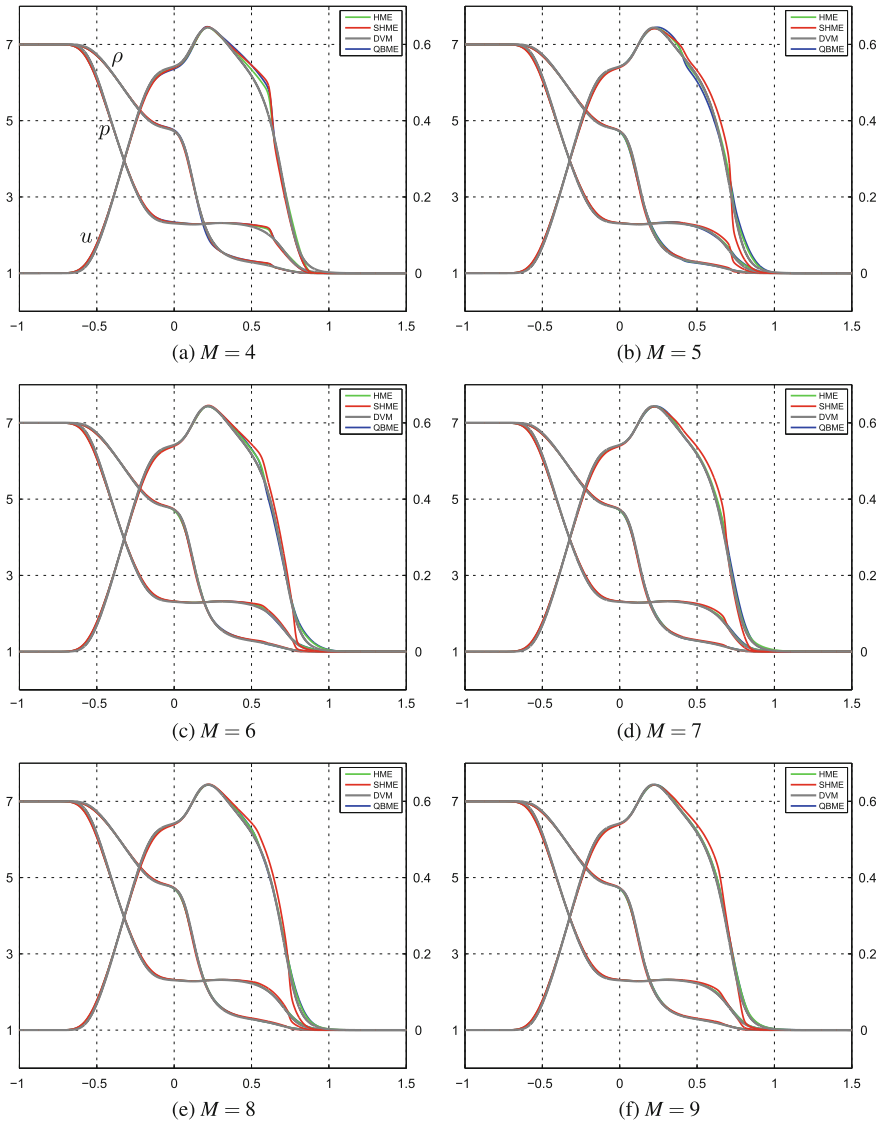


Fig. 1 Moment model comparison for SHME, HME, QBME, and DVM reference solution, Kn = 0.05. The left y-axis is for ρ and p, and the right y-axis is for u

only small differences with respect to the other methods. The approximation quality is good even for larger M. In the case of Kn = 0.5 in Fig. 1 we see that the SHME model is between HME and QBME for $M = 4$. However, the differences are larger for a larger number of moments M. This is as expected and due to the fact that more and more coefficients are neglected in the SHME approximation when increasing

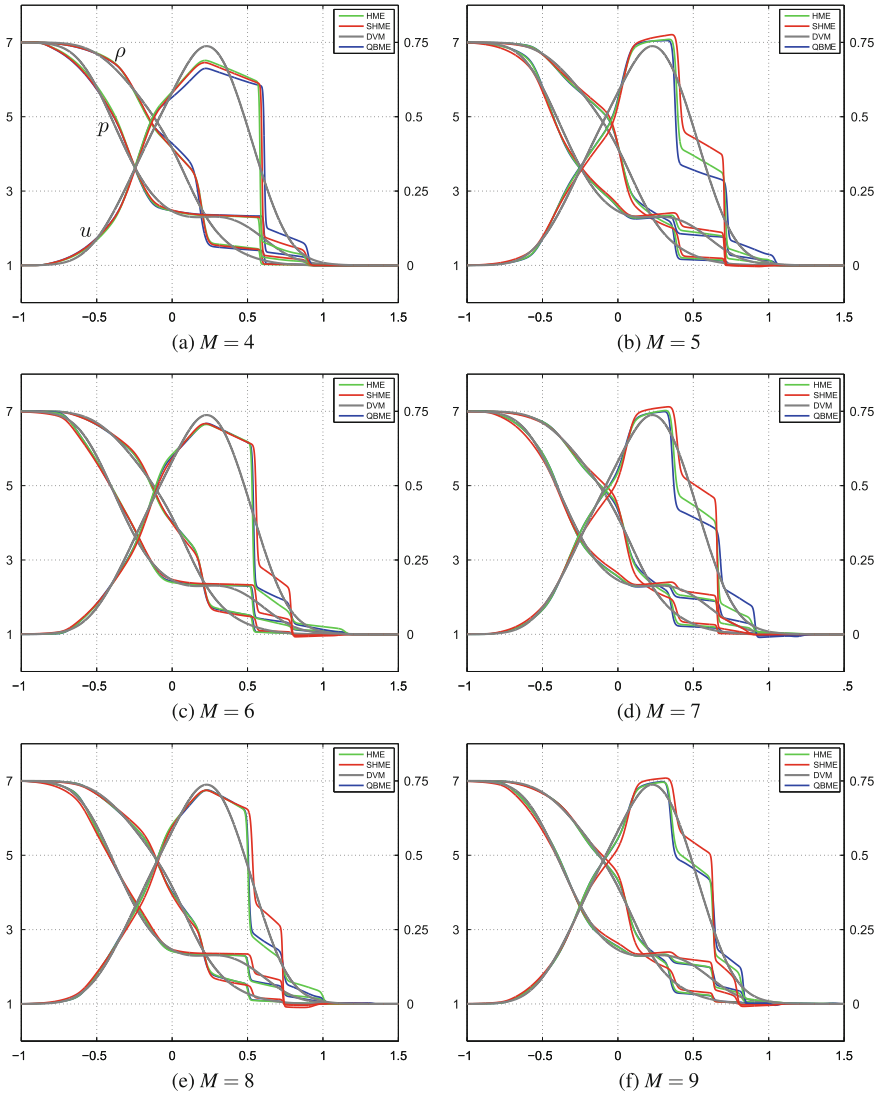


Fig. 2 Moment model comparison for SHME, HME, QBME, and DVM reference solution, Kn = 0.5. The left y-axis is for ρ and p, and the right y-axis is for u

M. Still, the model yields reasonably good results for small M, especially regarding the simplicity of the model. The SHME solution is not far from the DVM reference solution.

5 Conclusion

In this paper we derived a new hyperbolic moment model for the BGK Equation based on the approximation of several nonlinear terms during the derivation. The new model equations called Simplified Hyperbolic Moment Equations (SHME) have been shown to be globally hyperbolic. We compared the model with the discrete velocity model and motivated the use of the new equations by linearization of Grad's equations keeping as much of the adaptivity as possible. The results have shown that the model accuracy is good despite the reduced complexity and the simplified derivation. In order to characterize the new model in more detail additional investigations are necessary for example regarding the stability, convergence and more complex multi-dimensional test cases.

References

1. G.A. Bird, Direct simulation and the Boltzmann equation. Phys. Fluids **13**, 2676–2687 (1970)
2. P.L. Bhatnagar, E.P. Gross, M. Krook, A model for collision processes in gases. I. small amplitude processes in charged and neutral one-component systems. Phys. Rev. **94**(3), 511–525 (1954)
3. A. Canestrelli, *Numerical Modelling of Alluvial Rivers by Shock Capturing Methods*. Ph.D. thesis, Universita' Degli Studi di Padova (2008)
4. Z. Cai, Y. Fan, R. Li, Globally hyperbolic regularization of Grad's moment system in one dimensional space. Commun. Math. Sci. **11**(2), 547–571 (2013)
5. Y. Fan, J. Koellermeier, J. Li, R. Li, M. Torrilhon, Model reduction of kinetic equations by operator projection. J. Stat. Phys. **162**(2), 457–486 (2016)
6. H. Grad, On the kinetic theory of rarefied gases. Commun. Pure Appl. Math. **2**(4), 331–407 (1949)
7. J. Koellermeier, Hyperbolic approximation of kinetic equations using quadrature-based projection methods. Master's thesis, RWTH Aachen University (2013)
8. J. Koellermeier, M. Torrilhon, On new hyperbolic moment models for the boltzmann equation, in *Conference Proceedings of the YIC GACM 2015, Publication Server of RWTH Aachen University* (2015)
9. J. Koellermeier, M. Torrilhon, Numerical study of partially conservative moment equations in kinetic theory. Commun. Comput. Phys. **21**(04), 987–1011 (2017)
10. J. Koellermeier, R. Schaerer, M. Torrilhon, A framework for hyperbolic approximation of kinetic equations using quadrature-based projection methods. Kinet. Relat. Models **7**(3), 531–549 (2014)
11. C.D. Levermore, Moment closure hierarchies for kinetic theories. J. Stat. Phys. **83**, 1021–1065 (1996)
12. L. Mieussens, Discrete velocity model and implicit scheme for the BGK equation of rarefied gas dynamics. Math. Models Methods Appl. Sci. **10**(8), 1121–1149 (2000)
13. H. Struchtrup, M. Torrilhon, Regularization of Grad's 13 moment equations: derivation and linear analysis. Phys. Fluids **15**(9), 2668–2680 (2003)
14. M. Torrilhon, Hyperbolic moment equations in kinetic gas theory based on multi-variate Pearson-IV-distributions. Commun. Comput. Phys. **7**(4), 639–673 (2010)

Weakly Coupled Systems of Conservation Laws on Moving Surfaces

Andrea Korsch

Abstract We consider weakly coupled systems of nonlinear hyperbolic conservation laws on moving surfaces. As in the Euclidean space, see, for example, (Levy, Commun Partial Differ Equ 17(3–4):657–698, 1992, [9], Rohde, Weakly coupled systems of hyperbolic conservation laws. PhD thesis, Mathematische Fakultät der Albert-Ludwigs-Universität Freiburg 1996, [16]), the coupling is realized by a source term, which only depends on (x, t) and the unknown function $u(x, t)$ but not on its derivatives. Scalar conservation laws on moving surfaces were considered in Dziuk et al., Interfaces Free Boundaries **15**:202–236, 2013, [4]), Lengeler and Müller (J Differ Equ 254(4):1705–1727, 2013, [10]). The velocity of the surface is given by a smooth function, and we assume the surface to be compact. We prove the existence for an entropy solution. First, we consider the regularized parabolic problem with viscosity parameter ε and show that there exists a weak solution by decoupling and linearizing the problem. Then, we prove the boundedness of this solution in $L^\infty(G_T)$, use standard regularity results to prove that this solution is a solution in the classical sense, and show uniform boundedness in $L^\infty(G_T)$ and $W^{1,1}(G_T)$ with respect to ε.

Keywords Weakly coupled systems · Time dependent surface · Conservation laws · Entropy solution

1 Introduction

Weakly coupled systems of conservation laws are studied in many applications and as simplified models for more complex systems or are taken for numerical approximation. This type of hyperbolic differential equation systems is often used to describe physical problems of continuum mechanics, biomathematics, or chemical processes. For example, Majda's model for dynamical combustion was studied in [9]. Further

A. Korsch (✉)
Albert-Ludwigs Universität, Abteilung für angewandte Mathematik,
Hermann-Herder Str.10, 79106 Freiburg, Germany
e-mail: andrea.korsch@mathematik.uni-freiburg.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_18

applications are transport and adsorption in porous media, hyperbolic random walk systems, and resonant waves [16]. A relevant example for considering moving surfaces is transport problems on biomembranes. The coupling term could be interpreted as a first step to take coupling of partial differential equations on bulk phases and interfaces into account.

The results we present in this contribution are part of our PhD thesis, which is submitted [7]. We consider the following problem: On a given space–time surface $G_T := \bigcup_{t\in(0,T)} \Gamma(t) \times \{t\}$, where $\Gamma(t) \subset \mathbb{R}^{n+1}$ is a smooth time dependent compact hypersurface (i.e., without boundary), where the dimension $n \in \mathbb{N}$ is arbitrary and $(0, T)$ is a finite time interval, we consider the hyperbolic initial value problem of weakly coupled systems of conservation laws. Find a vector of unknown functions $\mathbf{u} : G_T \to \mathbb{R}^M$, $\mathbf{u} = (u^1, ..., u^M)$, which satisfies

$$\begin{aligned}
\dot{u}^1(x,t) + u^1(x,t)\nabla_\Gamma \cdot \nu(x,t) + \nabla_\Gamma \cdot (f^1(x,t,u^1)) &= r^1(x,t,\mathbf{u}),\\
\dot{u}^2(x,t) + u^2(x,t)\nabla_\Gamma \cdot \nu(x,t) + \nabla_\Gamma \cdot (f^2(x,t,u^2)) &= r^2(x,t,\mathbf{u}),\\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots & \qquad (1)\\
\dot{u}^M(x,t) + u^M(x,t)\nabla_\Gamma \cdot \nu(x,t) + \nabla_\Gamma \cdot (f^M(x,t,u^M)) &= r^M(x,t,\mathbf{u})
\end{aligned}$$

for $(x, t) \in G_T$ and admits the initial values $\mathbf{u}(x, 0) = \mathbf{u}_0(x)$ for $x \in \Gamma_0$. Here, $M \in \mathbb{N}$ is the number of the scalar problems, which are coupled to each other. The velocity of the surface ν, the nonlinear flux functions $\mathbf{f}_k = (f_k^1, ..., f_k^M)$ with $k = 1, ..., n + 1$, the coupling term $\mathbf{r} = (r^1, ..., r^M)$, and the initial value $\mathbf{u}_0$ are given. Each component of the system corresponds to a conservation law on moving surfaces. For the modeling and derivation, see, for example, [4]. We added a coupling term which can be considered as a source term. This additional term $\mathbf{r}$ depends on the variables (x, t) and on the unknown function $\mathbf{u}$ but not on its derivatives, which is the reason why we call such systems weakly coupled.

We present a summary of the existing work done in this field to the best of our knowledge: First, there is the Euclidean case where the weakly coupled systems are considered as Cauchy problems on $\mathbb{R}^2$ in Levy [9] and Rohde [16] or later in several space dimensions in Rohde [17]. Here, the authors looked at the model introduced by Majda that can be found in [11]. In the Euclidean case, for example, O. A. Ladyzenskaja and collaborators dealt with quasi-linear parabolic systems [8]. Further work, with interesting applications in radiation hydrodynamics, chemosensitive movement, and numerics for convection dominated parabolic systems can be found in [6, 14, 18].

Scalar conservation laws on surfaces without coupling has been studied in [4]. In the paper of Lengeler and Müller [10] as well as in the thesis of Müller [12], the authors proposed the scalar problem on Riemannian manifolds with time-dependent metric. This is a different approach to consider time-dependent manifolds. They proved for the compactness result a TV estimate. In the paper of Dziuk and Elliott [3], the authors derived a proof for the existence of a weak solution of the parabolic

scalar problem on moving surfaces, inter alia. More related research is done by Amorim, Ben-Artzi, LeFloch, and Panov in [1, 2, 15].

Assumptions Let Γ_0 be a compact, i.e., without boundary, smooth-oriented hypersurface in $\mathbb{R}^{n+1}$, and let $\{(U_i, \xi_i) | i \in I)\}$ be a parametrization of this surface. Let $T > 0$. We assume that there exists a diffeomorphism $\phi(\cdot, t) : \Gamma_0 \to \Gamma(t)$ with $\phi \in C^\infty(\Gamma_0 \times [0, T])$ and $\phi(\cdot, 0) = id_{\Gamma_0}$, which describes the movement of the surface. We define the space time area:

$$G_T := \bigcup_{t \in (0,T)} \Gamma(t) \times \{t\}. \tag{2}$$

Then, the parametrization of G_T is given by $\{(\psi_i, U_i) \, | i \in I\}$, with smooth functions $\psi_i(x, t) = (\phi(\xi_i(x), t), t)$. Let the flux functions $\mathbf{f}_l = (f_l^1, ..., f_l^M)$ with index $l \in \{1, ..., n + 1\}$ and $f_l^i : \overline{G_T} \times \mathbb{R} \to \mathbb{R}$ for $i = 1, ..., M$ be given. We assume that the functions f_l^i are in $C^3(G_T \times \mathbb{R})$ and f^i to be divergence free, which means $\nabla_\Gamma \cdot f^i(\cdot, t, s) = 0$ for all fixed $t \in \mathbb{R}^+$, $s \in \mathbb{R}$ and $i \in \{1, ..., M\}$. Furthermore, let the coupling term $\mathbf{r}$ be given with $\mathbf{r} = (r^1, ..., r^M)$, $r^i : \overline{G_T} \times \mathbb{R}^M \to \mathbb{R}$. We assume that the coupling term $\mathbf{r}$ is in $C^2(G_T \times \mathbb{R}^M)$, and it satisfies a global Lipschitz condition with respect to $\mathbf{u}$ in the following way: For all $i = 1, \ldots, M$ exists a constant $L_r^i \geq 0$, such that

$$|r^i(x, t, \mathbf{u}) - r^i(x, t, \mathbf{v})| \leq L_r^i[|u^1 - v^1| + \cdots + |u^M - v^M|] \tag{3}$$

holds for all $(x, t) \in \overline{G_T}$ and for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^M$. Moreover, the initial value $\mathbf{u}_0 \in L^\infty(G_T)^M$ is given.

Geometry Here, we define geometrical quantities on hypersurfaces and state basic calculation rules for moving surfaces.

Definition 1. Let $\Gamma \subset \mathbb{R}^{n+1}$ be an oriented C^1-hypersurface and $f : \Gamma \to \mathbb{R}$ a $C^1(\Gamma)$-function. Let $\tilde{f}$ be a C^1-extension of f in an open neighborhood of Γ. On Γ, we define the tangential gradient of f by

$$\nabla_\Gamma f(x) = \nabla \tilde{f}(x) - \nabla \tilde{f}(x) \cdot \nu(x)\nu(x) \quad x \in \Gamma, \tag{4}$$

where ∇ denotes the gradient in $\mathbb{R}^{n+1}$ and ν the outer unit normal on Γ. The components of the tangential gradient are denoted by $\nabla_\Gamma f = (\underline{D_1} f, ..., \underline{D_{n+1}} f)$. Let Γ be an oriented C^2-hypersurface in $\mathbb{R}^{n+1}$. For a real-valued function f, which is twice continuously differentiable in a neighborhood of Γ, we define the Laplace–Beltrami operator by

$$\Delta_\Gamma f(x) := \nabla_\Gamma \cdot \nabla_\Gamma f(x) = \sum_{i=1}^{n+1} \underline{D_i} \underline{D_i} f(x), \quad x \in \Gamma. \tag{5}$$

Theorem 1. *Let Γ be an oriented C^2-hypersurface with C^1-boundary $\partial\Gamma$, whose intrinsic unit outer normal to Γ is denoted by μ and $f \in C^1(\Gamma)$. Then, the formula of integration by parts on Γ is:*

$$\int_\Gamma \nabla_\Gamma f = -\int_\Gamma f H \nu + \int_{\partial\Gamma} f \mu. \tag{6}$$

Here, ν is the continuously differentiable outer unit normal of Γ, and the matrix H denotes the mean curvature of Γ, which is given by

$$H = -\nabla_\Gamma \cdot \nu. \tag{7}$$

Proof. Can be found in [5] for surfaces without boundary.

Definition 2. Let $U(t)$ be an open neighborhood of the hypersurface $\Gamma(t)$ and $U_T := \bigcup_{t\in(0,T)} U(t) \times \{t\}$. The material derivative of a continuously differentiable function $f(x,t)$ defined on U_T is given by:

$$\dot{f} = \frac{\partial f}{\partial t} + \upsilon \cdot \nabla f, \tag{8}$$

where we note that for all $t \in [0,T]$ we have

$$\frac{\partial}{\partial t}\phi(\cdot,t) = \upsilon(\phi(\cdot,t),t), \tag{9}$$

and ν describes the velocity of the surface. Therefore,

$$\dot{f}(\phi(\cdot,t),t) = \frac{d}{dt} f(\phi(\cdot,t),t). \tag{10}$$

Lemma 1. *Let $\Gamma(t)$ be a C^2-hypersurface and f a function defined on the C^2 hypersurface G_T, such that the appearing quantities exist, then we have the so-called Leibniz formula or transport theorem*

$$\frac{d}{dt}\int_{\Gamma_t} f = \int_{\Gamma_t} (\dot{f} + f\nabla_\Gamma \cdot \upsilon). \tag{11}$$

Proof. [3], page 291.

Problem formulation Since a classical solution of problem (1) does not exist in general, we define the class of weak solutions and we introduce the term of an entropy solution to achieve a unique solution.

Definition 3. A vector of functions $\mathbf{u} = (u^1, ..., u^M) \in L^\infty(G_T)^M$ is called weak solution of the weakly coupled hyperbolic initial value problem (1), if for all components of the system $i = 1, ..., M$

$$\int_0^T \int_{\Gamma_t} u^i \dot{\varphi} + f^i(\cdot, u^i) \cdot \nabla_\Gamma \varphi - r^i(\cdot, \mathbf{u})\varphi + \int_{\Gamma_0} u_0^i \varphi(\cdot, 0) = 0 \tag{12}$$

holds for all test functions $\varphi \in C^1(\overline{G_T})$ with $\varphi(\cdot, T) = 0$.

Definition 4. For $l \in \{1, ..., n+1\}$, let the flux functions $\mathbf{f}_l = (f_l^1, ..., f_l^M)$ be given as above, then the tuple $(\eta, \mathbf{q}_1, ..., \mathbf{q}_{n+1})$ is called entropy of $\mathbf{f}_1, ..., \mathbf{f}_{n+1}$, if the real-valued function $\eta \in C^2(\mathbb{R})$ is convex and the function $\mathbf{q}_l = (q_l^1, ..., q_l^M)$, with $q^i = q^i(x, t, s)$ is defined by

$$q_l^i(\cdot, s) := \int_{s_0}^s \eta'(\tau) \partial_u f_l^i(\cdot, \tau) d\tau$$

with $s_0 \in \mathbb{R}$ for all $i = 1, ..., M$ and $l \in \{1, ..., n+1\}$. The function η is called entropy function, and the functions $\mathbf{q}_1, ..., \mathbf{q}_{n+1}$ are called entropy fluxes.

Definition 5. A vector of functions $\mathbf{u} = (u^1, ..., u^M) \in L^\infty(G_T)^M$ is called entropy solution of (1), if for all components $i = 1, ..., M$

$$\int_0^T \int_{\Gamma_t} \left(-\eta(u^i)\dot{\varphi} - q^i(\cdot, u^i) \cdot \nabla_\Gamma \varphi + \varphi \nabla_\Gamma \cdot \nu \left(u^i \eta'(u^i) - \eta(u^i) \right) - r^i(\cdot, \mathbf{u}) \eta'(u^i) \varphi \right)$$
$$- \int_{\Gamma_0} \eta(u_0^i) \varphi(\cdot, 0) \leq 0 \tag{13}$$

holds for all test functions $\varphi \in H^1(G_T)$ with $\varphi \geq 0$ and $\varphi(\cdot, T) = 0$ and all entropies $(\eta, \mathbf{q}_1, ..., \mathbf{q}_{n+1})$ of $\mathbf{f}_1, ..., \mathbf{f}_{n+1}$.

2 The Parabolic System

We are using a classical viscosity approach to get a solution of the initial value problem (1). Therefore, we consider our problem with a viscosity term. In the following, we need a specific diffusive matrix $B = B(x, t)$, but we like to mention that it is possible to substitute it with the identity matrix; see [7]. The approach is subdivided into the following steps: At first, we show that a weak solution of the viscous problem exists and that it is uniformly bounded in $L^\infty(G_T)$. Then, we show that it is a classical solution. At last, we prove that it is uniformly bounded in $W^{1,1}(G_T)$.

Definition 6. Let the tuple $(\mathbf{f}_k, \mathbf{r}, \mathbf{u}_{0\varepsilon})$ with $k = 1, ..., n+1$ be given, such that $\mathbf{u}_{0\varepsilon} \to u_0$ in $L^1(\Gamma_0)$. Then, we call the following initial value problem to the viscosity parameter $\varepsilon \geq 0$ weakly coupled parabolic problem: Find a function $\mathbf{u} : G_T \to \mathbb{R}^M$, $\mathbf{u} = (u^1, .., u^M)$ with

$$\begin{aligned}
\dot{u}^1 \quad &+u^1\nabla_\Gamma\cdot\nu+\nabla_\Gamma\cdot(f^1(\cdot,u^1))=r^1(\cdot,\mathbf{u})+\varepsilon\nabla_\Gamma\cdot(B\nabla_\Gamma u^1),\\
\dot{u}^2 \quad &+u^2\nabla_\Gamma\cdot\nu+\nabla_\Gamma\cdot(f^2(\cdot,u^2))=r^2(\cdot,\mathbf{u})+\varepsilon\nabla_\Gamma\cdot(B\nabla_\Gamma u^2),\\
&\vdots \qquad\qquad\qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\quad \vdots\\
\dot{u}^M \quad &+u^M\nabla_\Gamma\cdot\nu+\nabla_\Gamma\cdot(f^M(\cdot,u^M))=r^M(\cdot,\mathbf{u})+\varepsilon\nabla_\Gamma\cdot(B\nabla_\Gamma u^M)
\end{aligned} \tag{14}$$

on G_T and $\mathbf{u}(x,0)=\mathbf{u}_{0\varepsilon}(x)$ for $x\in\Gamma_0$.

Note that the solution of this problem depends on the parameter ε; however, for the sake of simplicity, we will write $\mathbf{u}$ instead of $\mathbf{u}_\varepsilon$. In contrast to the Euclidean Cauchy problem, in general, there does not exist an explicit formula for solutions of problem (14) on surfaces. Therefore, we define a weak solution.

Lemma 2. *Let the tuple* $(\mathbf{f}_k,\mathbf{r},\mathbf{u}_{0\varepsilon})$, $k=1,...,n+1$ *and the diffusion matrix* B *be given. Let the parameter fulfill* $\varepsilon\geq 0$. *Under the assumptions of the introduction and* $\mathbf{f}$ *being global Lipschitz continuous in* u, *the matrix* B *is smooth, fulfills* $B\nu=\nu^*B=0$ *and maps a tangent vector on a tangent vector again and the initial data fulfills* $u_{0\varepsilon}\in H^1(\Gamma_0)$, *there exists a weak solution* $\mathbf{u}=(u^1,...,u^M)\in H^1(G_T)^M$ *of the weakly coupled parabolic initial value problem; i.e., for all components* $i=1,...,M$, *the following equality holds true*

$$(\dot{u}^i,\varphi)+(u^i,\nabla_{\Gamma(t)}\cdot\nu\varphi)+\varepsilon(B\nabla_{\Gamma(t)}u^i,\nabla_{\Gamma(t)}\varphi)=(r^i(\mathbf{u}),\varphi)-(\nabla_{\Gamma(t)}\cdot f^i(u^i),\varphi) \tag{15}$$

for all $\varphi(\cdot,t)\in H^1(\Gamma_t)$, *for almost all* $t\in[0,T]$, *and we have* $\mathbf{u}(x,0)=\mathbf{u}_{0\varepsilon}$ *for almost all* $x\in\Gamma_0$. *Here,* $(\cdot,\cdot)$ *denotes the scalar product of* $L^2(\Gamma_t)$.

Proof. At first, we decouple and linearize the problem which means we solve the following problem successively for $i=1,...,M$:

$$\begin{aligned}
&(\dot{u}^{i(m)}(\cdot,t),\varphi(\cdot,t))+(u^{i(m)}(\cdot,t),\nabla_{\Gamma(t)}\cdot\nu(\cdot,t)\varphi(\cdot,t))\\
&\qquad+\varepsilon(B(\cdot,t)\nabla_{\Gamma(t)}u^{i(m)}(\cdot,t),\nabla_{\Gamma(t)}\varphi(\cdot,t))\\
&=(r^i(\cdot,t,\mathbf{u}^{(m-1)}(\cdot,t)),\varphi(\cdot,t))-(\nabla_{\Gamma(t)}\cdot f^i(\cdot,t,u^{i(m-1)}(\cdot,t)),\varphi(\cdot,t)),
\end{aligned}$$

with $u^{i(m)}(x,0)=u^i_{0\varepsilon}$ on Γ_0 and $u^{i(0)}=0$. Such a solution exists in $H^1(G_T)$ due to [3]. They used a Galerkin approximation. Then we show, using energy estimates and the Lipschitz continuity of $\mathbf{f}$ and $\mathbf{r}$ in u, that the sequence $(u^{i(m)})_{m\in\mathbb{N}}$ converges to the weak solution in $H^1(G_T)$ of the nonlinear and coupled parabolic problem (14). □

Lemma 3. *Let the tuple* $(\mathbf{f}_k,\mathbf{r},\mathbf{u}_{0\varepsilon})$, $k=1,...,n+1$ *and the diffusion matrix* B, *which fulfills assumptions as in Lemma 2, be given. Let the parameter fulfill* $\varepsilon\geq 0$. *Under the assumptions from the introduction and the coupling term* $\mathbf{r}(x,t,\mathbf{u})=(r^1(x,t,\mathbf{u}),...,r^M(x,t,\mathbf{u}))$ *being quasi-monotone increasing in* $\mathbf{u}$, *which means that for every* $s\in 1,..,M$

$$r_s(x,t,\mathbf{u})\leq r_s(x,t,\bar{\mathbf{u}})\qquad \text{for}\quad u^i\leq\bar{u}^i,\quad i\neq s,\quad u^s=\bar{u}^s,$$

$u_{0\varepsilon} \in H^1(\Gamma_0)$ and with $\|u_{0\varepsilon}\|_{L^\infty(G_T)} < K$, where K is assumed to be independent of ε, the parabolic initial value problem (14) has a weak solution $\mathbf{u}_\varepsilon \in H^1(G_T)^M$, which fulfills

$$\|\mathbf{u}_\varepsilon\|_{L^\infty(G_T)} \leq C \tag{16}$$

with a constant C, which is independent of the viscosity parameter ε.

Proof. We rewrite the weak form such that we can solve the problem in

$$w^i(x,t) = e^{-\lambda t}\left(u^i(x,t) - \frac{\eta}{\lambda}\right) - \frac{\eta}{\lambda}$$

having two free parameters η, λ, use as the test function the positive part of the new unknown function w^i minus a constant, and apply Gronwall's inequality.

Note that we were able to omit the global Lipschitz continuity of $\mathbf{f}$ in u and that this proof is also applicable for $r^i < 0$ for all $i = 1, ..., M$. In [7], we showed that the quasi-monotonicity condition can be omitted, by first showing ε dependent $L^\infty(G_T)$ bounds using the Euclidean theory for quasi-linear systems of [8], proving regularity results and following the Euclidean proof for uniform boundedness, where we use the existence of the regularized solutions and transfer the Euclidean proof of C. Rohde in [16] to moving surfaces.

Lemma 4. *Let the assumptions of Lemma 3 hold true, and additionally, let $f^i \in C^k(G_T \times \mathbb{R})$, $r^i \in C^k(G_T \times \mathbb{R}^M)$, and $u^i_{0\varepsilon} \in H^k(\Gamma_0) \cap L^\infty(\Gamma_0)$ for $i = 1, ..., M$ with $k > \max\{\frac{n}{2} + 2, 3\}$. The parametrization ψ of the surface G_T is assumed to be in $C^\infty(\mathbb{R}^n \times [0,T])$. Then, the weak solution $\mathbf{u}_\varepsilon$ of the weakly coupled parabolic problem (14) is a solution in the classical sense.*

Proof. We use a localization function α on a map area U and follow a classical idea that can be found in [13], for which it is necessary, that the solution is bounded:

$$\int_U (\alpha u)^{\cdot}\varphi + \varepsilon B \nabla_{\Gamma(t)}(u\alpha)\cdot\nabla_{\Gamma(t)}\varphi + \alpha u \nabla_{\Gamma(t)}\cdot\nu\varphi = \int_U \nabla_{\Gamma(t)}\cdot f(u)\alpha\varphi + r(\mathbf{u})\alpha\varphi$$
$$- \int_U \varepsilon(\nabla_{\Gamma(t)}\cdot(B\nabla_{\Gamma(t)}\alpha)u\varphi + 2\varphi B\nabla_{\Gamma(t)}u\cdot\nabla_{\Gamma(t)}\alpha).$$

Due to the regularity of the right hand side, using results of parabolic problems with time dependent coefficients, we obtain higher regularity in the unknown function. This leads to higher regularity of the right hand side again. We reuse this argument until we get with an embedding theorem a classical solution. □

Lemma 5. *Let the assumptions of Lemma 4 with* $k > \max\{\frac{n}{2} + 2, 3\}$ *hold true, and let the initial data* $u_{0\varepsilon}$ *satisfy*

$$\|u_{0\varepsilon}\|_{L^\infty(\Gamma_0)} + \|\nabla_\Gamma u_{0\varepsilon}\|_{L^1(\Gamma_0)} + \varepsilon\|\nabla_\Gamma^2 u_{0\varepsilon}\|_{L^1(\Gamma_0)} \le C. \tag{17}$$

Let $\mathbf{u}_\varepsilon$ *be the classical solution of the parabolic initial value problem (14). Then, we have*

$$\sum_{i=1}^{M} \sup_{(0,T)} \int_{\Gamma_t} |\nabla_\Gamma u_\varepsilon^i| \le C, \tag{18}$$

where the constant C *is independent of the parameter* ε*. Now let the diffusion matrix* $B(x,t)$ *satisfy the following initial value problem*

$$\dot{B} = BA(\nu) + A(\nu)^* B + \lambda B, \quad B(\cdot,0) = B_0, \tag{19}$$

where $\lambda > 0$ *is a constant, the matrix* $A(\nu) = A(\nu)_{l,r} := \underline{D}_l \nu_r - \nu \cdot \nu_l \underline{D}_r \nu$*, and* B_0 *is a symmetric, tangentially positive definite* $(n+1) \times (n+1)$ *matrix. Then,*

$$\sum_{i=1}^{M} \sup_{(0,T)} \int_{\Gamma_t} |\dot{u}^i| \le C, \tag{20}$$

where C *is independent of* ε*.*

Proof. According to Lemma 7.1., p. 215 in [4], there exists a symmetric and positive definite matrix B, which fulfills this initial value problem and which maps the tangential space onto it self, such that $B\nu = \nu^* B = 0$ holds. The proof is analog to the one of the scalar cases in [4] since the coupling term only depends on the unknown function itself and not on the derivatives. In both estimates, the authors differentiated the regularized problem with respect to one component of the tangential gradient or the material derivative, then multiplied the equation with the normalized component of the tangential gradient or with the signum of the material derivative of the regularized solution, integrated over the space surface, and used some technical estimates. Finally, they used Gronwall's estimate. The PDE (19) for the diffusive matrix B is needed since derivatives on moving surfaces do not commute. In [7], we show with a compactness argument of Dafermos that this condition is not needed and the matrix B can be replaced by the identity matrix.

3 Existence of an Entropy Solution

In this section, we put all previous results together to arrive at the existence of an entropy solution of weakly coupled system of conservation laws on moving surfaces (1).

Theorem 2. *Let the assumptions of Lemma 4 with* $k > \max\{\frac{n}{2} + 2, 3\}$ *be satisfied. Then, there exists a subsequence* $\mathbf{u}_\varepsilon$ *of classical solutions of the regularized problem (14) with diffusive matrix B satisfying (19) and initial data* $u_{0\varepsilon}$ *satisfying (17), that converges in* $L^1(G_T)$ *and almost everywhere to a function* $\mathbf{u}$*, which is an entropy solution of the hyperbolic initial value problem of weakly coupled system of conservation laws (1).*

Proof. We use the result of Lemma 5 and the Theorem of Kondrakov to get a subsequence of solutions $\mathbf{u}_\varepsilon$ that converges for $\varepsilon \to 0$ in $L^1(G_T)$ and a.e. to a function $\mathbf{u}$. Then, we use the weak form of the parabolic system to show that this limit function $\mathbf{u}$ is a entropy solution of problem (1).

More details on this topics and further results including a different approach to compactness, studying the Euclidean proof of Dafermos and the uniqueness of the entropy solution, which can be shown with a localization argument and using the classical ideas of Kruzkov, can be found in [7].

References

1. P. Amorim, M. Ben-Artzi, P.G. LeFloch et al., Hyperbolic conservation laws on manifolds: total variation estimates and the finite volume method. Methods Appl. Anal. **12**(3), 291–324 (2005)
2. M. Ben-Artzi, P. G. Le Floch, Well-posedness theory for geometry-compatible hyperbolic conservation laws on manifolds. Ann. Inst. H. Poincaré Anal. Nonlinéaire **24**(6), 989–1008 (2007)
3. G. Dziuk, C.M. Elliott, Finite elements on evolving surfaces. IMA J. Numer. Anal. **27**(2), 262–292 (2007)
4. G. Dziuk, D. Kröner, T. Müller, Scalar conservation laws on moving hypersurfaces. Interfaces Free Boundaries **15**, 202–236 (2013)
5. D. Gilbarg, N. Trudinger, *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics (Springer, Berlin, 2015)
6. T. Hillen, C. Rohde, F. Lutscher, Existence of weak solutions for a hyperbolic model of chemosensitive movement. J. Math. Anal. Appl. **260**(1), 173–199 (2001)
7. A. Korsch, *Weakly Coupled Systems of Conservation Laws on Moving Surfaces*. Ph.D. thesis, Mathematische Fakultät der Albert-Ludwigs-Universität Freiburg (2016)
8. O. Ladyženskaja, V. Solonnikov, N. Ural'ceva, *Linear and Quasilinear Equations of Parablic Type*, vol. 23. Translations of Mathematical Monographs, Rhode Island (1968)
9. A. Levy, On Majda's model for dynamic combustion. Commun. Partial Differ. Equ. **17**(3–4), 657–698 (1992)
10. D. Lengeler, T. Müller, Scalar conservation laws on constant and time-dependent riemannian manifolds. J. Differ. Equ. **254**(4), 1705–1727 (2013)
11. A. Majda, A qualitative model for dynamic combustion. SIAM J. Appl. Math. **41**(1), 70–93 (1981)
12. T. Müller, *Scalar Conservation Laws on Time-Dependent Riemannian Manifolds*. Ph.D. thesis, Albert-Ludwigs-Universität Freiburg (2014)
13. J. Malek, J. Necas, M. Rokyta, M. Ruzicka. *Weak and measure-valued solutions to evolutionary PDEs*, volume 13 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall (1996)

14. M. Ohlberger, C. Rohde, Adaptive finite volume approximations for weakly coupled convection dominated parabolic systems. IMA J. Numer. Anal. **22**(2), 253–280 (2002)
15. E.Y. Panov, On the Dirichlet problem for first order quasilinear equations on a manifold. Trans. Am. Math. Soc. **363**, 2393–2446 (2011)
16. C. Rohde, *Weakly Coupled Systems of Hyperbolic Conservation Laws*. Ph.D. thesis, Mathematische Fakultät der Albert-Ludwigs-Universität Freiburg (1996)
17. C. Rohde, Entropy solutions for weakly coupled hyperbolic systems in several space dimensions. Zeitschrift für angewandte Mathematik und Physik ZAMP **49**(3), 470–499 (1998)
18. C. Rohde, W.-A. Yong, The nonrelativistic limit in radiation hydrodynamics: I. Weak entropy solutions for a model problem. J. Differ. Equ. **234**(1), 91–109 (2007)

A Phase-Field Model for Flows with Phase Transition

Mirko Kränkel and Dietmar Kröner

Abstract There are many mathematical models for describing compressible or incompressible flows with phase transition. In this contribution, we will focus on the Navier–Stokes–Korteweg model [12] (Appl Math Comput 272, part 2, 309–335, 2016) and a phase-field model: The compressible Navier–Stokes–Allen–Cahn model (NSAC) is able to model compressible two-phase flows including surface tension effects and phase transitions. In this contribution, we will present a discontinuous Galerkin scheme for the NSAC model. The scheme is designed to fulfill a discrete version of the free energy inequality, which is the second law of thermodynamics in the isothermal case. For situations near the thermodynamic equilibrium, this property suppresses so-called parasitic currents, which are unphysical velocity fields near the phase boundary.

Keywords Phase-field model · Phase transition · Two-phase flow · Energy inequality

1 Introduction

First, let us consider the Navier–Stokes–Korteweg model, which is given by

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\rho v) &= 0 \\ \partial_t (\rho v) + \nabla \cdot (\rho v \otimes v + p(\rho) I) &= \varepsilon \alpha \Delta v + \gamma \varepsilon^2 \rho \nabla \Delta \rho \quad \text{in} \quad \Omega \times [0, T] \end{aligned} \tag{1}$$

M. Kränkel · D. Kröner (✉)
Abteilung für angewandte Mathematik, University of Freiburg,
Hermann-Herder-Str 10, 79104 Freiburg, Germany
e-mail: dietmar.kroener@mathematik.uni-freiburg.de

M. Kränkel
e-mail: mirko.kraenkel@gmail.com

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_19

with additional boundary and initial conditions and $\Omega \subset \mathbb{R}^2$. Here, ρ, v, $p(\rho)$ are the density, the velocity, and the pressure of the fluid, respectively. The magnitude of $\rho(x,t)$ decides about the current phase state in x at time t. This model is similar to the compressible Navier–Stokes system with two main differences. The pressure $p(\rho)$ is not a monotone function of the density ρ, and there is the additional term $\gamma\varepsilon^2\rho\nabla\Delta\rho$ in the momentum equation (1). There are several theoretical results concerning existence and uniqueness for this problem [7–9, 17]. In order to see the background for this model, let us consider for a moment the static situation, which is characterized by $v = 0$ and ρ, v, p are time independent. In this, case (1) reduces to

$$\nabla p(\rho) = \gamma\varepsilon^2\rho\nabla\Delta\rho \quad \text{in} \quad \Omega \times [0,T]. \tag{2}$$

This equation is related to the Euler–Lagrange equation for the following minimum problem: Minimize the functional

$$\int_\Omega W(\rho) + \varepsilon^2|\nabla\rho|dx \tag{3}$$

under the constraint

$$\int_\Omega \rho(x)dx = constant, \tag{4}$$

where W is a double-well potential and the relation between $p(\rho)$ and W is given by thermodynamics as $p'(\rho) = \rho W''(\rho)$. Using this in (2), we obtain $\rho W''(\rho)\nabla\rho = \gamma\varepsilon^2\rho\nabla\Delta\rho$ or after some simple calculations

$$\nabla W'(\rho) = \gamma\varepsilon^2\nabla\Delta\rho \quad \text{or} \quad W'(\rho) = \gamma\varepsilon^2\Delta\rho + c_0(\varepsilon). \tag{5}$$

For this problem, the following can be proved [13]:

Theorem 1. *Let ρ_{ε_k} be a sequence of global minimizer of the variational problem (3), (4). Then, there is a subsequence ρ_{ε_k} and ρ_0 in $L^1(\Omega)$ such that ρ_{ε_k} converges to ρ_0 in $L^1(\Omega)$ for $k \to \infty$ and the image of ρ_0 consists of two values α, β. The boundary of $A := \{x \in \Omega | \rho_0(x) = \alpha\}$ is considered as the phase boundary. Let p_l, p_v denote the pressure on both sides of the phase boundary. Then, we have*

$$p_l - p_v = (n-1)c_1 k_m \varepsilon_k + 0(\varepsilon_k), \tag{6}$$

where n is the space dimension, c_1 a constant and k_m the constant mean curvature of the (reduced) boundary of A. In particular in the limit $\varepsilon_k \to 0 : p_l - p_v = 0$, i.e., the pressure is continuous across the phase boundary.

This is in some sense a contradiction to classical physical results which indicate that the pressure jump $p_l - p_v$ across the interface for two different fluids in the static case is proportional to the mean curvature [21].

Similar results as for (5) can be obtained for (1), but these results are not rigorous; they are based on the assumptions that there exist formal asymptotic expansions for ρ, v, p [5]. In [10], Daube could avoid these assumptions but instead of that he has to impose some other strong, but more general, assumptions on the solution of (1). A different approach is considered in [16]. In this paper, the authors consider a different scaling of (1), i.e.,

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\rho v) &= 0 \\ \partial_t (\rho v) + \nabla \cdot (\rho v \otimes v + p(\rho) I) &= \Delta v + \lambda \rho \nabla \Delta \rho. \end{aligned}$$

Using the assumption that there is an expansion of the pressure of the form $p = p_0 + M p_1 + M^2 p_2 + M^3 p_3 + \ldots$ and $\lambda = M^4$, where M is the Mach number, they obtain in the limit for $M \to 0$ in each phase:

$$\partial_t (\rho v) + \nabla \cdot (\rho v \otimes v + p_2 I) = \Delta v \quad \text{in} \quad \Omega \times [0, T], \tag{7}$$

and across the interface

$$p_l - p_v = const\ k_m M^2 \quad \text{and} \quad p_{2_l} - p_{2_v} = const\ k_m. \tag{8}$$

In this case, the pressure jump $p_{2_l} - p_{2_v}$ is consistent with [21]. Notice that the pressure p_2 is the pressure which appears in the remaining equation in the single phases; see (7).

In the case of Theorem 1, it turns out that across the interface

$$p_l - p_v = c k_m \varepsilon \quad \text{and} \quad \sigma = c_0 \sqrt{\gamma} \varepsilon, \tag{9}$$

where σ is the surface tension. Now, we should notice that ε is proportional to the thickness of the interface layer. In order to get an acceptable result for the resolution of the interface in numerical simulation, we need a small ε (i.e., a small width of the interface layer) and many cells in this layer.

As the interfacial layer has to be resolved sufficiently enough, e.g., 6–10 grid points/cells in [22], to guarantee a stable discretization, this poses server limitations on physical size of the computational domain, even when adaptive mesh refinement around the interface is used. The length scale which was achievable in [22] was several magnitudes below the size of physical experiments.

Now, if we want to increase the resolution, i.e., we take a smaller ε, we will also change physical quantities: the pressure jump and the surface tension due to (9). But it is not convenient that the physical quantities depend on the numerical solution. Therefore, we have to consider a different model, e.g., a phase-field model as described in the following section.

2 A Phase-Field Model

In this section, we will investigate a phase-field model of the following form (see also [1, 2, 24, 25])

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\rho v) &= 0 \\ \partial_t (\rho v) + \nabla \cdot (\rho v \otimes v + \mathbb{T}) &= \rho G \\ \rho(\partial_t \phi + v \cdot \nabla \phi) &= \tau \end{aligned} \tag{10}$$

in $\Omega \times [0, T]$, where

$$\begin{aligned} \rho &:= \text{ density, } v := \text{ velocity, } \phi := \text{ phasefield function, } \mathbb{T} := \mathbb{P} - \mathbb{D}, \\ \mathbb{P} &:= \mathbb{P}(\rho, \phi, \nabla, \phi) = p\mathbb{I} + \frac{\partial F}{\partial \nabla \phi} \otimes \nabla \phi, \\ p &:= \text{ thermodynamic pressure, } G := \text{ outer force,} \\ F &:= F(\rho, \phi, \nabla, \phi) = h_1(\rho) W(\phi) + \psi(\rho, \phi) + h_2(\rho) \frac{|\nabla \phi|^2}{2} \\ &\quad \text{(free energy density, see also [1, 2, 24, 25]),} \\ \psi &:= \psi(\rho, \phi) = \nu(\phi) F_1(\phi) + (1 - \phi) F_2(\phi), \\ &\quad F_1, F_2 \text{ are the free energies of the bulk phases,} \\ \nu(\phi) &:= \text{ monotone interpolation function,} \\ \lambda(\rho) &:= \text{ small parameter, which may depend on } \rho, \\ \mathbb{D} &:= \mu_1(\rho, \phi)(Dv + (Dv)^T) + \mu_2(\rho, \phi) \nabla \cdot v \mathbb{I}, \\ \mu_i &:= \text{ are chosen such that } \mathbb{D}(Dv) : Dv \geq 0 \text{ for all smooth vector fields v,} \\ &\quad \text{e.g., } \mu_1 > 0, \mu_2 > -\frac{2}{d}\mu_1, \\ \tau &:= -\eta \frac{\delta F}{\delta \phi}, \eta = \text{ reaction rate, where } \delta \text{ denotes the variational derivative.} \end{aligned}$$

Here, ϕ is the phase-field function with values in [0, 1], which allows us to distinguish between the two phases. Values $\phi = 0$ indicate that we are in one phase and values $\phi = 1$ that we are in the other phase.

This model goes back to [1, 2, 24, 25]. Similar models have been considered in [6, 14, 22]. In [18], the existence of strong solutions of (10) was shown.

Now, it is important to know that the system (10) satisfies the second law of thermodynamics. For the isothermal case, which we consider here, this means that an energy inequality holds. In particular, this property is important for developing a numerical scheme. Also, the numerical solution should satisfy a discrete energy inequality. The following theorem is due to [3, 24]; see also [19].

The total energy $\mathscr{E}(\rho, \phi, v, \nabla\phi)$ is given by the integral over Ω of the sum of the Helmholtz free energy $F(\rho, \phi) = \psi(\rho, \phi) + h_2(\rho)\frac{|\nabla\phi|^2}{2}$ and the kinetic energy $\rho\frac{|v|^2}{2}$:

$$\begin{aligned}\mathscr{E}(\rho, \phi, v, \nabla\phi) &= \int_\Omega E(\rho, \phi, v, \nabla\phi) = \int_\Omega F(\rho, \phi) + \rho\frac{|v|^2}{2}dx \\ &= \int_\Omega \psi(\rho, \phi) + h_2(\rho)\frac{|\nabla\phi|^2}{2} + \rho\frac{|v|^2}{2}dx.\end{aligned}$$

The following energy inequality holds [19]:

Theorem 2. *Let (ρ, v, p, ϕ) be a smooth solution of (10). Then, the total energy $\mathscr{E}$ is non-decreasing in time and the following inequality holds:*

$$\frac{\partial\mathscr{E}}{\partial t} \leq -\frac{\eta}{\rho}||\frac{\delta f}{\delta\phi}||_2^2 - C||\nabla v||_2^2 \leq 0. \tag{11}$$

3 Non-conservative Mixed Form

Since the exact solution satisfies an energy inequality (11), it turns that it is important to have a numerical scheme such that the numerical solution $(\rho_h, v_h, \phi_h, \sigma_h)$ also satisfies a discrete energy inequality of the form

$$E(\rho_{\cdot}^n v_{\cdot}^n \phi_h^n, \sigma_h^n) \leq E(\rho_h^{n+1}, v_h^{n+1}, \phi_h^{n+1}, \sigma_h^{n+1}), \tag{12}$$

where $(\rho_{\cdot}^n v_{\cdot}^n \phi_h^n, \sigma_h^n)$ denotes the numerical solution at time n. To devise a numerical scheme fulfilling (12), we use the ideas from [15], where such a scheme for the Navier–Stokes–Korteweg model was derived. In the proof of the energy inequality (11), it was necessary to use the mass balance equation multiplied with the quantity $\frac{\partial F}{\partial\rho} + \frac{|v|^2}{2}$, the momentum balanced multiplied with v, and the phase-field equation multiplied with $\frac{\delta F}{\delta\phi}$. As these quantities depend nonlinearly on the variables $(\rho, \rho v, \phi)$, we cannot use this quantity as test functions in a numerical formulation of system (10) directly, and they will not belong to the piecewise polynomial discontinuous Galerkin space, used for the test and trial functions. Furthermore, we had to compute explicitly the divergence of the pressure tensor P, to relate the resulting terms to $\frac{\partial F}{\partial\rho}$ and $\frac{\delta F}{\delta\phi}$. Therefore, we rewrite the system in a mixed form, so that terms that will be needed as test functions appear in the formulation and can be discretized in the discontinuous Galerkin space and the divergence of the pressure tensor appears in its explicit form. Using

$$\nabla\cdot\mathbb{P} = \rho\nabla(\frac{\partial F}{\partial\rho}) - \nabla\phi\frac{\delta f}{\delta\phi} \text{ and } \mu := -F + \frac{\partial F}{\partial\rho},\ \tau := \frac{\delta F}{\delta\phi}$$

we can rewrite (10) as

$$\begin{aligned} \partial_t \rho + \nabla \cdot \rho v &= 0 \\ \rho \partial_t v + \nabla \cdot (\rho v \otimes v) - \nabla \cdot (\rho v) v + \rho \nabla \mu & \\ -\tau \nabla \phi - \frac{1}{2} \rho \nabla |v|^2 - \nabla \cdot \mathbb{D}(\nabla v) &= 0 \\ \partial_t \phi + \nabla \phi \cdot v + \frac{\tau}{\rho} &= 0 \\ \tau - \frac{\partial F}{\partial \phi} + \nabla \cdot (h_2 \sigma) &= 0 \\ \mu - \frac{\partial F}{\partial \rho} - \frac{1}{2} |v|^2 &= 0 \\ \sigma - \nabla \phi &= 0. \end{aligned} \tag{13}$$

3.1 Discretization

First, let us fix some notations. Let $\Omega \subset \mathbb{R}^d$ be a polygonal domain, and let $\mathcal{T}$ be a computational grid such that $\bar{\Omega} = \cup_{E \in \mathcal{T}} E$ where E are the cells of the triangulation. The maximal diameter of all cells is $h := sup_{E \in \mathcal{T}} diam(E)$. Let $n_E, n_{E'}$ be the outer normals to the edge $e := \partial E \cap \partial E'$, respectively, and let ϕ be a function which is smooth on E and E', but might be discontinuous across e. Then, the inner and outer traces $\phi^+(x)$ and $\phi^-(x)$ for $x \in e$ are given by

$$\phi^+(x) := lim_{\varepsilon \to 0} \phi(x + \varepsilon n_E), \;\; \phi^-(x) := lim_{\varepsilon \to 0} \phi(x + \varepsilon n_{E'}).$$

If $x \in \partial\Omega$, then for Dirichlet data g we define $\phi^-(x) := g(x)$. The discontinuous Galerkin space is defined as

$$\mathcal{V}_h := \{u \in L^2(\Omega) : u|_E \in \mathbb{P}_k \text{ for all } E \in \mathcal{T}\},$$

where $\mathbb{P}_k$ is the space of polynomials of degree $\leq k$. The mean value $\{\{\phi\}\}$ of a piecewise smooth function ϕ on the edge $e := \partial E \cap \partial E'$ is defined by

$$\{\{\phi\}\} := \frac{1}{2}(\phi^+ + \phi^-) \text{ and the jump } [[\phi]] \text{ by } [[\phi]] := \phi^+ n_E + \phi^- n_{E'}.$$

The set of all intersections of $\mathcal{T}$ is denoted as:

$$\Gamma := \bigcup_{E \in \mathcal{T}} \bigcup_{e \subset \partial E} e.$$

To devise a numerical scheme which is energy consistent, in the sense that a discrete version of the energy inequality is satisfied, we impose the equations of system (13) on each element by multiplying with test functions from the discontinuous Galerkin

space and then integrating over the element. Then, we choose appropriate numerical fluxes to couple the degrees of freedom on each element. Note that for this scheme, no partial integration is performed and the scheme can be seen as a weighted residual method and the numerical fluxes are used to impose a weak form of continuity over the element boundaries.

For the time discrete scheme, we use the following notation. We start by subdividing the time interval $[0, T]$ by a sequence of time steps $t_0 = 0 < t_1 < t_2 < \cdots < t_N = T$. The backward difference quotient for a time-dependent function u is denoted by $D_t u^{n+1} := \frac{u(\cdot,t_{n+1}) - u(\cdot,t_n)}{\Delta t^n}$, and for the average of two consecutive time steps, we write $u^{n+\frac{1}{2}} := \frac{u(\cdot,t_{n+1}) + u(\cdot,t_n)}{2}$.

The fully discrete discontinuous Galerkin scheme for (10) reads:

Problem 1. For initial values ρ^0, v^0, ϕ^0 and for all $n = 1 \dots N$ find

$$(\rho_h^n, v_h^n, \phi_h^n, \mu_h^n, \tau_h^n, \sigma_h^n) \in \mathscr{V}_h \times \mathscr{V}_h^d \times \mathscr{V}_h \times \mathscr{V}_h \times \mathscr{V}_h \times \mathscr{V}_h^d$$

such that:

$$\begin{aligned}
0 &= \sum_{E\in\mathscr{T}} \int_E (D_t \rho^{n+1} + \nabla \cdot (\rho^{n+\frac{1}{2}} v^{n+\frac{1}{2}}))\psi \, dx \qquad (14)\\
&+ \int_\Gamma G_1(\rho^{n+\frac{1}{2}}, \phi^{n+\frac{1}{2}}, v^{n+\frac{1}{2}}, \mu^{n+\frac{1}{2}}, \tau^{n+\frac{1}{2}}, \sigma^{n+\frac{1}{2}}; \psi) ds\\
0 &= \sum_{E\in\mathscr{T}} \int_\Omega \Big(\rho^{n+\frac{1}{2}} \partial_t v + \nabla \cdot (\rho^{n+\frac{1}{2}} v^{n+\frac{1}{2}} \otimes v^{n+\frac{1}{2}}) - \nabla \cdot (\rho^{n+\frac{1}{2}} v^{n+\frac{1}{2}}) v^{n+\frac{1}{2}}\\
&+ \rho^{n+\frac{1}{2}} \nabla \mu^{n+\frac{1}{2}} - \tau^{n+\frac{1}{2}} \nabla \phi^{n+\frac{1}{2}} - \frac{1}{2} \rho^{n+\frac{1}{2}} \nabla |v^{n+\frac{1}{2}}|^2 \Big) \chi \, dx\\
&+ \int_\Gamma G_2(\rho^{n+\frac{1}{2}}, \phi^{n+\frac{1}{2}}, v^{n+\frac{1}{2}}, \mu^{n+\frac{1}{2}}, \tau^{n+\frac{1}{2}}, \sigma^{n+\frac{1}{2}}; \chi) ds + \mathbb{B}(v^{n+\frac{1}{2}}, \chi^{n+\frac{1}{2}})\\
0 &= \sum_{E\in\mathscr{T}} \int_E \left(D_t \phi^{n+1} + \nabla \phi^{n+\frac{1}{2}} \cdot v^{n+\frac{1}{2}} + \frac{\tau^{n+\frac{1}{2}}}{\rho^{n+\frac{1}{2}}} \right) \theta \, dx\\
&+ \int_\Gamma G_{3.1}(\rho^{n+\frac{1}{2}}, \phi^{n+\frac{1}{2}}, v^{n+\frac{1}{2}}, \mu^{n+\frac{1}{2}}, \tau^{n+\frac{1}{2}}, \sigma^{n+\frac{1}{2}}; \theta) ds\\
0 &= \sum_{E\in\mathscr{T}} \int_E \left(\tau^{n+\frac{1}{2}} - \frac{F(\rho^{n+1}, \phi^{n+1}) - F(\rho^{n+1}, \phi^n)}{\phi^{n+1} - \phi^n} + \nabla \cdot \sigma^{n+\frac{1}{2}} \right) \zeta \, dx \qquad (15)\\
&+ \int_\Gamma G_{3.2}(\rho^{n+\frac{1}{2}}, \phi^{n+\frac{1}{2}}, v^{n+\frac{1}{2}}, \mu^{n+\frac{1}{2}}, \tau^{n+\frac{1}{2}}, \sigma^{n+\frac{1}{2}}; \zeta) ds\\
0 &= \sum_{E\in\mathscr{T}} \int_E \left(\mu^{n+\frac{1}{2}} - \frac{F(\rho^{n+1}, \phi^n) - F(\rho^n, \phi^n)}{\rho^{n+1} - \rho^n} - \frac{1}{4}(|v^{n+1}|^2 + |v^n|^2) \right) \eta \, dx\\
0 &= \sum_{E\in\mathscr{T}} \int_E \left(\sigma^{n+1} - \nabla \phi^{n+1} \right) \xi \, dx\\
&+ \int_\Gamma G_4(\rho^{n+1}, \phi^{n+1}, v^{n+1}, \mu^{n+1}, \tau^{n+1}, \sigma^{n+1}; \xi) ds
\end{aligned}$$

for all $(\psi, \chi, \theta, \zeta, \eta, \xi) \in \mathscr{V}_h \times \mathscr{V}_h^d \times \mathscr{V}_h \times \mathscr{V}_h \times \mathscr{V}_h \times \mathscr{V}_h^d$, where the numerical fluxes are given as

$$\begin{aligned}
G_1(\rho, \phi, v, \mu, \tau, \sigma; \psi) &= -[[\rho v]]\{\{\psi\}\} + \alpha[[\mu]]\{\{\psi\}\},\\
G_2(\rho, \phi, v, \mu, \tau, \sigma; \chi) &= F_{2.1}(\rho, \mu; \chi) + F_{2.2}(\phi, \tau; \chi),\\
G_{2.1}(\rho, \mu; \chi) &= -[[\mu]]\{\{\rho\chi\}\},\\
G_{2.2}(\phi, \tau; \chi) &= [[\phi]]\{\{\tau\chi\}\},\\
G_{3.1}(\rho, \phi, v, \mu, \tau, \sigma; \theta) &= -[[\phi]]\{\{\theta v\}\},\\
G_{3.2}(\rho, \phi, v, \mu, \tau, \sigma; \zeta) &= -h_2[[\sigma]]\{\{\zeta\}\},\\
G_4(\rho, \phi, v, \mu, \tau, \sigma; \xi) &= [[\phi]]\{\{\xi\}\}
\end{aligned}$$

with $\alpha > 0$. The diffusion part of the momentum equation is discretized by the interior penalty bilinear form

$$\mathbb{B}(v, \chi) := \sum_{E\in\mathscr{T}} \int_E \mathbb{D}(\nabla v)\nabla\chi dx - \sum_{e\in\Gamma} \int_e \{\{\mathbb{D}(\nabla v)\}\}[[\chi]] + \{\{\mathbb{D}(\nabla\chi)\}\}[[v]] + \int_\Gamma \frac{\beta}{|e|}[[v]][[\chi]]ds$$

with $\beta > 0$ sufficiently large; see [4].

Combining the space and time discretization, one can show a discrete version of the energy inequality (11) [19]:

Theorem 3. *The discrete solution $(\rho_h^n, \phi_h^n, \sigma_h^n)$ of scheme (1) can be shown to fulfill the following discrete energy equation*

$$\begin{aligned}
&\sum_{E\in\mathscr{T}} \int_E F(\rho_h^{n+1}, \phi_h^{n+1}) + h_2\frac{|\sigma_h^{n+1}|^2}{2} + \rho_h^{n+1}\frac{|v_h^{n+1}|^2}{2}dx\\
&- \int_E F(\rho_h^n, \phi_h^n) + h_2\frac{|\sigma_h^n|^2}{2} + \rho_h^n\frac{|v_h^n|^2}{2}dx = -\frac{(\tau_h^{n-\frac{1}{2}})^2}{\rho_h^{n-\frac{1}{2}}} - \mathbb{B}(\nabla v_h^{n\frac{1}{2}}, \nabla v_h^{n-\frac{1}{2}}).
\end{aligned} \tag{16}$$

4 Numerical Examples

For the numerical examples, we set $h_1 = \frac{A}{\delta}$ and $h_2 = A\delta$. Thus, we arrive at a free energy of the form

$$F(\rho, \phi, \nabla\phi) = \frac{A}{\delta}W(\phi) + \nu(\phi)(F_1(\rho) + (1 - \nu(\phi))F_2(\rho) + A\delta\frac{|\nabla\phi|^2}{2}, \tag{17}$$

where the double-well potential is chosen as $W(\phi) = \phi^2(\phi - 1)^2$ and the free energy of the bulk phases is of the stiffened gas form $F_i = a_i \rho \log(\rho) + (b_i - a_i)\rho + c_i$. The parameters A and δ will be chosen accordingly in the numerical examples.

4.1 Convergence Test

In the first example, we test the convergence properties of our scheme in one space dimension. Therefore, we use the manufactured solution

$$\begin{aligned} \phi_{exact}(x,t) &= \frac{1}{2}\cos(5\pi t)\cos(2\pi x) + 0.5, \\ v_{exact}(x,t) &= \cos(5\pi t)\cos(4\pi x), \\ \rho_{exact}(x,t) &= \frac{1}{2}\cos(5\pi t)\cos(2\pi x) + 1.5 \end{aligned} \tag{18}$$

and compute source terms $S_\phi(x,t)$, $S_v(x,t)$, $S_\rho(x,t)$ for system (10), so that $(\phi_{exact}, v_{exact}, \rho_{exact})^t$ is an exact solution. For this, we used the Sympy python library [23] for symbolical calculations computing the partial derivatives und nonlinearities occurring in (10). The source terms are discretized by using the L^2-projection of the terms on the discontinuous Galerkin space.

The bulk energies are chosen as

$$F_1 = 1.5\rho\log(\rho) + (\log(2) - 1.5)\rho \text{ and } F_2 := \rho\log(\rho) - \rho + 0.5. \tag{19}$$

For the other parameters, we choose $\nu = 10^{-3}$ for the viscosity, $\delta = 0.05$, $A = 5$ for the scaling parameters of the free energy, and $\eta = 1$ for the reaction rate. For the computations, a time step of $\Delta t = 10^{-5}$ and periodic boundary conditions were used. We ran the simulation up to $T = 0.025$ on grids with 16, 32, 64, 128, 256 elements and studied the experimental order of convergence. The results in Tables 1 and 2 indicate that the scheme converges with order $k + 1$ in space, where k is the degree of the polynomials of the discontinuous Galerkin space.

Table 1 Polynomial degree 1

Size	$\|\|\phi_h - \phi\|\|_{L^2}$	eoc	$\|\|v_h - v\|\|_{L^2}$	eoc	$\|\|\rho_h - \rho\|\|_{L^2}$	eoc
16	$1.60837e-03$	$---$	$7.96245e-02$	$---$	$5.23601e-02$	$---$
32	$2.52897e-04$	2.67	$3.70140e-02$	1.11	$3.25914e-02$	0.684
64	$3.46324e-05$	2.87	$1.57573e-021$	1.23	$1.50968e-02$	1.11
128	$4.11489e-06$	3.07	$4.89976e-03$	1.69	$5.05711e-03$	1.58
256	$1.00037e-06$	2.04	$8.95160e-04$	2.45	$9.60924e-04$	2.4

Table 2 Polynomial degree 2

Size	$\|\|\phi_h - \phi\|\|_{L^2}$	eoc	$\|\|v_h - v\|\|_{L^2}$	eoc	$\|\|\rho_h - \rho\|\|_{L^2}$	eoc
16	$3.03399e-05$	$---$	$5.03631e-03$	$---$	$4.65518e-03$	$---$
32	$2.07955e-06$	3.87	$5.72892e-04$	3.14	$5.88707e-04$	2.98
64	$2.46924e-07$	3.07	$5.38299e-05$	3.41	$5.99462e-05$	3.3
128	$3.09367e-08$	3.0	$4.67366e-06$	3.53	$5.78404e-06$	3.37
256	$3.88461e-09$	2.99	$3.97984e-07$	3.55	$7.29653e-07$	2.99

4.2 Bubble Ensemble

For the next numerical example, we fitted the free energies of the bulk phases to the equilibrium values of the van der Waals potential at a given temperature. To be precise, let ρ_v and ρ_l be the densities of the vapor and liquid phase, so that the equilibrium condition

$$p_{vdW}(\rho_l) = p_{vdW}(\rho_v) \tag{20}$$

$$\mu_{vdW}(\rho_l) = \mu_{vdW}(\rho_v) \tag{21}$$

is fulfilled; see [11]. Then, we can compute the values for the speed of sound c_l, c_v for the liquid and the vapor phase by $c_{l,v} = \sqrt{\partial_\rho p_{l,v}}$.

For the free energy F_1, we determine the coefficients a_1, b_1, c_1 by solving the following system of equations:

$$\begin{aligned} p_1(\rho_v) = a_1\rho_v - c_1 &= p_{vdW}, \\ \mu_1(\rho_v) = a_1 \log(\rho_v) - b1 &= \mu_{vdW}, \\ c_1 = \sqrt{a_1} &= c_v. \end{aligned}$$

The same can be done for F_2 using $\rho_l, c_l, p_{vdW}, \mu_{vdW}$. Furthermore, we have chosen $\delta = 0.01$, $A = 10^{-4}$, $\mu_1 = 0.001$, and $\eta = 100$. The simulation starts with an ensemble of bubbles where the densities are set to ρ_v inside of the bubbles and to ρ_l in the fluid surrounding the bubbles. There is a small transition layer between the inside and the outside of a bubble. In this region, the densities are interpolated by the phase-field variable which was set to $\phi \approx 1$ in the liquid region and $\phi \approx 0$ in the vapor bubbles. To have a steep but smooth transition profile, the constant states of the phase-field variable are connected with a $tanh$-profile over a width $\approx \delta$. We can see in Fig. 1 where the vapor phase is blue and the liquid phase is red that smaller bubbles collapse over time and tend to a system with a fewer but larger bubbles. This is what we expected, as the system tries to minimize the length of the transition layer and therefore the overall surface energy.

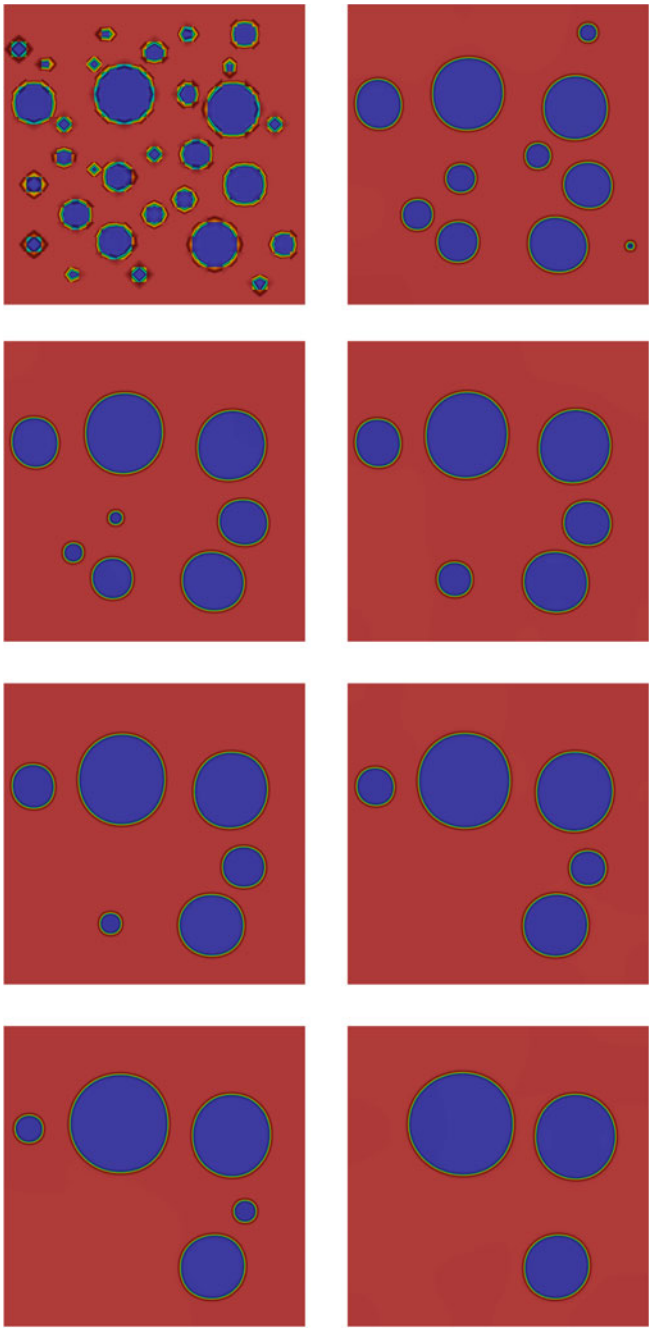


Fig. 1 Evolution of a bubble ensemble. Blue indicates the vapor phase and red the surrounding liquid

References

1. H.W. Alt, The entropy principle for interfaces. Fluids and solids. Adv. Math. Sci. Appl. **19**(2), 585–663 (2009)
2. H.W. Alt, W. Alt, Phase boundary dynamics: transitions between ordered and disordered lipid monolayers. Interfaces Free Bound. **11**(1), 1–36 (2009)
3. H.W. Alt, G. Witterstein, Distributional equation in the limit of phase transition for fluids. Interfaces Free Bound. **13**(4), 531–554 (2011)
4. D. Arnold, F. Brezzi, B. Cockburn, D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**(5), 1749–1779 (2002)
5. G. Aki, J. Daube, W. Dreyer, J. Giesselmann, M. Kränkel, C. Kraus, A diffuse interface model for quasi-incompressible flows: sharp interface limits and numerics. ESAIM Proc. **38**, 54–77 (2012)
6. T. Blesgen, *Generalization of the Navier-Stokes equations to two-phase flows. (Eine Verallgemeinerung der Navier–Stokes–Gleichungen auf Zweiphasenstrmungen.)* Ph.D. thesis Bonn, Hohe Mathematisch-Naturwissenschaftliche Fakultät, 100 S (1997)
7. D. Bresch, B. Desjardins, C.-K. Lin, On some compressible fluid models: Korteweg, lubrication and shallow water systems. Commun. Partial Differ. Equ. **28**(3), 843–868 (2003)
8. S. Benzoni-Gavage, R. Danchin, S. Descombes, *Well-posedness of One-dimensional Korteweg Models*, preprint (2004)
9. S. Benzoni-Gavage, R. Danchin, S. Descombes, *On the Well-posedness for the Euler–Korteweg Model in Several Space Dimensions*, preprint (2005)
10. J. Daube, *Sharp Interface Limit for the Navier–Stokes–Korteweg Equations*, thesis, Freiburg (2016)
11. D. Diehl, *Higher order schemes for simulation of compressible liquid-vapor flows with phase change.* Ph.D thesis, Universität Freiburg (2007), https://freidok.uni-freiburg.de/volltexte/3762/
12. D. Diehl, J. Kremser, D. Kröner, C. Rohde, Numerical solution of Navier–Stokes–Korteweg systems by local discontinuous Galerkin methods in multiple space dimensions. Appl. Math. Comput. **272**, part 2, 309–335 (2016)
13. W. Dreyer, C. Kraus, On the van der Waals–Cahn–Hilliard phase-field model and its equilibria conditions in the sharp interface limit. Proc. R. Soc. Edinb. Sect. A, Math. **140**(6), 1161–1186 (2010)
14. E. Feireisl, H. Petzeltova, E. Rocca, G. Schimperna, Analysis of a phase-field model for two-phase compressible fluids. Math. Models Methods Appl. Sci. **20**(07), 1129–1160 (2010)
15. J. Giesselmann, C. Makridakis, T. Pryer, Energy consistent DG methods for the Navier–Stokes–Korteweg system. Math. Comput. **83**(289), 2071–2099 (2014)
16. K. Hermsdörfer, C. Kraus, D. Kröner, Interface conditions for limits of the Navier–Stokes–Korteweg model. Interfaces Free Bound. **13**(2), 239–254 (2011)
17. M. Kotschote, *Strong Well-posedness for a Korteweg-Type Model for the Dynamics of a Compressible Non-isothermal Fluid*, Preprint Leipzig (2006)
18. M. Kotschote, Strong solutions of the Navier–Stokes equations for a compressible fluid of Allen–Cahn type. Arch. Ration. Mech. Anal. **206**(2), 489–514 (2012)
19. M. Kraenkel, *Discontinuous Galerkin Schemes for compressible Phasefield Flow*, thesis, Freiburg (2017)
20. D. Kröner, *Numerical Schemes for Conservation Laws* (Wiley-Teubner, 1997)
21. L.D. Landau, E.M. Lifschitz, *Hydrodynamik, 5*, überarbeitete edn. (Akademie Verlag, Berlin, 1991)
22. J. Lowengrub, L. Truskinovsky, Quasi-incompressible Cahn–Hilliard fluids and topological transitions. Proc. R. Soc. Lond. Ser. A., Math. Phys. Eng. Sci **454**(1978), 2617–2654 (1998)
23. A. Meurera et al., SymPy: symbolic computing in Python. Peer J. Comput. Sci. **3**, e103 (2017)
24. G. Witterstein, Sharp interface limit of phase change flows. Adv. Math. Sci. Appl. **20**(2), 585–629 (2010)
25. G. Witterstein, Phase change flows with mass exchange. Adv. Math. Sci. Appl. **21**(2), 559–611 (2011)

Mathematical Theory of Two-Phase Geochemical Flow with Chemical Species

W. J. Lambert, A. C. Alvarez, D. Marchesin and J. Bruining

Abstract In this work, we introduce a formalism for two-phase geochemical flow. Here, we admit that the chemical species flow in both phases. Moreover, we consider chemical interaction and chemical equilibrium laws for which it is possible to obtain algebraic relationships between the chemical species. In this work, we consider that we have only one free chemical species, i.e., by using equilibrium laws, we admit that all chemical species can be written as function of only one, which we denote as y. We present a formalism for this kind of flow, moreover, we obtain the eigenvalues, eigenvectors, and bifurcations structures. We also show the structure of integral and Hugoniot curves in the saturation versus chemical species plane.

Keywords Geochemical flow · Enhanced oil recovery
General structure for shocks · Rarefactions and bifurcation loci

This work was supported in part by: CNPq under Grants 402299/2012-4, 301564/2009-4, 470635/2012-6, FAPERJ under Grants E-26/111.416/2010, E-26/102.965/2011, E-26/110.658/2012, E-26/111.369/2012, E-26/110114.110/2013, ANP-731948/2010, PRH32-6000.0069459.11.4, CAPES Nuffic-024/2011, TUDelft, Section Petroleum Engineering.

W. J. Lambert (✉)
ICT - UNIFAL, Cidade Universitária, BR 267 Km 533, Rodovia José Aurélio Vilela, Poços de Caldas, MG, Brazil
e-mail: wanderson.lambert@unifal-mg.edu.br

A. C. Alvarez · D. Marchesin
Instituto Nacional de Matemática Pura e Aplicada, Estrada Dona, Castorina 110, Rio de Janeiro, RJ 22460-320, Brazil
e-mail: meissa98@impa.br

D. Marchesin
e-mail: marchesi@impa.br

J. Bruining
TU Delft, Civil Engineering and Geosciences, Stevinweg 1, 2628 CE Delft, The Netherlands
e-mail: J.Bruining@tudelft.nl

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_20

1 Introduction

One-dimensional multicomponent geochemical flow is modeled by system of equations of form $\frac{\partial G(V)}{\partial t} + \frac{\partial u F(V)}{\partial x} = 0$, in which $V = V(x,t) : \mathbb{R} \times \mathbb{R}^+ \longrightarrow \Omega \subset \mathbb{R}^n$, $G(V) = (G_1(V), \cdots, G_{n+1}(V)) : \Omega \longrightarrow \mathbb{R}^{n+1}$ and $F(V) = (F_1(V), \cdots, F_{n+1}(V)) : \Omega \longrightarrow \mathbb{R}^{n+1}$, see [3, 4, 13]. Here, we can decouple the variable V into two subgroups, the saturation variables s_i, for $i = 1, \cdots, p$ and the chemical species variables y_i, for $i = 1, \cdots, m$. The compatibility condition implies that $p + m + 1 = n$. Moreover, each G_i and F_i, for $i = 1, \cdots, n+1$, is written as:

$$G_i = \sum_{j=1}^{m} s_j C_{ij}(y) + a_i(y) \quad \text{and} \quad F_i = \sum_{j=1}^{m} f_j C_{ij}(y), \tag{1}$$

in which $a_i(y)$ are called adsorption functions. We have two different class of problems. First, f_j, for $j = 1, \cdots, p$, depends only on $s = (s_1, \cdots, s_p)$. The other scenario is f_j depending on s and $y = (y_1, \cdots, y_m)$, f_j is called *fractional flux*. In [10], we develop theory dealing with problems for general accumulation and fluxes terms G and F. In this paper, we consider the two-phase flow problem, i.e., $s = (s_w, s_o)$ and $s_w + s_o = 1$, with G and F are given by (1), in which we disregard adsorption phenomena, i.e., $a_i(y) = 0$ for all i. Here, we admit that the fractional fluxes f_w and f_o depend on s_w, s_0 and y. Under this assumptions, we obtain several important general structures appearing in this class of equations, in which we generalize several results and theory known in literature, such as integral and Hugoniot curves, coincidences, inflection. Moreover, we give a geometrical interpretation for these results.

Historically, the first models studied (from mathematical view point) for two-phase flow admitted that the chemical species flow only in one phase, for which we call s. Moreover, these models admitted that the chemical species appear as a linear function in the system of equations and they did not consider any adsorption effects. Under these hypothesis, one can prove that u is constant and G and F are written as:

$$G_1 = s, \quad F_1 = f, \quad G_{i+1} = s y_i \quad \text{and} \quad F_{i+1} = f y_i, \quad \text{for} \quad i = 1, 2, \cdots, n. \tag{2}$$

The eigenpairs for states (s, y), $y = (y_1, \cdots, y_n)$ are: one *saturation* eigenpair of form $\lambda_s = \frac{\partial f}{\partial s}$ $\mathbf{r}_s = (1, 0, \cdots, 0)$ and n chemical eigenvalues of form $\lambda_c = \frac{f}{s}$. Notice that all λ_c are equals; however, we have n different associated eigenvectors. We can take these eigenvectors of form e_i, for $i = 2, \cdots, n+1$ for which e_i for $i = 1, \cdots, n+1$ is the canonical basis of $\mathbb{R}^n$. In this model, there is a coincidence surface between eigenvalues of different families, i.e., there is (s, y) for which $\lambda_s = \lambda_c$. This system loses strict hyperbolicity on this coincidence surface. This phenomenon appeared also in elasticity problems, such as in the famous work [9]. In this work, authors have introduced an extension of Lax's and Liu's entropy conditions. This

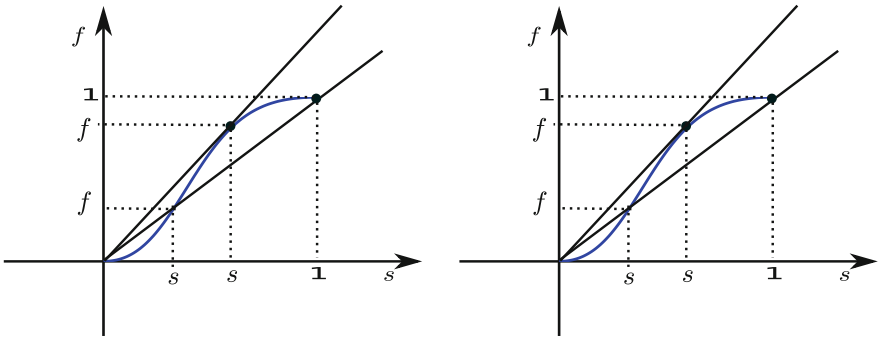


Fig. 1 **a**–*Left*. Coincidence between eigenvalues λ_s and λ_c for model without adsorption. **b**–*Right*. Coincidence between eigenvalues λ_s and λ_c for model with adsorption

coincidence has a interesting geometrical interpretation, see Fig. 1a. The eigenvalue λ_c for state (s, y) is the slope of line connecting $(0, 0)$ and $(s, f(s, y))$ and λ_c is the tangent to the graph $f(s, y)$, notice that there is a state (s, y) for which both line coincide.

Some papers discuss the mathematical theory for this model. For example, [15] in which the 2×2 Riemann and Cauchy problem is solved, and [5] in which the $(n + 1) \times (n + 1)$ Riemann and Cauchy problem is solved.

Posteriorly, models considered the interaction between the porous media and the chemical species, i.e., such models considered the existence of adsorption functions. For these, G and F are written as:

$$G_1 = s, \quad F_1 = f, \quad G_{i+1} = sy_i + a_i(y) \quad \text{and} \quad F_{i+1} = fy_i, \quad \text{for } i = 1, 2, \cdots, n. \tag{3}$$

$a_i = a_i(y)$, for $i = 1, \cdots, n$ are the adsorption functions. The eigenvalues are: one λ_s and n eigenvalues of form $\lambda_{c_i} = \dfrac{f}{s + h_i}$, in which h_i are the eigenvalues of Jacobian matrix of $a = (a_1, \cdots, a_n)^T$. This model exhibits n coincidence surfaces $\lambda_s = \lambda_{c_i}$. We also have a geometrical interpretation of this coincidence, see Fig. 1b for this interpretation. If $h_i \neq h_j$ for all $i, j = 1, \cdots, n$ thus $\lambda_{c_i} \neq \lambda_{c_j}$.

Some works were devoted to this problem, for which authors developed a mathematical theory. We can cite, for example, [2, 6–8]. In the [2], authors developed a algorithm to solve $n \times n$ system. Moreover, some other methods have been developed to obtain analytical solution exploring the possibility of decoupling between variable s (hydrodynamical) and y (thermodynamical); in this sense, a very interesting method was proposed by [14].

In the present work, we consider that the chemical species can flow in both phases. Moreover, we use equilibrium laws to obtain, from chemical species y, the so-called concentration functions. Under equilibrium, we can use Gibbs phase rule (see, e.g., [11]) that states that the number of degrees of freedom is given by

$$N_f = N_s - N_r - N_c + 2 - p, \tag{4}$$

where N_s is the number of different chemical species, N_r is the number of possible equilibrium reactions within the phases, N_c is the number of constraints, e.g., the charge balance. The number 2 represents the temperature and pressure and p the number of phases. We follow Appelo and Parkhurst [1, 12] to obtain the algebraic relationships between chemical species. From this modeling, we have, generically, nonlinear concentration functions. In order to obtain our formalism, we assume that in this flow there is only one free chemical species, denoted by y.

Thus, we have four unknowns: s_w, s_o, y and u. Moreover, we assume that $s_o = 1 - s_w$. Thus, disregarding any desorption effects, we have 3×3

$$\frac{\partial}{\partial t}(s_w C_{i1}(y) + s_o C_{i2}(y)) + \frac{\partial}{\partial x} u(f_w C_{i1}(y) + f_o C_{i2}(y)) = 0, \tag{5}$$

in which each concentration is given by C_{ij}, for $i = 1, 2, 3$. The fractional fluxes of water, f_w, and oil, f_o, depending on s, s_o and y, i.e., $f_w = f_w(s, y)$. From physical considerations $0 \leq s \leq 1$ and $0 \leq y_i \leq 1$ for $i = 1, \cdots, n-1$, thus we denote $\Omega = \{(s, y)\} = [0, 1] \times [0, 1]$ as the phase space.

In Sect. 2, we obtain the eigenvalues of the system, we also obtain many important structures and we prove important results on the topology of integral curves. In Section rsk, we obtain the structure of Hugoniot locus. These structures are the main ingredients to obtain the Riemann solutions and to analyze the stability of solution. The focus of this work is to use this formalism (and to extend to $n \times n$ system) to solve important applied problems. In the Sect. 4, we draw our conclusions.

2 Eigenvalues, Eigenvectors and Bifurcations

The system of eigenvalues is written as $A\mathbf{r} = B\lambda\mathbf{r}$, where $\mathbf{r} = (s, y, u)^T$ and B, A are:

$$B = \left([C_i]\ s_w C'_{i1} + s_o C'_{i2}\ 0\right)_{1 \leq i \leq 3}, \tag{6}$$

$$A = \left(u \frac{\partial f_w}{\partial s_w}[C_i]\ u\left(f_w C'_{i1} + f_o C'_{i2} + \frac{\partial f_w}{\partial y}[C_i]\right)\ F_i\right)_{1 \leq i \leq 3}, \tag{7}$$

in which $[C_i] = C_{i1} - C_{i2}$. Notice that the matrix B is singular, because it has a complete row of zeros, this condition was analyzed in [10].

To obtain the eigenvalues, we solve $det(A - \lambda B) = 0$, where $A - \lambda B$ is:

$$\left(\left(u \frac{\partial f_w}{\partial s_w} - \lambda\right)[C_i]\ (u f_w - \lambda s_w) C'_{i1} + (u f_o - \lambda s_o) C'_{i2} + u \frac{\partial f_w}{\partial y}[C_i]\ F_i\right)_{1 \leq i \leq 3} \tag{8}$$

Substituting the second row by the sum of the first row multiplied by $[C_2]$ with second row multiplied by $[C_1]$. Similarly, we substitute the third row by the sum of the first row multiplied by $[C_3]$ with second third row multiplied by $[C_1]$. After, we substitute the second row by the sum of the third row multiplied by $-\nu_1$ with the second row multiplied by ν_2, and we obtain:

$$\begin{pmatrix} \left(u\frac{\partial f_w}{\partial s_w}-\lambda\right)[C_1] & (uf_w-\lambda s_w)C'_{11}+(uf_o-\lambda s_o)C'_{12}+u\frac{\partial f_w}{\partial y}[C_i] & F_1 \\ 0 & (uf_w-\lambda s_w)\vartheta_1+(uf_o-\lambda s_o)\vartheta_2 & 0 \\ 0 & (uf_w-\lambda s_w)\gamma_{31}+(uf_o-\lambda s_o)\gamma_{32} & \nu_2 \end{pmatrix} \tag{9}$$

where $\vartheta_i=\gamma_{2i}\nu_3-\gamma_{3i}\nu_1$, $\gamma_{ij}=C'_{ij}[C_1]+C'_{1j}[C_i]$ and $\nu_i=C_{12}C_{i1}-C_{11}C_{i2}$.

Thus, we have two eigenpairs. The first one is $(\lambda_s, \mathbf{r}_s)$:

$$\lambda_s=u\frac{\partial f_w}{\partial s_w} \quad \text{and} \quad \mathbf{r}_s=(1,0,0). \tag{10}$$

For this eigenpair, only saturation changes and we identify this family wave as *saturation wave* or Buckley–Leverett type wave.

If $\vartheta_1-\vartheta_2\neq 0$, we can write the second eigenpair, λ_Δ, as:

$$\lambda_\Delta=u\frac{f_w-\Delta}{s_w-\Delta}, \quad \text{where} \quad \Delta=\frac{\vartheta_2}{\vartheta_2-\vartheta_1}. \tag{11}$$

The form (11.*a*) has an interesting geometrical interpretation. For each s_w^* fixed, λ_w corresponds to the slope of line in the (s_w, f_w)-plane connecting the point $(-\Delta, -\Delta)$ to the point $(s^*, f_w(s^*))$. To obtain the corresponding eigenvector, we substitute λ_Δ given by (11) into (9) and using that

$$uf_w-\lambda_\Delta s_w=u\frac{f_w(s_w-\Delta)-(f_w-\Delta)s_w}{s_w-\Delta}=u\Delta\frac{s_w-f_w}{s_w-\Delta}. \tag{12}$$

$$uf_o-\lambda_\Delta s_o=u(1-f_w)-\lambda_\Delta(1-s_w)=u\frac{(1-f_w)(s_w-\Delta)-(f_w-\Delta)(1-s_w)}{s_w-\Delta}=u(1-\Delta)\frac{s_w-f_w}{s_w-\Delta}, \tag{13}$$

after some tedious calculations, we obtain:

$$\begin{pmatrix} \left(u\frac{\partial f_w}{\partial s}-\lambda\right)[C_1] & u\frac{s_w-f_w}{s_w-\Delta}\left(\Delta C'_{11}+(1-\Delta)C'_{12}\right)+u\frac{\partial f_w}{\partial y}[C_1] & F_1 \\ 0 & u\frac{s_w-f_w}{s_w-\Delta}\left(\Delta\gamma_{31}+(1-\Delta)\gamma_{32}\right) & \nu_2 \end{pmatrix}\begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}=0 \tag{14}$$

By solving (14) and using that $(1-\Delta)=-\vartheta_1/\vartheta_2\Delta$, after some calculations the eigenvector $\mathbf{r}_\Delta$ can be written as:

$$\mathbf{r}_\Delta=\left(-\frac{s_w-f_w}{(s_w-\Delta)}\hat{r}_1-\frac{\partial f_w}{\partial y},\frac{\partial f_w}{\partial s_w}-\left(\frac{f_w-\Delta}{s_w-\Delta}\right),u\frac{s_w-f_w}{s_w-\Delta}\left(\frac{\partial f_w}{\partial s_w}-\left(\frac{f_w-\Delta}{s_w-\Delta}\right)\right)\hat{r}_3\right). \tag{15}$$

where

$$\hat{r}_1 = \frac{\Delta}{\vartheta_2}\left(\frac{C'_{11}\vartheta_1 - C'_{12}\vartheta_1 + (\gamma_{22}\gamma_{31} - \gamma_{21}\gamma_{32})\,F_1}{[C_1]}\right), \quad \hat{r}_3 = \frac{\Delta}{\vartheta_2}\,(\gamma_{22}\gamma_{31} - \gamma_{21}\gamma_{32})\,. \tag{16}$$

From the eigenvectors $\mathbf{r}_\Delta$, we are able to obtain the integral curves solutions of:

$$\frac{ds_w}{d\xi} = -\frac{s_w - f_w}{(s_w - \Delta)}\hat{r}_1 - \frac{\partial f_w}{\partial y}, \quad \frac{dy}{d\xi} = \frac{\partial f_w}{\partial s_w} - \left(\frac{f_w - \Delta}{s_w - \Delta}\right), \tag{17}$$

$$\frac{du}{d\xi} = u\frac{s_w - f_w}{s_w - \Delta}\left(\frac{\partial f_w}{\partial s_w} - \left(\frac{f_w - \Delta}{s_w - \Delta}\right)\right)\hat{r}_3. \tag{18}$$

In the next sections, we are able to identify the structure of each integral wave in the phase plane (s_w, y); moreover, we obtain conditions to identify the rarefaction branches of each integral curve.

Remark 1. In some models, it is possible that $\vartheta_2 - \vartheta_1 = 0$, which leads to discontinuities (at least numerically). To overcome this problem, we rewrite the eigenpair in a similar form removing this singularity. First of all, we write $\Delta = \vartheta_2/\Delta_1$, in which $\Delta_1 = \vartheta_2 - \vartheta_1$. From similar calculations, we write the eigenpair $(\lambda_\Delta, \mathbf{r}_\Delta)$ as

$$\lambda_\Delta = \frac{f_w\Delta_1 - \vartheta_2}{s_w\Delta_1 - \vartheta_2}. \tag{19}$$

$$\mathbf{r}_\Delta = \left(-\frac{s_w - f_w}{(s_w\Delta_1 - \vartheta_2)}\bar{r}_1 - \frac{\partial f_w}{\partial y}, \frac{\partial f_w}{\partial s_w} - \frac{f_w\Delta_1 - \vartheta_2}{s_w\Delta_1 - \vartheta_2}, u\frac{s_w - f_w}{s_w\Delta_1 - \vartheta_2}\left(\frac{\partial f_w}{\partial s_w} - \left(\frac{f_w\Delta_1 - \vartheta_2}{s_w\Delta_1 - \vartheta_2}\right)\right)\bar{r}_3\right). \tag{20}$$

where

$$\bar{r}_1 = \left(\frac{C'_{11}\vartheta_2 - C'_{21}\vartheta_1 + (\gamma_{22}\gamma_{31} - \gamma_{21}\gamma_{32})\,F_1}{[C_1]}\right), \quad \bar{r}_3 = (\gamma_{22}\gamma_{31} - \gamma_{21}\gamma_{32})\,. \tag{21}$$

Moreover, notice that $\lim_{\Delta_1 \longrightarrow 0}\lambda_\Delta = 1$.

2.1 Bifurcation Structures in the Model

The are several important structures in the phase space, called *bifurcations*. These structures, generically, are used to divide the phase space in subregions in which the sequence of waves for the Riemann solution is the same. The main bifurcation structures appearing in this model are: inflections and coincidences.

The inflections are co-dimension-1 structures, in which the increasing of characteristic speed fails, i.e., $\nabla\lambda \cdot \mathbf{r} = 0$. In this model, we have two fields. For the field λ_s, it is easy to see that $\nabla\lambda_s \cdot \mathbf{r} = \dfrac{\partial^2 f_w}{\partial s_w^2}$, thus the inflection states are the

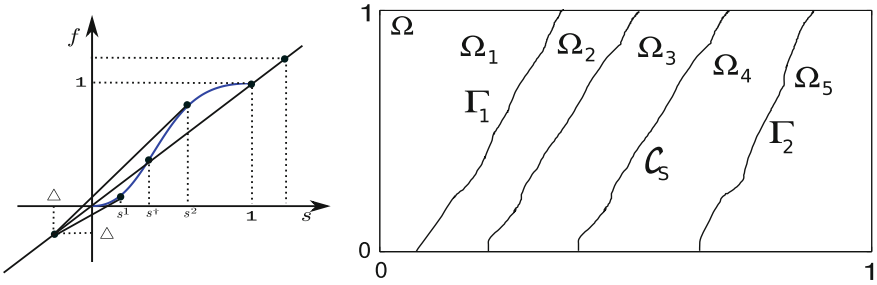


Fig. 2 **left a**—States s^1, s^* and s^2. **left b**—Regions Ω_i and curves Γ_1, $\mathscr{C}_s$, $\mathscr{I}_s$ and Γ_2. In this regions $s^* < s^\dagger$, however, it is possible also that $s^\dagger \leq s^*$

$s^* = s^*(y)$, satisfying $\dfrac{\partial^2 f_w}{\partial s_w^2}(s^*, \cdot) = 0$. We denote the inflection for the field $(\lambda_s, \mathbf{r}_s)$ as $\mathscr{I}_s = \left\{(s^*, y);\ \text{such that } \dfrac{\partial^2 f_w}{\partial s_w^2}(s^*, y) = 0\right\}$. If f_w does not depends on y, thus $\mathscr{I}_s$ is a vertical line in the (s_w, y) plane.

Before we obtain the inflection of field $(\lambda_\Delta, \mathbf{r}_\Delta)$, we obtain coincidence structures that are fundamental to obtain the inflection locus.

The first coincidence is between the eigenvalues λ_s and λ_Δ, which we denote as Γ. From geometrical arguments, see Fig. 2, we state the following result.

Lemma 1. *If $\Delta > 0$ or $\Delta < -1$, for each state y, there are saturations $s^1(y)$ and $s^2(y)$, satisfying $s^1(y) < s^*(y) < s^2(y)$, such that*

$$\frac{\partial f_w(s^1, y)}{\partial s_w} = \frac{f_w(s^1, y) - \Delta(y)}{s_w^1 - \Delta(y)} \quad \text{and} \quad \frac{\partial f_w(s^2, y)}{\partial s_w} = \frac{f_w(s^2, y) - \Delta(y)}{s_w^2 - \Delta(y)}. \tag{22}$$

Remark 2. From geometrical arguments, we also can see that there are one state $s^\dagger = s^\dagger(y)$ such that $f_w(s^\dagger, y) = s^\dagger$, such that, for each y, $s^1(y) < s(y)^\dagger < s^2(y)$. However we can not identify if $s^\dagger(y) \leq s^*(y)$ or $s^\dagger(y) \geq s^*(y)$. Thus we define $\mathscr{C}_s = \left\{(s^\dagger, y);\ \text{such that } f_w(s^\dagger, y) = s^\dagger\right\}$.

From Lemma 1 and Remark 2, we state the following result:

Proposition 1. *We identify four curves partitioning the (s_w, y)-plane, which are $\Gamma_1 = \{(s^1(y), y) \in \Omega\}$, $\Gamma_2 = \{(s^2(y), y) \in \Omega\}$, $\mathscr{I}_s = \{(s^*(y), y) \in \Omega\}$ and $\mathscr{C}_s = \{(s^\dagger(y), y) \in \Omega\}$. In the situation that, f_w does not depend on y $\mathscr{I}_s$ and $\mathscr{C}_s$ are straight lines which are parallels to axis y.*

The curve $\mathscr{C}_s$ is parametrized as function of y, then differentiating $s_w(y) = f_w(s_w, y)$ with relationship with y, we obtain:

$$\frac{ds_w}{dy} = \frac{\partial f_w}{\partial s_w}\frac{ds_w}{dy} + \frac{\partial y}{\partial y} \tag{23}$$

thus, rearranging (23), we write the following Lemma:

Lemma 2. *Let* (s_w^θ, y^θ) *one point satisfying* $s_w^\theta = f_w(s_w^\theta, y^\theta)$*, thus curve* $\mathscr{C}_s$ *is obtained as solution of:*

$$\frac{ds_w}{dy} = -\frac{\frac{\partial f_w}{\partial y}}{1 - \frac{\partial f_w}{\partial s_w}} \qquad \text{and } s_w^\theta = f_w(s_w^\theta, y^\theta). \tag{24}$$

It is also useful to define five regions Ω_i, for $i = 1, \cdots, 3$:

$$\begin{aligned}
\Omega_1 &= \{(s_w, y) \quad \text{such that} \quad 0 \le s_w < s^1(y)\},\\
\Omega_2 &= \{(s_w, y) \quad \text{such that} \quad s^1(y) \le s_w < \min(s^*, s^\dagger)\},\\
\Omega_3 &= \{(s_w, y) \quad \text{such that} \quad \min(s^*, s^\dagger) \le s_w < \max(s^*, s^\dagger)\},\\
\Omega_4 &= \{(s_w, y) \quad \text{such that} \quad s^1 \le s_w < s^2(y)\},\\
\Omega_5 &= \{(s_w, y) \quad \text{such that} \quad s^2(y) < s_w \le 1\}.
\end{aligned}$$

The curves $s = 0$, Γ_1, $\mathscr{C}_s$, $\mathscr{I}_s$, Γ_2 and $s = 1$ are boundaries of Ω_i, see Fig. 2b.

Now, we are able to prove some important results about the topology of the integral curves for the fields $(\lambda_s, \mathbf{r}_s)$ and $(\lambda_\Delta, \mathbf{r}_\Delta)$. For $(\lambda_s, \mathbf{r}_s)$, the integral curves are straight lines in the (s_w, y) plane which are parallel to s_w axis. For the field $(\lambda_\Delta, \mathbf{r}_\Delta)$, we use Eqs. (17)–(18), so we can identify the shape of each integral curve. Notice that from Eq. (17.*b*) for any state in Ω_1 and Ω_5, since $\lambda_s < \lambda_\Delta$, we have that $\dfrac{dy}{d\xi} < 0$; in Ω_2, Ω_3 and Ω_4, since $\lambda_s < \lambda_\Delta$, we have that $\dfrac{dy}{d\xi} > 0$. From (17.*a*) and Lemma 2, one can see that the integral curve has the curve $\mathscr{C}_s$ as a asymptotic curve, i.e., the integral curve for the field $(\lambda_\Delta, \mathbf{r}_\Delta)$ does not cross $\mathscr{C}_s$. To prove this, let a state (s_w^σ, y^σ) before $\mathscr{C}_s$. The integral curve from this point satisfy the system (17)–(18), and we denote them as $\gamma = \gamma(s_w^\sigma, y^\sigma)$. Suppose by absurd that γ reaches $\mathscr{C}_s$, then, there is a state $(\tilde{s}_w^\sigma, \tilde{y}^\sigma)$ in $\mathscr{C}_s$ and γ. Then this state satisfy $\tilde{s}_w^\sigma = f_w(\tilde{s}_w^\sigma, \tilde{y}^\sigma)$ and Eqs. (17.*a*) and (17.*b*) reduce to

$$\frac{ds_w}{d\xi} = -\frac{\partial f_w}{\partial y} \quad \text{and} \quad \frac{dy}{d\xi} = \frac{\partial f_w}{\partial s_w} - 1, \tag{25}$$

Dividing (25.*a*) by (25.*b*), we obtain (24), which is the equation for $\mathscr{C}_s$, i.e., if γ reached $\mathscr{C}_s$ the uniqueness of ODE system fails, which is a contraction. Thus, the integral curve from any state before (s_w, y) does not reach $\mathscr{C}_s$. We can say that $\mathscr{C}_s$ is a *barrier* for this integral curves.

Moreover, notice that from (17.*a*) that $\dfrac{ds}{d\xi}$ depends on $s_w - f_w$, thus for states on the same horizontal line in the (s_w, f_w) plane, the sign of $\dfrac{ds}{d\xi}$ change when we

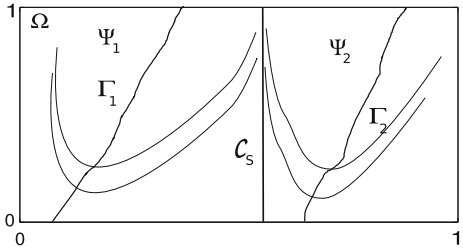


Fig. 3 left a—Integral curves for eigenpair $(\lambda_\Delta, \mathbf{r}_\Delta)$. Notice that the integral curve change the behavior when crosses Γ_i and that $\mathscr{C}_s$ is a barrier for this curve. We denote the invariant regions Ψ_i

take states crossing $\mathscr{C}_s$. From these conditions, we can see that all integral curves for eingenpair $(\lambda_\Delta, \mathbf{r}_\Delta)$ has a minimum on Γ_1 and Γ_2 or a maximum on these curves.

We can summarize these results as:

Proposition 2. *The curve $\mathscr{C}_s$ is a barrier for the integral curves of $(\lambda_\Delta, \mathbf{r}_\Delta)$, i.e., integral curve of any state taken out from $\mathscr{C}_s$ does not cross $\mathscr{C}_s$. Moreover, the integral curves have either a maximum or a minimum on $\mathscr{I}_1$ and $\mathscr{I}_2$.*

The shape of these curves is shown in Fig. 3.

From Proposition 2, we can identify some invariant regions for the integral curves (thus rarefaction curves) of eigenpair $(\lambda_\Delta, \mathbf{r}_\Delta)$ and identify the behavior and form of these integral curves. Using previous regions Ω_i, for $i = 1, \cdots, 5$ we define:

$$\Psi_1 = \{(s, y) \quad \text{such that} \quad 0 \leq s < s^\dagger\}, \tag{26}$$

$$\Psi_2 = \{(s, y) \quad \text{such that} \quad s^\dagger \leq s < 1\}. \tag{27}$$

Notice that when $s^* < s^\dagger$, then $\Psi_1 = \Omega_1 \cup \Omega_2 \cup \Omega_3$ and $\Psi_2 = \Omega_4 \cup \Omega_5$. In other hand, if $s^* > s^\dagger$, then $\Psi_1 = \Omega_1 \cup \Omega_2$ and $\Psi_2 = \Omega_3 \cup \Omega_4 \cup \Omega_5$.

Previous results leads to the following result:

Corollary 1. *The integral curves with initial states in Ψ_i remains in Ψ_i, i.e., Ψ_i is invariant for integral curves associated to $(\mathbf{r}_\Delta, \lambda_\Delta)$ field. Moreover, y modify the behavior when the integral curve crosses Γ_i.*

Another important results are on the inflection locus associated to λ_Δ. Several authors, for example, [10], have noticed that the coincidence states satisfying $\lambda_s = \lambda_\Delta$ lies on the inflection locus of $(\lambda_\Delta, \mathbf{r}_\Delta)$, in some models $\mathscr{C}_s$ is also in the same inflection locus. Here, we obtain a general result that clarifies this discussion, moreover, we decouple this inflection locus in two parts one that depends on s_w and y and another that depends only on y. These results help us to explain the topology of this structure.

Lemma 3. *The states satisfying $\lambda_s = \lambda_\Delta$ and $f_w = s_w$ are in the inflection locus of eigenpair $(\lambda_\Delta, \mathbf{r}_\Delta)$ field.*

Proof: Let λ_Δ given by (11), calculating $\nabla\lambda_\Delta$, we obtain:

$$\nabla\lambda_\Delta = \left(\frac{1}{s_w - \Delta}(\lambda_s - \lambda_\Delta), \frac{-u\Delta'}{(s_w - \Delta)^2}(s_w - f_w) - u\frac{\partial f_w}{\partial y}, \frac{\lambda_\Delta}{u}\right). \tag{28}$$

Using $\mathbf{r}_\Delta$ given in (15), we obtain, after some calculations:

$$\nabla\lambda_\Delta \cdot \mathbf{r}_\Delta = \frac{s_w - f_w}{(s_w - \Delta)^2}(\lambda_s - \lambda_\Delta)\mathscr{H}(s, y), \tag{29}$$

where $\mathscr{H} = -\hat{r}_1 - \Delta' + \hat{r}_3(f_w - \Delta)$. Thus the curves Γ_1, Γ_2 and $\mathscr{C}_s$ are in the inflection locus of $(\lambda_\Delta, \mathbf{r}_\Delta)$. □.

Lemma 4. *The function $\mathscr{H} = -\hat{r}_1 - \Delta' + \hat{r}_3(f_w - \Delta)$ does not depend on s_w.*

Proof: Using $\hat{r}_1$ and $\hat{r}_3$ defined in (16), we can write $-\hat{r}_1 + \hat{r}_3 f$ as:

$$\begin{aligned}
&-\frac{\Delta}{\vartheta_{24}}\left(\frac{C_1'\vartheta_{24} - C_2'\vartheta_{13} + (\gamma_2\gamma_3 - \gamma_1\gamma_4)\,F_1}{C_1 - C_2}\right) + f_w\frac{\Delta}{\vartheta_{24}}\,(\gamma_2\gamma_3 - \gamma_1\gamma_4) = \\
&-\frac{\Delta}{\vartheta_{24}}\left(\frac{C_1'\vartheta_{24} - C_2'\vartheta_{13} + (\gamma_2\gamma_3 - \gamma_1\gamma_4)\,(f_w(C_1 - C_2) + C_2)}{C_1 - C_2} - f_w\,(\gamma_2\gamma_3 - \gamma_1\gamma_4)\right) = \\
&-\frac{\Delta}{\vartheta_{24}}\left(\frac{C_1'\vartheta_{24} - C_2'\vartheta_{13} + (\gamma_2\gamma_3 - \gamma_1\gamma_4)\,(C_2)}{C_1 - C_2}\right).
\end{aligned} \tag{30}$$

Substituting (30) in $\mathscr{H}$ and since $\hat{r}_3$ and Δ' does not depend on s the Lemma is proved. □

We define the curve $\mathscr{J}_{\mathscr{H}} = \{(s, y),\ \text{such that } \mathscr{H}(y) = 0\}$. Notice that $\mathscr{J}_{\mathscr{H}}$ are straight lines parallels to axis s.

From previous results, we can state the following result:

Proposition 3. *The inflection locus of field $(\lambda_\Delta, \mathbf{r}_\Delta)$, denoted as $\mathscr{I}_\Delta$ is the union of curves Γ_1, Γ_2, $\mathscr{C}_s$ and $\mathscr{J}_{\mathscr{H}}$.*

Remark 3. We can obtain a similar equation that we obtain in Remark 1 for the inflection locus $\mathscr{I}_\Delta$. Writing $\Delta = \vartheta_2/\Delta_1$ and differentiating with respect to y:

$$\Delta' = \left(\frac{\vartheta_2}{\Delta}\right)' = \frac{\vartheta_2'\Delta_1 - \vartheta_2\Delta_1'}{\Delta_1^2}. \tag{31}$$

Substituting (31) in $\mathscr{H} = -\hat{r}_1 - \Delta' + \hat{r}_3(f_w - \Delta)$, using $\bar{r}_1$ and $\bar{r}_3$ given by (21) and (20), after some tedious calculations, using Eq. (29), $\nabla\lambda_\Delta \cdot \mathbf{r}_\Delta$ becomes:

$$\frac{s_w - f_w}{(\Delta_1 s_w - \vartheta_2)^2}(\lambda_s - \lambda_\Delta)\left(\Delta_1(-\bar{r}_1 + \bar{r}_3 f_w)) - \left(\vartheta_2'\Delta_1 - \vartheta_2\Delta_1'\right) - \bar{r}_3\vartheta_2\right) \tag{32}$$

If we take $\Delta_1 \longrightarrow 0$, we obtain that $\lim_{\Delta_1 \longrightarrow 0} \nabla\lambda_\Delta \cdot \mathbf{r}_\Delta = \dfrac{s_w - f_w}{\vartheta_2}(\lambda_s - \lambda_\Delta)\left(\Delta_1' - \bar{r}_3\right)$. Thus, $\nabla\lambda_\Delta \cdot \mathbf{r}_\Delta$ is a continuous function of s_w and y.

3 Rankine–Hugoniot Locus

We want to obtain discontinuous solutions associated to system (5) satisfying the Rankine–Hugoniot condition, which is written, for $i = 1, 2, 3$, as:

$$v^s(s_w^+ C_{i1}^+ + s_o^+ C_{i2}^+ - (s_w^- C_{i1}^- + s_o^- C_{i2}^-)) = u^+(f_w^+ C_{i1}^+ + f_o^+ C_{i2}^+) - u^-(f_w^- C_{i1}^- + f_o^- C_{i2}^-). \quad (33)$$

where v^s is the speed of discontinuity; $C^\pm = C(y^\pm)$ and $f^\pm = f(s^\pm)$.

For each fixed state (s^-, y^-), we obtain a set of states (s^+, y^+) satisfying (33), which we call *Hugoniot–locus*, which we denote as $\mathscr{H}\mathscr{L}(s^+, y^+)$. Following [10], for each fixed (s^-, y^-), the states satisfying (33) are obtained solving:

$$\det\left(s^+\mathscr{A}_{123}^+ + \mathscr{B}_{123}^+ - (s^-\mathscr{A}_{123}^- + \mathscr{B}_{123}^-) \; -f^+\mathscr{A}_{123}^+ - \mathscr{B}_{123}^+ \; f^-\mathscr{A}_{123}^- + \mathscr{B}_{123}^-\right) = 0, \quad (34)$$

for which we substitute $s_o^\pm = 1 - s_w^\pm$ and $f_o^\pm = 1 - f_w^\pm$ and we define:

$$\mathscr{A}_{ijk} = \mathscr{A}_{ijk}(y) = (C_{i1} - C_{i2}, C_{j1} - C_{j2}, C_{k1} - C_{k2})^T, \quad (35)$$

$$\mathscr{B}_{ijk} = \mathscr{B}_{ijk}(y) = (C_{i2}, C_{j2}, C_{k2})^T \quad \text{and} \quad \mathscr{D}_{ijk} = \mathscr{D}_{ijk}(y) = (C_{i1}, C_{j1}, C_{k1})^T. \quad (36)$$

Notice that

$$\mathscr{A}_{ijk} = \mathscr{B}_{ijk} + \mathscr{D}_{ijk}. \quad (37)$$

We can simplify Eq. (34). To do so, we sum second column the third column multiplied by -1 with first column and we obtain:

$$\det\left((s^+ - f^+)\mathscr{A}_{123}^+ - (s^- - f^-)\mathscr{A}_{123}^- \; -f^+\mathscr{A}_{123}^+ - \mathscr{B}_{123}^+ \; f^-\mathscr{A}_{123}^- + \mathscr{B}_{123}^-\right) = 0, \quad (38)$$

Applying determinant property on (38), we can write (38) as:

$$\begin{aligned}(s^+ - f^+)\left(\det\left(\mathscr{A}_{123}^+ \; -f^+\mathscr{A}_{123}^+ - \mathscr{B}_{123}^+ \; f^-\mathscr{A}_{123}^- + \mathscr{B}_{123}^-\right)\right) + \\ -(s^- - f^-)\left(\det\left(\mathscr{A}_{123}^- \; -f^+\mathscr{A}_{123}^+ - \mathscr{B}_{123}^+ \; f^-\mathscr{A}_{123}^- + \mathscr{B}_{123}^-\right)\right) = 0, \quad (39)\end{aligned}$$

Now, using (37) and applying some properties of determinant, we can write (39) as:

$$\begin{aligned}&(s^+ - f^+)\left(\det\left(\mathscr{D}_{123}^- \; \mathscr{D}_{123}^+ \; \mathscr{B}_{123}^+\right) + (1 - f^-)\det\left(\mathscr{D}_{123}^+ \; \mathscr{B}_{123}^+ \; \mathscr{B}_{123}^-\right)\right) \\ =&(s^- - f^-)\left(\det\left(\mathscr{D}_{123}^- \; \mathscr{D}_{123}^+ \; \mathscr{B}_{123}^-\right) + (1 - f^+)\det\left(\mathscr{D}_{123}^- \; \mathscr{B}_{123}^+ \; \mathscr{B}_{123}^-\right)\right). \quad (40)\end{aligned}$$

For each fixed (s^-, y^-), from Eq. (40), it is clear that a branch of $\mathscr{R}\mathscr{H}(s^-, y^-)$ is (s, y^-) (here, we drop the upper index $+$), i.e., the saturation Branch, that satisfy identically (40). In order hand, if we admit that y is not constant, than we obtain the other branch. Defining $\mathscr{F}(s, y(s))$ as:

$$\mathscr{F}(s, y(s)) = (s^+ - f^+)\left(\det\left(\mathscr{D}_{123}^- \mathscr{D}_{123}^+ \mathscr{B}_{123}^+\right) + (1 - f^-)\det\left(\mathscr{D}_{123}^+ \mathscr{B}_{123}^+ \mathscr{B}_{123}^-\right)\right)$$
$$- (s^- - f^-)\left(\det\left(\mathscr{D}_{123}^- \mathscr{D}_{123}^+ \mathscr{B}_{123}^-\right) + (1 - f^+)\det\left(\mathscr{D}_{123}^- \mathscr{B}_{123}^+ \mathscr{B}_{123}^-\right)\right). \quad (41)$$

Differentiating $\mathscr{F}$ with respect to s, we can prove the following result:

Lemma 5. *For each fixed* (s^-, y^-)*, the* $\mathscr{RH}(s^-, y^-)$ *has two branch: one branch is associated to saturation variation and* y *is constant. The other branch where* y *is not constant changes this behavior for states satisfying:* $\mathscr{F}(s, y) = 0$, $\dfrac{df}{ds} = \dfrac{\mathscr{M}}{\mathscr{M} - \mathscr{N}}$. *When there is not states satisfying these equalities, then* $\mathscr{RH}(s^-, y^-)$ *reaches the boundaries* $s = 0$ *and* $s = 1$.

From the form (40), under suitable hypothesis, we can study the behavior of this other branch of Rankine–Hugoniot locus. One can prove the following result:

Lemma 6. *If* $\det\left(\mathscr{D}_{123}^- \mathscr{D}_{123}^+ \mathscr{B}_{123}^+\right) + (1 - f^-)\det\left(\mathscr{D}_{123}^+ \mathscr{B}_{123}^+ \mathscr{B}_{123}^-\right)$ *and* $\det\left(\mathscr{D}_{123}^- \mathscr{D}_{123}^+ \mathscr{B}_{123}^-\right) + (1 - f^+)\det\left(\mathscr{D}_{123}^- \mathscr{B}_{123}^+ \mathscr{B}_{123}^-\right)$ *have the same sign, then* Ψ_i *are invariant regions for the Rankine–Hugoniot locus, i.e., if* $(s^-, y^-) \in \Psi_i$*, thus* $\mathscr{RH}(s^-, y^-) \subset \Psi_i$.

Proof: Let $(s^-, y^-) \in \Psi_1$, then $s^- - f^- < 0$. Since $\det\left(\mathscr{D}_{123}^- \mathscr{D}_{123}^+ \mathscr{B}_{123}^+\right) + (1 - f^-)\det\left(\mathscr{D}_{123}^+ \mathscr{B}_{123}^+ \mathscr{B}_{123}^-\right)$ and $\det\left(\mathscr{D}_{123}^- \mathscr{D}_{123}^+ \mathscr{B}_{123}^-\right) + (1 - f^+)\det\left(\mathscr{D}_{123}^- \mathscr{B}_{123}^+ \mathscr{B}_{123}^-\right)$ have the same sign, then using (40) we have that $s^+ - f^+ < 0$, i.e., $(s^+, y^+) \in \Psi_1$. Using the same argument we prove that Ψ_2 is invariant for the Rankine–Hugoniot locus, see Fig. 3b. □.

4 Conclusions

We introduce a formalism for the two-phase geochemical flow for one freedom degree of chemical species, denoted as y, flowing in both phases. Moreover, we admit that the concentrations functions are nonlinear functions of y. We obtain several bifurcation structures, and we describe the behavior of integral and Hugoniot curves. These structures can be applied in real models to obtain analytical solutions.

References

1. C.A.J. Appelo, D. Postma, *Geochemistry, Groundwater and Pollution* (Taylor & Francis, 2005)
2. O. Dahl, T. Johansen, A. Tveito, R. Winther, Multicomponent chromatography in a two phase environment. SIAM J. Appl. Math. **52**(1), 65–104 (1992)
3. F.G. Helfferich et al., Theory of multicomponent, multiphase displacement in porous media. Soc. Pet. Eng. J. **21**(01), 51–62 (1981)
4. G. Hirasaki et al., Ion exchange with clays in the presence of surfactant. Soc. Pet. Eng. J. **22**(02), 181–192 (1982)

5. E.L. Isaacson, J.B. Temple, Analysis of a singular hyperbolic system of conservation laws. J. Differ. Equ. **65**(2), 250–268 (1986)
6. T. Johansen, A. Tveito, R. Winther, A riemann solver for a two-phase multicomponent process. SIAM J. Sci. Stat. Comput. **10**(5), 846–879 (1989)
7. T. Johansen, R. Winther, The solution of the riemann problem for a hyperbolic system of conservation laws modeling polymer flooding. SIAM J. Math. Anal. **19**(3), 541–566 (1988)
8. T. Johansen, R. Winther, The riemann problem for multicomponent polymer flooding. SIAM J. Math. Anal. **20**(4), 908–929 (1989)
9. B.L. Keyfitz, H.C. Kranzer, A system of non-strictly hyperbolic conservation laws arising in elasticity theory. Arch. Ration. Mech. Anal. **72**(3), 219–241 (1980)
10. W. Lambert, D. Marchesin, The riemann problem for multiphase flows in porous media with mass transfer between phases. J. Hyperbolic Equ. **8**(02), 149–156 (2009)
11. B.J. Merkel, B. Planer-Friedrich, D.K. Nordstrom, *Groundwater Geochemistry* (Springer, 2005)
12. D.L. Parkhurst, C. Appelo, *Description of Input and Examples for PHREEQC Version Computer Program for Speciation, Batch-Reaction, One-Dimensional Transport, and Inverse Geochemical Calculations* (US Geological Survey, Denver, 2013)
13. G. Pope, L. Lake, F. Helfferich et al., Cation exchange in chemical flooding: Part 1-basic theory without dispersion. Soc. Pet. Eng. J. **18**(06), 418–434 (1978)
14. A. Puime. *INDEPENDÊNCIA ENTRE TERMODINÂMICA E HIDRODINÂMICA EM PROCESSOS DE RECUPERACÃO AVANCADA DE PETRÓLEO*. PhD thesis, Universidade Estadual do Norte Fluminense, 2003
15. B. Temple, Global solution of the cauchy problem for a class of $2\times$ 2 nonstrictly hyperbolic conservation laws. Adv. Appl. Math. **3**(3), 335–375 (1982)

Localization of Adiabatic Deformations in Thermoviscoplastic Materials

Min-Gi Lee, Theodoros Katsaounis and Athanasios E. Tzavaras

Abstract We study an instability occurring at high strain-rate deformations, induced by thermal softening properties of metals, and leading to the formation of shear bands. We consider adiabatic shear deformations of thermoviscoplastic materials and establish the existence of a family of focusing self-similar solutions that capture this instability. The self-similar solutions emerge as the net response resulting from the competition between Hadamard instability and viscosity. Their existence is turned into a problem of constructing a heteroclinic orbit for an associated dynamical system, which is achieved with the help of geometric singular perturbation theory.

Keywords Localization · Shear bands · Self similar solutions · Geometric singular perturbations

1 Introduction

Shear bands are regions of intensely localized shear deformation appearing when metals are deforming at high strain rates. This type of material instability has attracted attention in the mechanics and mathematical literature [3, 5, 12–15]. In the mechanics literature, such material instability is often called *Hadamard instability* and is associated with an ill-posed initial value problem. However, it should be noted that although Hadamard instability indicates the catastrophic growth of oscillations around a mean state, coherent localized structures, *the shear bands*, emerge in an

M.-G. Lee (✉) · T. Katsaounis · A. E. Tzavaras
Computer, Electrical and Mathematical Sciences and Engineering Division, KAUST, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
e-mail: mingi.lee@kaust.edu.sa

T. Katsaounis
e-mail: theodoros.katsaounis@kaust.edu.sa

A. E. Tzavaras
e-mail: athanasios.tzavaras@kaust.edu.sa

T. Katsaounis
IACM, FORTH, Heraklion, Greece

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_21

orderly fashion. This is a highly nonlinear phenomenon resulting from the competition between Hadamard instability and viscosity.

Under isothermal conditions, metals, in general, strain harden and exhibit a stable response. As the deformation speed increases, the heat produced by the plastic work causes an increase in the temperature. For certain metals, the tendency for thermal softening may outweigh the tendency for strain hardening and deliver net softening. A destabilizing feedback mechanism is then induced, which operates as follows (see [2]): Nonuniformities in the strain rate result in nonuniform heating. If heat diffusion is too weak to equalize the temperatures, the initial nonuniformities in the strain rate are, in turn, amplified. This mechanism tends to localize the total deformation into narrow regions (shear bands). On the other hand, there is opposition to this process by "viscous effects" induced by strain-rate sensitivity. The outcome of the competition depends mainly on the relative weights of thermal softening, strain hardening, and strain-rate sensitivity, as well as the loading circumstances. This qualitative scenario is widely accepted as the mechanism of shear band formation.

We work with a simple model that captures the mechanism of shear band formation: We consider the adiabatic shear deformation of a thermoviscoplastic material that occupies the slab between two parallel plates. The relevant quantities are the velocity $v(t, x)$, the shear strain $\gamma(t, x)$, the shear strain rate $u(t, x)$, the shear stress $\tau(t, x)$, and the temperature $\theta(t, x)$. The system of equations describing the motion takes the form

$$\begin{aligned} \gamma_t &= u \quad \text{(kinematic compatibility)}, \\ v_t &= \tau_x \quad \text{(momentum conservation)}, \\ \theta_t &= \tau u \quad \text{(energy equation)}, \\ \tau &= \tau(\theta, \gamma, u) \quad \text{(constitutive law)}, \end{aligned} \tag{A}$$

with $(t, x) \in \mathbb{R}^+ \times \mathbb{R}$. In terms of classification, the model (A) belongs to the framework of one-dimensional thermoviscoelasticity. It is also instructive to interpret $(A)_4$ as a constitutive law for thermoviscoplastic materials viewing $\gamma \equiv \gamma_p$ as the plastic strain; see the hierarchy of models in [6, 14]. This context suggests the terminology: The material exhibits thermal softening at (θ, γ, u) when $\tau_\theta(\theta, \gamma, u) < 0$, strain hardening if $\tau_\gamma(\theta, \gamma, u) > 0$, and strain softening if $\tau_\gamma(\theta, \gamma, u) < 0$.

In this study, we focus on a constitutive hypothesis in the form of a *power* law

$$\tau = \varphi(\theta, \gamma)u^n = \theta^{-\alpha}\gamma^m u^n, \tag{1}$$

where n is the strain-rate sensitivity which is assumed to be very small $0 < n \ll 1$, α measures the degree of thermal softening, while m measures the degree of strain hardening. We further introduce two subclasses of (A), where $\varphi(\theta, \gamma)$ is independent of either θ or γ, respectively. These are the strain-independent model (B) $(m = 0)$ consisting of

$$v_t = \tau_x, \quad \theta_t = \tau u, \quad \tau = \mu(\theta)u^n, \tag{B}$$

and the temperature-independent model (C) $(\alpha = 0)$ consisting of

$$\gamma_t = u, \quad v_t = \tau_x, \quad \tau = \varphi(\gamma)u^n. \tag{C}$$

Model (A) with power law (1) admits a special class of solutions describing uniform shearing, which can be written explicitly

$$\gamma_s = t + \gamma_0,\ v_s = x,\ \theta_s = \left[\theta_0^{1+\alpha} + \frac{1+\alpha}{1+m}\left[\gamma_s^{1+m} - \gamma_0^{1+m}\right]\right]^{\frac{1}{1+\alpha}},\ \tau_s = \theta_s^{-\alpha}\gamma_s^m, \tag{2}$$

where $\gamma_0,\ \theta_0$ denote initial values of shear strain and temperature, respectively.

The linear stability analysis [5] around the *uniform shearing solutions* (2) indicates that system (A) becomes unstable in the regime $q := -\alpha + m + n < 0$, while it is asymptotically stable in the complementary region $q > 0$. Using the Chapman–Enskog expansion and a relaxation theory approach, the authors in [6] obtained an effective equation for the shear strain rate that changes type from forward parabolic when $q > 0$ to backward parabolic when $q < 0$.

Nonlinear stability of model (B) ($m = 0$) in the regime $q > 0$ has been studied in [3] when $n = 1$, and in [13] for $n \neq 1$. System (B) has an interpretation that the model describes a fluid with temperature-dependent viscosity $\mu(\theta)$ in the rectlinear shear motion. Similar result for the problem (C) ($\alpha = 0$) in the regime $q > 0$ is obtained in [14]. For the problem (B) in the regime $q < 0$, the failure of the asymptotic stability is treated in [1] when $n = 1$ and for $n \neq 1$ in [6].

The main result of this paper is the construction of a family of self-similar solutions of *focusing type* to models (B) and (C) that capture the underlying instability. It provides a survey of several related results concerning the construction of focusing solutions that have recently appeared in the literature [7, 9, 11]. The exposition is organized mostly around the general system of three Eqs. (A). The last step of the construction of focusing self-similar solutions for the general power law (1) is work in progress [10].

The paper is organized as follows: Sect. 2 contains the main results of this work. The existence of focusing type self-similar solutions for systems (B) and (C) is stated precisely in Theorems 1 and 2. Figure 1a, b, c, and d depicts the typical behavior of such focusing solutions. The existence of focusing self-similar solutions is turned into a problem of constructing a heteroclinic orbit for an associated dynamical system. The main tool of proving the latter is the theory of geometric singular perturbations [4], which is discussed briefly in Sect. 3.

2 Main Results

2.1 Self-similar Structure

We investigate the scale invariance property of the system (A), and consequently that of (B) and (C) too. Suppose $(\gamma, u, v, \theta, \tau)$ is a solution of system (A). Then, a

rescaled version of it $(\gamma_\rho, u_\rho, v_\rho, \theta_\rho, \tau_\rho)$ given by

$$\begin{aligned}
\gamma_\rho(t,x) &= \rho^a \gamma(\rho^{-1}t, \rho^\lambda x), & v_\rho(t,x) &= \rho^b v(\rho^{-1}t, \rho^\lambda x),\\
\theta_\rho(t,x) &= \rho^c \theta(\rho^{-1}t, \rho^\lambda x), & \tau_\rho(t,x) &= \rho^d \tau(\rho^{-1}t, \rho^\lambda x),\\
u_\rho(t,x) &= \rho^{b+\lambda} \gamma(\rho^{-1}t, \rho^\lambda x),
\end{aligned}$$

is also a solution of (A) provided that

$$\begin{aligned}
a &= \frac{2+2\alpha-n}{D} + \frac{2+2\alpha}{D}\lambda =: a_0 + a_1\lambda,\\
b &= \frac{1+m}{D} + \frac{1+m+n}{D}\lambda =: b_0 + b_1\lambda,\\
c &= \frac{2(1+m)}{D} + \frac{2(1+m+n)}{D}\lambda =: c_0 + c_1\lambda,\\
d &= \frac{-2\alpha+2m+n}{D} + \frac{-2\alpha+2m+2n}{D}\lambda =: d_0 + d_1\lambda,
\end{aligned}$$

for each $\lambda \in \mathbb{R}$, where $D = 1 + 2\alpha - m - n$. Motivated by the scale invariance property parametrized by λ, we look for the solutions of the form

$$\begin{aligned}
\gamma(t,x) &= t^a \Gamma(t^\lambda x), & v(t,x) &= t^b V(t^\lambda x), & \theta(t,x) &= t^c \Theta(t^\lambda x),\\
\tau(t,x) &= t^d \Sigma(t^\lambda x), & u(t,x) &= t^{b+\lambda} U(t^\lambda x),
\end{aligned}$$

and set $\xi = t^\lambda x$. In this format, $\lambda > 0$ accounts for the focusing behavior as time increases, whereas $\lambda < 0$ accounts for the de-focusing behavior. This family includes the uniform shearing solution at $\lambda = -\frac{1+m}{2(1+\alpha)}$. Since we are interested in the focusing solutions, we consider $\lambda > 0$ in the rest of the paper.

Using this ansatz to the system (A), we obtain a system of ordinary differential and algebraic equations that $\big(\Gamma(\xi), V(\xi), \Theta(\xi), \Sigma(\xi), U(\xi)\big)$ satisfies

$$\begin{aligned}
a\Gamma(\xi) + \lambda\xi\Gamma'(\xi) &= U(\xi),\\
bV(\xi) + \lambda\xi V'(\xi) &= \Sigma'(\xi),\\
c\Theta(\xi) + \lambda\xi\Theta'(\xi) &= \Sigma(\xi)U(\xi),\\
\Sigma(\xi) &= \Theta(\xi)^{-\alpha}\Gamma(\xi)^m U(\xi)^n,\\
V'(\xi) &= U(\xi).
\end{aligned} \tag{3}$$

2.2 Main Theorem

We first state the existence of two parameters' family of solutions for (B) where $m = 0$. See [10] for the detailed discussion.

Theorem 1. *Let $\alpha, n > 0$, $\alpha \neq 2n+1$ be the given material parameters and fix $U_0 > 0$ and $\Theta_0 > 0$. Suppose that*

$$\frac{2}{1+2\alpha-n} < \frac{U_0^{1+n}}{\Theta_0^{1+\alpha}} < \frac{2}{1+n}, \tag{4}$$

$-\alpha + n < 0$, and n is sufficiently small. Then, there is a focusing self-similar solution to system (B) of the form

$$v(t,x) = (t+1)^b V((t+1)^\lambda x), \quad \theta(t,x) = (t+1)^c \Theta((t+1)^\lambda x),$$
$$\tau(t,x) = (t+1)^d \Sigma((t+1)^\lambda x), \quad u(t,x) = (t+1)^{b+\lambda} U((t+1)^\lambda x)$$

where the focusing rate is

$$\lambda = \frac{1+2\alpha-n}{2+2n}\frac{U_0^{1+n}}{\Theta_0^{1+\alpha}} - \frac{2}{2+2n} > 0. \tag{5}$$

Furthermore, the self-similar profile $\big(V(\xi), \Theta(\xi), \Sigma(\xi), U(\xi)\big)$, $\xi = (t+1)^\lambda x$, has the following properties:

(i) Satisfies the boundary condition at $\xi = 0$,

$$V(0) = \Theta_\xi(0) = \Sigma_\xi(0) = U_\xi(0) = 0, \quad U(0) = U_0, \Theta(0) = \Theta_0.$$

(ii) Its asymptotic behavior as $\xi \to 0$ is given by

$$\begin{aligned}
&\Theta(\xi) = \Theta(0) + \Theta''(0)\frac{\xi^2}{2} + o(\xi^2), && \Theta''(0) < 0,\\
&\Sigma(\xi) = \Theta(0)^{-\alpha}U(0)^n + \Sigma''(0)\frac{\xi^2}{2} + o(\xi^2), && \Sigma''(0) > 0,\\
&U(\xi) = U(0) + U''(0)\frac{\xi^2}{2} + o(\xi^2), && U''(0) < 0,\\
&V(\xi) = U(0)\xi + U''(0)\frac{\xi^3}{6} + o(\xi^3), && U''(0) < 0.
\end{aligned} \tag{6}$$

(iii) Its asymptotic behavior as $\xi \to \infty$ is given by

$$\begin{aligned}
&V(\xi) = O(1), \quad \Theta(\xi) = O(\xi^{-\frac{1+n}{\alpha-n}}),\\
&\Sigma(\xi) = O(\xi), \quad U(\xi) = O(\xi^{-\frac{1+\alpha}{\alpha-n}}).
\end{aligned} \tag{7}$$

In the case of system (C), where $\alpha = 0$, there exists a two-parameter family of solutions; see [9, 11] for the detailed discussion.

Theorem 2. *Let $-1 \le m < 0$ and $n > 0$, $m + n \neq -\frac{1}{2}$ be the given material parameters and fix $U_0 > 0$ and $\Gamma_0 > 0$. Suppose that*

$$\frac{2-n}{1-m-n} < \frac{U_0}{\Gamma_0} < \frac{2-n}{1+m+n},$$

$m + n < 0$, and n is sufficiently small. Then, there is a focusing self-similar solution to system (C) of the form

$$\begin{aligned} \gamma(t,x) &= (t+1)^a \Gamma((t+1)^\lambda x), \quad v(t,x) = (t+1)^b V((t+1)^\lambda x), \\ \tau(t,x) &= (t+1)^d \Sigma((t+1)^\lambda x), \quad u(t,x) = (t+1)^{b+\lambda} U((t+1)^\lambda x), \end{aligned}$$

where the focusing rate is

$$\lambda = \frac{1-m-n}{2}\Big(\frac{U_0}{\Gamma_0} - \frac{2-n}{1-m-n}\Big) > 0. \tag{8}$$

Furthermore, the self-similar profile $\big(V(\xi), \Theta(\xi), \Sigma(\xi), U(\xi)\big)$, $\xi = (t+1)^\lambda x$, has the following properties:

(i) Satisfies the boundary condition at $\xi = 0$,

$$V(0) = \Gamma_\xi(0) = \Sigma_\xi(0) = U_\xi(0) = 0, \quad U(0) = U_0, \Gamma(0) = \Gamma_0.$$

(ii) Its asymptotic behavior as $\xi \to 0$ is given by

$$\begin{aligned} \Gamma(\xi) &= \frac{1}{a}U(0) + \Gamma''(0)\frac{\xi^2}{2} + o(\xi^2), && \Gamma''(0) < 0, \\ \Sigma(\xi) &= \Gamma(0)^m U(0)^n + \Sigma''(0)\frac{\xi^2}{2} + o(\xi^2), && \Sigma''(0) > 0, \\ U(\xi) &= U(0) + U''(0)\frac{\xi^2}{2} + o(\xi^2), && U''(0) < 0, \\ V(\xi) &= U(0)\xi + U''(0)\frac{\xi^3}{6} + o(\xi^3), && U''(0) < 0. \end{aligned} \tag{9}$$

(iii) Its asymptotic behavior as $\xi \to \infty$ is given by

$$\begin{aligned} \Gamma(\xi) &= O(\xi^{\frac{1}{m+n}}), \quad V(\xi) = O(1), \\ \Sigma(\xi) &= O(\xi), \qquad U(\xi) = O(\xi^{\frac{1}{m+n}}). \end{aligned} \tag{10}$$

2.3 Emergence of Localization

We describe the emergence of localization of the family of solutions for system (B) constructed by Theorem 1. The corresponding localized solutions for system (C) constructed by Theorem 2 are similar, thus omitted.

In both cases, we replace $t \leftarrow t+1$,

$$v(t,x) = (t+1)^b V((t+1)^\lambda x), \quad \theta(t,x) = (t+1)^c \Theta((t+1)^\lambda x),$$
$$\tau(t,x) = (t+1)^d \Sigma((t+1)^\lambda x), \quad u(t,x) = (t+1)^{b+\lambda} U((t+1)^\lambda x),$$

so that we interpret

$$\big(V(\xi), \Theta(\xi), \Sigma(\xi), U(\xi)\big) = \big(v(0,x), \theta(0,x), \tau(0,x), u(0,x)\big)|_{x=\xi},$$

the initial states of the self-similar solutions.

- Initial nonuniformities : The profile $\big(V(\xi), \Theta(\xi), \Sigma(\xi), U(\xi)\big)$ is the initial profile of the self-similar solution. $\Theta(\xi)$ and $U(\xi)$ have a small bump at the origin from the asymptotically flat state. The tip sizes at the origin Θ_0 and U_0 are the two parameters that fix the solution. The velocity $V(\xi)$ is an odd function of ξ that connects $-V_\infty$ and V_∞ as ξ spans from $-\infty$ to ∞, where $V_\infty \triangleq \lim_{\xi\to\infty} V(\xi)$. The slope near the origin is slightly steeper, which reflects the initial nonuniformity in the velocity.
- Temperature: The temperature is an increasing function of t for a fixed x. The growth rate at the origin is faster than any other x, which dictates the localization near the origin, see Fig. 1a

$$\theta(t,0) = (1+t)^{\frac{2}{D}+\frac{2+2n}{D}\lambda}\Theta(0), \qquad \theta(t,x) \sim t^{\frac{2}{D}-\frac{(1+n)^2}{D(\alpha-n)}\lambda}|x|^{-\frac{1+\alpha}{\alpha-n}}, \quad \text{as } t\to\infty, x\neq 0.$$

- Strain rate: The growth rate at the origin is faster than the rest of the points, which dictates the localization near the origin, see Fig. 1b

$$u(t,0) = (1+t)^{\frac{1}{D}+\frac{2+2\alpha}{D}\lambda}U(0),$$
$$u(t,x) \sim t^{\frac{1}{D}-\frac{(1+\alpha)(1+n)}{D(\alpha-n)}\lambda}|x|^{-\frac{1+\alpha}{\alpha-n}}, \quad \text{as } t\to\infty, x\neq 0.$$

- Stress: The stress is a decreasing function of t for fixed x. However, the decay rate at the origin is faster than the rest of the points, see Fig. 1d

$$\tau(t,0) = (1+t)^{\frac{-2\alpha+n}{D}+\frac{-2\alpha+2n}{D}\lambda}\Sigma(0),$$
$$\tau(t,x) \sim t^{\frac{-2\alpha+n}{D}+\frac{1+n}{D}\lambda}|x|^{-\frac{1+\alpha}{\alpha-n}}, \quad \text{as } t\to\infty, x\neq 0.$$

Note that the rate of the latter is always less than $-\frac{n}{1+n}$ in the valid range of λ.

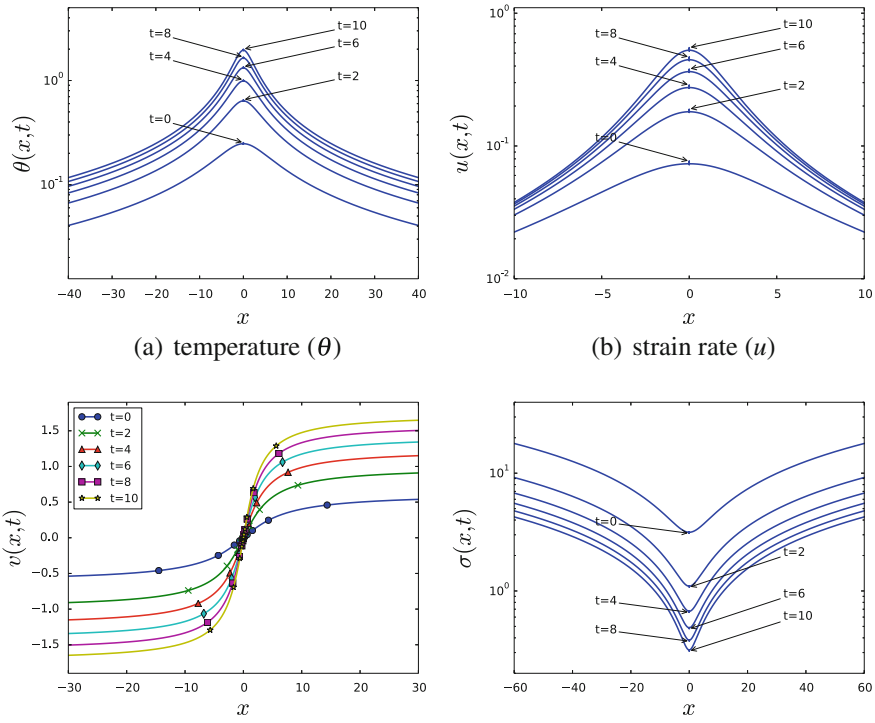


Fig. 1 Localizing solutions for system (B), for $\mu(\theta) = \theta^{-\alpha}$, $\alpha = 1.4$, $n = 0.3$, and $\lambda = 0.39$. All graphs except v are in logarithmic scale. See [11] for the system (C)

- Velocity: The velocity $v(x, t)$ is an odd function of x. It connects $-v_\infty$ to v_∞, as x runs from $-\infty$ to ∞, where $v_\infty \triangleq \lim_{x\to\infty} v(t, x)$. Because of the scaling law $\xi = (1 + t)^\lambda x$, the transition from $-v_\infty$ to v_∞ localizes around the origin as time increases. The slope becomes steeper and steeper and develops a step function-type singularity, see Fig. 1c. The far field velocity

$$v_\infty(t) = (1 + t)^b V_\infty = (1 + t)^{\frac{1}{D} + \frac{1+n}{D}\lambda} V_\infty$$

itself grows at a polynomial rate. This is not in agreement with the uniform shearing motion. This deviation is a consequence of our simplifying assumption for the self-similarity.

3 Existence via Geometric Singular Perturbation Theory

The main step in proving Theorems 1 and 2 is to show the existence of a heteroclinic orbit to an associated dynamical system, which, in the case of system (B), reads as follows:

$$\begin{aligned}
\frac{\dot{p}}{p} &= \Big[\frac{1+\alpha}{1+n}\,\frac{1}{\lambda}\Big(r^{1+n}-c_0\Big)\Big]-\Big[d_1+q+\lambda pr\Big],\\
\frac{\dot{q}}{q} &= \Big[b_1+\frac{bpr}{q}\Big]-\Big[d_1+q+\lambda pr\Big],\\
n\frac{\dot{r}}{r} &= \Big[\frac{\alpha-n}{\lambda(1+n)}\Big(r^{1+n}-c_0\Big)\Big]+\Big[d_1+q+\lambda pr\Big].
\end{aligned} \tag{P}$$

The objective is to construct the heteroclinic orbit that connects equilibrium points

$$M_0=\Big(0,0,\big(\frac{2}{D}+\frac{2(1+n)}{D}\lambda\big)^{\frac{1}{1+n}}\Big),\quad M_1=\Big(0,1,\big(\frac{2}{D}-\frac{(1+n)^2}{D(\alpha-n)}\lambda\big)^{\frac{1}{1+n}}\Big).$$

First, we describe now briefly how system (P) is derived. The technique we employ was mainly developed in [7]. Following [7, 9, 10], we introduce a series of nonlinear transformations described by (11) and (12), while the definition of (p,q,r)-variables is given by (13)

$$\begin{aligned}
&\bar{\gamma}(\xi)=\xi^{a_1}\Gamma(\xi),\quad \bar{v}(\xi)=\xi^{b_1}V(\xi),\qquad \bar{\theta}(\xi)=\xi^{c_1}\Theta(\xi),\\
&\bar{\tau}(\xi)=\xi^{d_1}\Sigma(\xi),\quad \bar{u}(\xi)=\xi^{b_1+1}U(\xi).
\end{aligned} \tag{11}$$

$$\begin{aligned}
&\tilde{\gamma}(\log\xi)=\bar{\gamma}(\xi),\quad \tilde{v}(\log\xi)=\bar{v}(\xi),\quad \tilde{\theta}(\log\xi)=\bar{\theta}(\xi),\\
&\tilde{\tau}(\log\xi)=\bar{\tau}(\xi),\quad \tilde{u}(\log\xi)=\bar{u}(\xi),
\end{aligned} \tag{12}$$

$$p\triangleq\frac{\tilde{\theta}^{\frac{1+\alpha}{1+n}}}{\tilde{\tau}},\quad q\triangleq b\frac{\tilde{v}}{\tilde{\tau}},\quad r\triangleq\frac{\tilde{u}}{\tilde{\theta}^{\frac{1+\alpha}{1+n}}}, \tag{13}$$

with $\eta\triangleq\log\xi$ being the new independent variable $\Big(\frac{df}{d\eta}=\dot{f}\Big)$. The system (P) has a *fast–slow* structure due to parameter n in front of $\dot{r}$. We conduct a Chapman–Enskog-type reduction via geometric singular perturbation theory [4, 8]. The reduced problem becomes a planar dynamical system, and the heteroclinic orbit is obtained by phase-space analysis [10].

3.1 Critical Manifold

The surface the orbit relaxes is near the zero set of the right-hand side of $(\mathrm{P})_3$. The zero set, that is away from $r=0$ plane, is the surface specified by

$$r = \frac{\frac{\alpha c_0}{\lambda} - d_1 - q}{\frac{\alpha}{\lambda} + \lambda p} \triangleq h(p, q; n = 0), \quad \text{or} \quad q + \lambda r p + \frac{\alpha}{\lambda}\Big(r - r_0\Big) = 0.$$

We take the triangle T in the first quadrant enclosed by p-axis, q-axis and the contour line $\underline{r} = h(p, q; n = 0)$ and a compact set $K \supset\supset T$. We set the critical manifold

$$G(\lambda, \alpha, n = 0) \triangleq \left\{ (p, q, r) \mid (p, q) \in K, \text{ and } r = \frac{\frac{\alpha c_0}{\lambda} - d_1 - q}{\frac{\alpha}{\lambda} + \lambda p} \right\}. \tag{14}$$

The system in *fast scale* with the independent variable $\tilde{\eta} = \eta/n$ is

$$\begin{aligned} p' &= np\Big(\Big[\frac{1+\alpha}{1+n}\,\frac{1}{\lambda}\Big(r^{1+n} - c_0\Big)\Big] - \Big[d_1 + q + \lambda p r\Big]\Big), \\ q' &= nq\Big(\Big[b_1 + \frac{bpr}{q}\Big] - \Big[d_1 + q + \lambda p r\Big]\Big), \\ r' &= r\Big(\Big[\frac{\alpha - n}{\lambda(1+n)}\Big(r^{1+n} - c_0\Big)\Big] + \Big[d_1 + q + \lambda p r\Big]\Big), \end{aligned} \tag{$\tilde{P}$}$$

where we denoted $(\cdot)' = \dfrac{d}{d\tilde{\eta}}(\cdot)$. When $n = 0$, $(\tilde{P})|_{n=0}$ reads

$$p' = 0, \quad q' = 0, \quad r' = r\Big(\Big[\frac{\alpha}{\lambda}\Big(r - c_0\Big)\Big] + \Big[d_1 + q + \lambda p r\Big]\Big).$$

Lemma 1. *$G(\lambda, \alpha, 0)$ is a normally hyperbolic invariant manifold with respect to the system $(\tilde{P})|_{n=0}$.*

3.2 Chapman–Enskog-Type Reduction

By the theorem of geometric singular perturbation theory [4, 8], if n is sufficiently small, there exists the locally invariant manifold $G(\lambda, \alpha, n)$ with respect to (P). Then, on this manifold, $\big(p(\eta), q(\eta)\big)$ satisfies the planar system

$$\begin{aligned} \dot{p} &= p\left\{\Big[\frac{1+\alpha}{1+n}\,\frac{1}{\lambda}\Big(h^{1+n} - c_0\Big)\Big] - \Big[d_1 + q + \lambda p h\Big]\right\}, \\ \dot{q} &= q\Big(1 - q - \lambda p h\Big) + bph, \end{aligned} \tag{R}$$

where h stands for $h(p, q; n)$.

3.3 Confinement of the Orbit

Lemma 2. *The triangle T is positively invariant for the system (R) when $n = 0$.*

We can compute the inward normal component of $(\dot{p}, \dot{q})$ on the boundary of the triangle T for $(R)|_{n=0}$:

$$\dot{p} = -\frac{D}{\alpha} p\Big(d_1 + q + \lambda p h\Big),$$
$$\dot{q} = q\Big(1 - q - \lambda p h\Big) + bph.$$

Essential calculation is on the hypotenuse, and the fact that it is the contour line $\underline{r} = h(p, q, n = 0)$ helps us obtain the estimate. Define $\underline{p}$ and $\underline{q}$ to be the p-intercept and q-intercept of the contour line, respectively: $\underline{q} = \lambda \underline{p} \underline{r} = \frac{\alpha}{\lambda}(r_0 - \underline{r})$. With $(-\underline{q}, -\underline{p})$ being the inward normal vector, the inward normal component on the hypotenuse is

$$\begin{aligned}
(\dot{p}, \dot{q}) \cdot (-\underline{q}, -\underline{p}) &= \frac{D}{\alpha} \underline{q}\, p\Big(d_1 + q + \lambda p \underline{r}\Big) - \underline{p}\Big\{q\Big(1 - q - \lambda p \underline{r}\Big) + bp\underline{r}\Big\} \\
&= \frac{D}{\alpha} \underline{q}\, p\Big(d_1 + \underline{q}\Big) - \underline{p}\Big\{\Big(\underline{q} - \lambda p \underline{r}\Big)\Big(1 - \underline{q}\Big) + bp\underline{r}\Big\} \\
&= -\underline{pq}(1 - \underline{q}) + \underline{q}\, p\Big(\frac{D}{\alpha} d_1 + \frac{D}{\alpha}\underline{q} + (1 - \underline{q}) - \frac{b}{\lambda}\Big) \\
&= -\underline{pq}(1 - \underline{q}) + \underline{q}\, p \frac{1 + \alpha}{\lambda}\Big(\frac{1}{1 + \alpha} - \underline{r}\Big) \\
&\geq -\underline{pq}(1 - \underline{q}) \quad \text{for } \underline{r} < \frac{1}{1 + \alpha} \\
&\geq \delta_0 > 0.
\end{aligned}$$

Lemma 3. *The triangle T is positively invariant for the system (R) provided n is sufficiently small.*

Now the hypotenuse is not anymore a contour line of the function $h(p, q; \lambda, \alpha, n)$. We arrange terms of right-hand sides of (R) in the form

$$\begin{aligned}
\dot{p} &= p\Big\{\Big[\frac{1 + \alpha}{\lambda}\Big(\underline{r} - c_0\Big)\Big] - \Big[d_1 + q + \lambda p \underline{r}\Big]\Big\} \\
&\quad + \underbrace{p\Big\{\Big[\frac{1 + \alpha}{1 + n}\,\frac{1}{\lambda}\Big(h^{1+n} - c_0\Big)\Big] - \Big[\frac{1 + \alpha}{\lambda}\Big(\underline{r} - c_0\Big)\Big] - \lambda p(h - \underline{r})\Big]\Big\}}_{\triangleq g_1(p,q,n)}, \\
\dot{q} &= q\Big(1 - q - \lambda p \underline{r}\Big) + bp\underline{r} + \underbrace{(-q\lambda p + b)(h - \underline{r})}_{\triangleq g_2(p,q,n)}.
\end{aligned}$$

Since h is a smooth function of n and D is compact, provided n is sufficiently small, we have an estimate

$$|g_1(p,q,n)| + |g_2(p,q,n)| \le C_0 n, \quad \text{where } C_0 \text{does not depend on } p, q, \text{ and } n.$$

Therefore

$$(\dot{p}, \dot{q}) \cdot (-\underline{q}, -\underline{p}) \ge \delta_0 + C_0' n \quad \text{for another uniform constant } C_0'.$$

For n sufficiently small, the last expression is positive. After having the orbit confined in the positive invariant set T, we further conduct the phase-space analysis to capture the heteroclinic orbit; for details, we refer to [9, 10].

Acknowledgements This research was supported by King Abdullah University of Science and Technology (KAUST).

References

1. M. Bertsch, L. Peletier, S. Verduyn Lunel, The effect of temperature dependent viscosity on shear flow of incompressible fluids. SIAM J. Math. Anal. **22**, 328–343 (1991)
2. R.J. Clifton, J. Duffy, K.A. Hartley, T.G. Shawki, On critical conditions for shear band formation at high strain rates. Scripta Met. **18**, 443–448 (1984)
3. C.M. Dafermos, L. Hsiao, Adiabatic shearing of incompressible fluids with temperature-dependent viscosity. Q. App. Math. **41**, 45–58 (1983)
4. N. Fenichel, Geometric singular perturbation theory for ordinary differential equations. J. Differ. Equ. **31**, 53–98 (1979)
5. C. Fressengeas, A. Molinari, Instability and localization of plastic flow in shear at high strain rates. J. Mech. Phys. Solids **35**, 185–211 (1987)
6. Th. Katsaounis, A.E. Tzavaras, Effective equations for localization and shear band formation. SIAM J. Appl. Math. **69**, 1618–1643 (2009)
7. Th. Katsaounis, J. Olivier, A.E. Tzavaras, Emergence of coherent localized structures in shear deformations of temperature dependent fluids. Arch. Ration. Mech. Anal. https://doi.org/10.1007/s00205-016-1071-2. (online)
8. C. Kuehn, *Multiple Time Scale Dynamics*. Applied Mathematical Sciences, vol. 191 (Springer, Basel, 2015)
9. M-G. Lee, A.E. Tzavaras, Existence of localizing solutions in plasticity via the geometric singular perturbation theory. SIAM J. Appl. Dyn. Syst. arXiv:1608.00198. (to appear)
10. M-G. Lee, Th. Katsaounis, A.E. Tzavaras, Localization and the formation of shear bands in thermoviscoplastic deformations. (in preparation)
11. Th. Katsaounis, M.-G. Lee, A.E. Tzavaras, Localization in inelastic rate dependent shearing deformations. J. Mech. Phys. Solids **98**, 106–125 (2017)
12. T.G. Shawki, R.J. Clifton, Shear band formation in thermal viscoplastic materials. Mech. Mater. **8**, 13–43 (1989)
13. A.E. Tzavaras, Shearing of materials exhibiting thermal softening or temperature dependent viscosity. Q. Appl. Math. **44**, 1–12 (1986)
14. A.E. Tzavaras, Nonlinear analysis techniques for shear band formation at high strain-rates. Appl. Mech. Rev. **45**, S82–S94 (1992)
15. C. Zener, J.H. Hollomon, Effect of strain rate upon plastic flow of steel. J. Appl. Phys. **15**, 22–32 (1944)

The Global Nonlinear Stability of Minkowski Spacetime for Self-gravitating Massive Fields

Philippe G. LeFloch

Abstract We address the global evolution problem for the Einstein equations of general relativity and investigate the global geometry of matter spacetimes that are initially close to Minkowski spacetime. First, we provide a review the equations of Einstein's gravity and then f(R)-gravity. We present their relationship and, next, the wave-Klein-Gordon formalism. Finally, we discuss our new statements of nonlinear stability of massive fields.

Keywords Einstein equations · Massive field · Global existence · Minkowski spacetime · Modified gravity theory

1 Introduction

This is a short review of the series of papers [18–20] which, in collaboration with Yue Ma, establish several novel existence results for systems of coupled-wave-Klein-Gordon equation. Our method—the Hyperbolic Hyperboloidal Method—has allowed us to address the global evolution problem for the Einstein equations of general relativity and investigate the global geometry of *matter spacetimes* that are initially close to Minkowski spacetime. The Einstein equations (when expressed in wave gauge) take the form of nonlinear system of partial differential equations of hyperbolic type and, in presence of self-gravitating massive matter fields, involve a strong coupling between wave equations (for the geometry) and Klein–Gordon equations (for the matter fields). Our method also provides a global existence theory for the field equations of the $f(R)$-theory of gravity, which is a natural generalization of Einstein's gravity theory (see below).

The global nonlinear stability problem for Minkowski spacetime is formulated from initial data which are prescribed on a spacelike hypersurface and are a small

P. G. LeFloch (✉)
Laboratoire Jacques-Louis Lions, Centre National de la Recherche Scientifique, Sorbonne Université, 4 Place Jussieu, 75258 Paris, France
e-mail: contact@philippelefloch.org

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_22

perturbation of an asymptotically flat slice in Minkowski space. This problem is equivalent to a global-in-time existence problem for a system of nonlinear wave equations with sufficiently small data in a weighted Sobolev space.

There are several major challenges to be overcome. Gravitational waves are perturbations propagating in the curved spacetime and may relate to either the Weyl curvature (in vacuum spacetimes) or the Ricci curvature (in matter spacetimes). It is required to understand the effect of nonlinear wave interactions on the possible growth of the energy, in order to be able to exclude dynamical instabilities and self-gravitating massive modes and, therefore, to avoid gravitational collapse (trapped surfaces, black holes)—a generic phenomenon in general relativity [5, 23].

The global dynamics is particularly complex in general, but *sufficiently small* perturbations in Minkowski spacetime are expected to disperse in timelike directions and an asymptotic convergence to Minkowski spacetime to be observed. More precisely, in order to prove its stability, it is necessary to establish that the spacetime is future timelike geodesically complete.

We will begin by reviewing Einstein's gravity and the f(R)-gravity theory and explain their relation. Next, the wave-Klein-Gordon formalism will be presented, and finally the global nonlinear stability will be stated. For further reading, we refer to the works by Donninger and Zenginoglu [7], Fajman et al. [8], Wang [24], and Zenginoglu [25].

2 Self-gravitating Massive Fields

For simplicity in the presentation, the theory is developed for a massive scalar field, whereas our method should generalize to other massive models and, for instance, one should be able to also analyze the coupling between the Einstein equations with massive Yang–Mills fields.

Throughout, we are interested in Lorentzian manifolds (or spacetime) $(M, g_{\alpha\beta})$ with signature $(-,+,+,+)$ and in local coordinates we write $g = g_{\alpha\beta}dx^\alpha dx^\beta$. For instance, Minkowski spacetime $M = \mathbb{R}^{3+1}$ is described in standard coordinates by $g_M = -(dx^0)^2 + \sum_{a=1}^3 (dx^a)^2$. In a spacetime, the covariant derivative operator allows to write schematically $\nabla_\alpha X = \partial_\alpha X + \Gamma \star X$ with $\Gamma \simeq \partial g$. (The exact expressions in coordinates will be given only later in this text.) The Ricci curvature also reads schematically $R_{\alpha\beta} = \partial^2 g + \partial g \star \partial g$, and one also defines the scalar curvature $R := R^\alpha_\alpha = g^{\alpha\beta}R_{\alpha\beta}$ by taking the trace of the Ricci curvature. Here, $\alpha, \beta = 0, 1, 2, 3$ and, whenever relevant, Einstein's summation convention on repeated indices is in order.

The Einstein equations for self-gravitating matter have the form

$$G_{\alpha\beta} = 8\pi T_{\alpha\beta}, \tag{1}$$

in which $G_{\alpha\beta} := R_{\alpha\beta} - (R/2)g_{\alpha\beta}$ is the Einstein curvature tensor and $T_{\alpha\beta}$ denotes the energy-momentum of the matter. A (minimally coupled) massive scalar field with

potential $U(\phi)$, for instance with the quadratic potential $U(\phi) = \frac{c^2}{2}\phi^2$, is described by the energy-momentum tensor

$$T_{\alpha\beta} := \nabla_\alpha\phi\nabla_\beta\phi - \Big(\frac{1}{2}g^{\alpha'\beta'}\nabla_{\alpha'}\phi\nabla_{\beta'}\phi + U(\phi)\Big)g_{\alpha\beta}. \tag{2}$$

Consequently, the Einstein–Klein–Gordon system for the unknown $(M, g_{\alpha\beta}, \phi)$ reads

$$\begin{aligned} R_{\alpha\beta} - 8\pi\Big(\nabla_\alpha\phi\nabla_\beta\phi + U(\phi)\, g_{\alpha\beta}\Big) &= 0, \\ \Box_g\phi - U'(\phi) &= 0, \end{aligned} \tag{3}$$

where $\Box_g = \nabla_\alpha\nabla^\alpha$ denotes the wave operator associated with the unknown metric g. The above equation is a system of geometric PDE's, which enjoys a gauge invariance property.

On the other hand, the field equations for the $f(R)$-theory of modified gravity are based on the following generalized Hilbert–Einstein functional:

$$\int_M \Big(f(R) + 16\pi L[\phi, g]\Big)\, dV_g, \tag{4}$$

in which the nonlinear function $f(R) = R + \frac{\kappa}{2}R^2 + \kappa^2\mathcal{O}(R^3)$ is prescribed with $\kappa > 0$. The condition $\kappa := f''(0) > 0$ will be essential for global stability. This theory has a long history in physics, beginning with Weyl (1918), Pauli (1919), Eddington (1924), and many others [2, 3]. We emphasize that alternative theories of gravity are relevant in view of recent observational data, which have demonstrated the accelerated expansion of the universe and have identified instabilities in galaxies in our universe. The $f(R)$-theory allows for the gravitation field to be mediated by an additional field without explicitly introducing a notion of "dark matter."

Numerical evidence and physical heuristics have led to the conjecture that asymptotically flat, matter spacetimes should be stable [22], even though the existence of a family of "oscillating soliton stars" had first suggested a possible instability mechanism within small perturbations of massive fields. Advanced numerical methods were necessary to handle the long-time evolution of oscillating soliton stars. During an initial phase, the matter *tends to collapse*, but during an intermediate phase (below a certain threshold in the mass density) the *collapse slows down*, until finally the *dispersion* becomes of the main feature of the evolution of the matter field.

Observe that in asymptotically anti-Sitter (AdS) spacetimes, such instabilities are observed [11] and the effect of gravity is dominant so that generic (even arbitrarily small) initial data lead to black hole formation. In AdS spacetime, the matter is confined and cannot disperse: The timelike boundary is reached in finite proper time.

3 The Wave-Klein-Gordon Formulation

Unless we introduce a specific gauge, the field equations $G_{\alpha\beta} = 8\pi T_{\alpha\beta}$ in coordinates take the form of a second-order system with no specific PDE type. By imposing the wave gauge $\Box_g x^\gamma = 0$, we find

$$2\, g^{\alpha\beta}\, \partial_\beta g_{\alpha\gamma} - g^{\alpha\beta}\, \partial_\gamma g_{\alpha\beta} = 0, \qquad \gamma = 0, \dots, 3, \tag{5}$$

leading us to an expression of the Ricci curvature $R_{\alpha\beta} \simeq \Box_g g_{\alpha\beta}$ (after Einstein, Choquet-Bruhat, De Turck, etc.). The Einstein-massive field system is then equivalent to a second-order system of 11 nonlinear wave-Klein-Gordon equations, supplemented with the Hamiltonian-momentum Einstein's constraints. Namely, in wave gauge, the Einstein equations for a self-gravitating massive field read

$$\begin{aligned} \widetilde{\Box}_g g_{\alpha\beta} &= F_{\alpha\beta}(g, \partial g) - 8\pi \left(2\partial_\alpha\phi\partial_\beta\phi + c^2\phi^2\, g_{\alpha\beta}\right), \\ \widetilde{\Box}_g \phi - c^2\phi &= 0, \end{aligned} \tag{6}$$

with $\widetilde{\Box}_g \psi := g^{\alpha'\beta'}\partial_{\alpha'}\partial_{\beta'}\psi$.

The expression of the quadratic nonlinearities $F_{\alpha\beta}(g, \partial g)$ is given in the following lemma and involves null terms of the general form $g^{\alpha\beta}\partial_\alpha u \partial_\beta u$ or $\partial_\alpha u \partial_\beta v - \partial_\beta u \partial_\alpha v$, as well as terms that do not satisfy the null condition and require a specific analysis. Recall first that

$$\begin{aligned} R_{\alpha\beta} &= \partial_\lambda \Gamma^\lambda_{\alpha\beta} - \partial_\alpha \Gamma^\lambda_{\beta\lambda} + \Gamma^\lambda_{\alpha\beta}\Gamma^\delta_{\lambda\delta} - \Gamma^\lambda_{\alpha\delta}\Gamma^\delta_{\beta\lambda}, \\ \Gamma^\lambda_{\alpha\beta} &= \frac{1}{2} g^{\lambda\lambda'}\left(\partial_\alpha g_{\beta\lambda'} + \partial_\beta g_{\alpha\lambda'} - \partial_{\lambda'} g_{\alpha\beta}\right). \end{aligned}$$

Following Lindblad and Rodnianski [21] for vacuum Einstein spacetime, we have the following result.

Lemma 1. *With $F_{\alpha\beta} = Q_{\alpha\beta} + P_{\alpha\beta}$, the Ricci curvature in wave gauge reads*

$$2\, R_{\alpha\beta} = -\widetilde{\Box}_g g_{\alpha\beta} + Q_{\alpha\beta} + P_{\alpha\beta},$$

which contains:

1. *null terms satisfying Klainerman's null condition (and enjoying good decay in time)*

$$\begin{aligned} Q_{\alpha\beta} := {} & g^{\lambda\lambda'} g^{\delta\delta'} \partial_\delta g_{\alpha\lambda'} \partial_{\delta'} g_{\beta\lambda} \\ & - g^{\lambda\lambda'} g^{\delta\delta'} \left(\partial_\delta g_{\alpha\lambda'} \partial_\lambda g_{\beta\delta'} - \partial_\delta g_{\beta\delta'} \partial_\lambda g_{\alpha\lambda'}\right) \\ & + g^{\lambda\lambda'} g^{\delta\delta'} \left(\partial_\alpha g_{\lambda'\delta'} \partial_\delta g_{\lambda\beta} - \partial_\alpha g_{\lambda\beta} \partial_\delta g_{\lambda'\delta'}\right) + \dots, \end{aligned} \tag{7}$$

2. *and quasi-null terms (as they are called by the authors)*

$$P_{\alpha\beta} := -\frac{1}{2} g^{\lambda\lambda'} g^{\delta\delta'} \partial_\alpha g_{\delta\lambda'} \partial_\beta g_{\lambda\delta'} + \frac{1}{4} g^{\delta\delta'} g^{\lambda\lambda'} \partial_\beta g_{\delta\delta'} \partial_\alpha g_{\lambda\lambda'} \tag{8}$$

(which will require a further investigation based on the wave gauge condition).

A similar decomposition can be written for the conformal metric of the f(R)-theory of gravity, which we now introduce. The modified gravity equations read

$$N_{\alpha\beta} = 8\pi T_{\alpha\beta} \tag{9}$$

and are based on a choice of a function $f(R) = R + \frac{\kappa}{2} R^2 + \ldots$, and take the form of a fourth-order system with no specific PDE type. We propose to rely on an augmented formulation with unknown $({g^\dagger}_{\alpha\beta}, \rho)$ defined as follows, by regarding the spacetime curvature as an independent unknown and by working with the conformal metric

$${g^\dagger}_{\alpha\beta} := f'(R_g) g_{\alpha\beta}, \tag{10}$$

in which $\rho := \frac{1}{\kappa} \ln f'(R_g)$. In view of the standard relation between the Ricci curvature tensors of g and $g^\dagger$, i.e.,

$${R^\dagger}_{\alpha\beta} = R_{\alpha\beta} - 2\big(\nabla_\alpha \nabla_\beta \rho - \nabla_\alpha \rho \nabla_\beta \rho\big) - \big(\Box_g \rho + 2g(\nabla\rho, \nabla\rho)\big) g_{\alpha\beta},$$

we arrive at a third-order system. In addition, from the trace of the field equation, we derive an evolution equation for the scalar curvature which is a new degree of freedom in the theory and must be supplemented with suitable initial data. Finally, in wave coordinates

$$\Box_{g^\dagger} x^\alpha = 0 \tag{11}$$

we arrive at a second-order system of 12 nonlinear wave-Klein-Gordon equations. The structure is analogous to the one of the Einstein-massive field system, but has a significantly more involved algebraic structure and admits additional constraints.

Proposition 1. *The equations of f(R)-gravity for a self-gravitating massive field take the form*

$$\begin{aligned}
\widetilde{\Box}_{g^\dagger} {g^\dagger}_{\alpha\beta} &= F_{\alpha\beta}(g^\dagger, \partial g^\dagger) - 8\pi \left(2e^{-\kappa\rho} \partial_\alpha\phi \partial_\beta\phi + c^2\phi^2 e^{-2\kappa\rho} {g^\dagger}_{\alpha\beta}\right) \\
&\quad - 3\kappa^2 \partial_\alpha\rho \partial_\beta\rho + \kappa\, \mathcal{O}(\rho^2) {g^\dagger}_{\alpha\beta}, \\
\widetilde{\Box}_{g^\dagger}\phi - c^2\phi &= c^2\big(e^{-\kappa\rho} - 1\big)\phi + \kappa g^{\dagger\alpha\beta} \partial_\alpha\phi \partial_\beta\rho, \\
3\kappa\, \widetilde{\Box}_{g^\dagger}\rho - \rho &= \kappa\, \mathcal{O}(\rho^2) - 8\pi\Big(g^{\dagger\alpha\beta} \partial_\alpha\phi \partial_\beta\phi + \frac{c^2}{2} e^{-\kappa\rho}\phi^2\Big),
\end{aligned} \tag{12}$$

supplemented with:

1. *the wave gauge conditions* $g^{\dagger\alpha\beta} {\Gamma^\dagger}^\lambda_{\alpha\beta} = 0$,
2. *the curvature compatibility condition* $e^{\kappa\rho} = f'(R_{e^{-\kappa\rho} g^\dagger})$,

3. and the Hamiltonian and momentum constraints.

These three sets of conditions can be propagated from a Cauchy hypersurface.

Proposition 2. *In the limit* $\kappa \to 0$ *one finds*

$$g^{\dagger} \to g \quad \text{and} \quad \rho \to 8\pi\big(g^{\alpha\beta}\nabla_\alpha\phi\nabla_\beta\phi + \frac{c^2}{2}\phi^2\big), \tag{13}$$

and the Einstein system for a self-gravitating massive field is (6).

This completes the formulation of the field equations in a PDE form, and the geometric problem of interest can be reformulated as a global existence problem for coupled nonlinear wave equations. Our main challenge is that the system is *not invariant by scaling*, and one must rely on *fewer symmetries* in, for instance, defining weighted energy-like functionals. The analysis of coupled wave equations and Klein-Gordon equations is particularly challenging and drastically *different time asymptotic behavior* arise for the unknown components of the system: $O(t^{-1})$ for wave equations and $O(t^{-3/2})$ for Klein–Gordon equations.

We also need to investigate the dependence in f and determine the *singular limit* $f(R) \to R$ which, as we can see, transforms a second-order PDE into an algebraic equation.

4 The Global Nonlinear Stability

For the global existence theory, we need to establish that there is a sufficient rate of time decay for all the nonlinearities of interest. The nonlinear coupling between the geometry and massive matter leads to strong interactions at the PDE level and, consequently, it is necessary to be able to establish (almost) sharp L^2 and L^∞ time decay for the metric and matter field. Understanding the quasi-null structure of the Einstein equations is fundamental since the standard null condition is violated and an amplification phenomenon arise for the energy.

We thus consider the initial value problem for the Einstein equations (and its generalization). An initial data set, by definition, provides us with the geometry of the initial hypersurface $(M_0 \simeq \mathbb{R}^3, g_0, k_0)$ and the initial data for the matter field ϕ_0, ϕ_1. We assume that these data are sufficiently close to a spacelike, asymptotically flat slice in Minkowski spacetime. The local existence is standard and goes back to Choquet-Bruhat [4]: to each initial data set, one can associate a unique maximal, globally hyperbolic Cauchy development (i.e., intuitively, the maximal part of the spacetime which is uniquely determined by the prescribed initial data and remains smooth).

The fundamental work on the stability of Minkowski spacetime for the vacuum Einstein equations (or massless matter) was done by Christodoulou and Klainerman [6] (and later generalized in [1]):

1. They introduced a fully geometric proof, in which the Bianchi identities are regarded as the main evolution equations,
2. they analyzed the geometry of null cones and defined a double null-maximal foliation,
3. and they relied on all of the Killing fields of Minkowski spacetime.

The earlier work by Friedrich [9] also addressed the global existence problem for the vacuum Einstein equations and established the nonlinear stability of De Sitter spacetime. More recently, Lindblad and Rodnianski [21] obtained the first global existence result for the vacuum Einstein equations in wave coordinates (despite an "instability" result by Choquet-Bruhat) and again relied on all of the Killing fields of Minkowski spacetime. They introduced a foliation by asymptotically flat hypersurfaces.

In contrast, the recent work [18–20] addresses this stability problem for *self-gravitating massive matter fields*:

1. The proposed new method (the Hyperboloidal Foliation Method) does not rely on Minkowski's scaling field $r\partial_r + t\partial_t$,
2. which is the key of be able to tackle massive matter fields,
3. is based on an asymptotically hyperbolic foliation,
4. and leads to a somewhat simpler proof for the case of massless fields.

The positive mass theorem restricts the possible behavior of solutions at spacelike infinity. No solution can be exactly Minkowski "at infinity," but can coincide with the Schwarzschild metric outside a spatially compact region. More generally, solutions are assumed to approach the Schwarzschild metric near space infinity (with ADM mass $m \ll 1$). We only provide here informal statements of our results, and we refer to [18–20] for the precise statements.

Theorem 1 (Nonlinear stability of Minkowski spacetime with self-gravitating massive fields). *Consider the Einstein-massive field system when the initial data set* $(M_0 \simeq \mathbb{R}^3, g_0, k_0, \phi_0, \phi_1)$ *is asymptotically Schwarzschild and sufficiently close to Minkowski data and satisfies the Einstein constraint equations. Then, the initial value problem*

1. *admits a globally hyperbolic Cauchy development,*
2. *which is foliated by asymptotically hyperbolic hypersurfaces,*
3. *and is future causally geodesically complete and asymptotically approaches Minkowski spacetime.*

Theorem 2 (Nonlinear stability of Minkowski spacetime in f(R)-gravity). *Consider the field equations of* $f(R)$*-modified gravity when the initial data set* $(M_0 \simeq \mathbb{R}^3, g_0, k_0, R_0, R_1, \phi_0, \phi_1)$ *is asymptotically Schwarzschild and sufficiently close to Minkowski data and satisfies the constraint equations of modified gravity. Then, the initial value problem*

1. *admits a globally hyperbolic Cauchy development,*

2. *which is foliated by asymptotically hyperbolic hypersurfaces,*
3. *and is future causally geodesically complete and asymptotically approaches Minkowski spacetime.*

The limit problem $\kappa \to 0$ can be viewed as a relaxation phenomenon for the spacetime scalar curvature. We pass from the second-order wave equation

$$3\kappa\, \widetilde{\Box}_{g^\dagger}\rho - \rho = \kappa\, \mathcal{O}(\rho^2) - 8\pi\Big(g^{\dagger\alpha\beta}\partial_\alpha\phi\partial_\beta\phi + \frac{c^2}{2}\, e^{-\kappa\rho}\phi^2\Big) \tag{14}$$

to the purely algebraic equation

$$\rho \to 8\pi\big(g^{\alpha\beta}\nabla_\alpha\phi\nabla_\beta\phi + \frac{c^2}{2}\phi^2\big). \tag{15}$$

Theorem 3 (f(R)-spacetimes converge toward Einstein spacetimes). *In the limit $\kappa \to 0$, when the nonlinear function $f = f(R)$ (the integrand in the Hilbert–Einstein action) approaches the scalar curvature function R, the Cauchy developments of modified gravity (given in the previous theorem) converge (in every bounded time interval, in a sense specified quantitatively in Sobolev norms) to Cauchy developments of Einstein's gravity theory.*

The proofs rely on weighted norms associated with the asymptotically hyperboloidal foliation which we construct. Our energy norms are solely based on the translations ∂_α and the Lorentzian boosts L_a of Minkowski spacetime. These fields enjoy good commutator properties even in curved space and allow us to decompose the wave operators, the metric, etc. On each hyperboloidal hypersurface $\mathcal{H}^n[s]$ at any hyperboloidal time s, in wave coordinates, we use the boosts to define the norm

$$\big(\|u\|_{\mathcal{H}^n[s]}\big)^2 := \sup_{a=1,2,3} \sum_{|J|\leq n} \int_{\mathcal{H}_s \simeq \mathbb{R}^3} |L_a^J u|^2\, dx \tag{16}$$

and, within the spacetime, we use the translations to define the norm

$$\|u\|_{\mathcal{H}^N[s_0,s_1]} := \sup_{s\in[s_0,s_1]} \sum_{|I|+n\leq N} \big\|\partial^I u\big\|_{\mathcal{H}^n[s]}. \tag{17}$$

We introduce a suitable bootstrap argument, which shows that the total contribution of the interaction terms contributes only a finite amount to the growth of the total energy. We derive global time-integrability properties for the source terms, which are established from sharp pointwise estimates—required to handle the strong geometry–matter interactions under consideration. Sobolev inequalities and Hardy inequalities are adapted to the hyperboloidal foliation, and a hierarchy of energy bounds distinguishes between various orders of differentiation and growth rates in the hyperboloidal time s.

References

1. L. Bieri, N. Zipser, *Extensions of the Stability Theorem of the Minkowski Space in General Relativity*, vol. 45, AMS/IP Studies in Advanced Mathematics, American Mathematical Society (International Press, Cambridge, 2009)
2. C. Brans, R.H. Dicke, Mach principle and a relativistic theory of gravitation. Phys. Rev. **124**, 925–935 (1961)
3. H.A. Buchdahl, Non-linear Lagrangians and cosmological theory. Monthly Notices R. Astr. Soc. **150**, 1–8 (1070)
4. Y. Choquet-Bruhat, *General Relativity and the Einstein Equations*, Oxford Mathematical Monograph (Oxford University Press, Oxford, 2009)
5. D. Christodoulou, *The Formation of Black Holes in General Relativity*. European Mathematical Society (EMS) Series (Zurich, 2008)
6. D. Christodoulou, S. Klainerman, *The Global Nonlinear Stability of the Minkowski Space*, vol. 41, Princeton Mathematical Series (1993)
7. R. Donninger, A. Zenginoglu, Nondispersive decay for the cubic wave equation. Anal. PDE **7**, 461–495 (2014)
8. D. Fajman, J. Joudioux, J. Smulevici, A vector field method for relativistic transport equations with applications, ArXiv:1510.04939
9. H. Friedrich, Cauchy problems for the conformal vacuum field equations in general relativity. Commun. Math. Phys. **91**, 445–472 (1983)
10. L. Hörmander, *Lectures on Nonlinear Hyperbolic Differential Equations* (Springer, Berlin, 1997)
11. G.T. Horowitz, J.E. Santos, Geons and the instability of Anti-de Sitter spacetime, ArXiv:1408.5906
12. S. Katayama, Asymptotic pointwise behavior for systems of semilinear wave equations in three space dimensions. J. Hyperbolic Differ. Equ. **9**, 263–323 (2012)
13. S. Katayama, Global existence for coupled systems of nonlinear wave and Klein-Gordon equations in three space dimensions. Mathematische Zeitschrift **270**, 487–513 (2012)
14. S. Klainerman, Global existence for nonlinear wave equations. Commun. Pure Appl. Math. **33**, 43–101 (1980)
15. S. Klainerman, Global existence of small amplitude solutions to nonlinear Klein-Gordon equations in four spacetime dimensions. Commun. Pure Appl. Math. **38**, 631–641 (1985)
16. P.G. LeFloch, *An introduction to self-gravitating matter, Graduate course given at Institute Henri Poincaré* (Paris, Fall, 2015), http://www.youtube.com/user/PoincareInstitute
17. P.G. LeFloch, *An Introduction to the Einstein-Euler Equations* (in preparation)
18. P.G. LeFloch, Y. Ma, The global nonlinear stability of Minkowski space for self-gravitating massive fields. The wave-Klein-Gordon model. Commun. Math. Phys. **346**, 603–665 (2016)
19. P.G. LeFloch and Y. Ma, The global nonlinear stability of Minkowski space for self-gravitating massive fields. Compact Schwarzschild perturbations, ArXiv:1511.03324
20. P.G. LeFloch, Y. Ma, *The Global Nonlinear Stability of Minkowski Space for Self-gravitating Massive Fields* (in preparation)
21. H. Lindblad, I. Rodnianski, The global stability of Minkowski spacetime in harmonic gauge. Ann. Math. **171**, 1401–1477 (2010)
22. H. Okawa, V. Cardoso, P. Pani, Collapse of self-interacting fields in asymptotically flat spacetimes: do self-interactions render Minkowski spacetime unstable? Phys. Rev. D **89**, 041502 (2014)
23. R.M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984)
24. Q. Wang, An intrinsic hyperboloid approach for Einstein Klein-Gordon equations, ArXiv:1607.01466
25. A. Zenginoglu, Hyperboloidal evolution with the Einstein equations. Class. Quantum Grav. **25**, 195025 (2008)

A Particle-Based Multiscale Solver for Compressible Liquid–Vapor Flow

Jim Magiera and Christian Rohde

Abstract To describe complex flow systems accurately, it is in many cases important to account for the properties of fluid flows on a microscopic scale. In this work, we focus on the description of liquid–vapor flow with a sharp interface between the phases. The local phase dynamics at the interface can be interpreted as a Riemann problem for which we develop a multiscale solver in the spirit of the heterogeneous multiscale method (HMM) [7], using a particle-based microscale model to augment the macroscopic two-phase flow system. The application of a microscale model makes it possible to use the intrinsic properties of the fluid at the microscale, instead of formulating (ad hoc) constitutive relations.

Keywords Multiscale modeling · Heterogeneous multiscale method
Conservation laws · Compressible two-phase flow · Liquid–vapor flow
Sharp interface resolution · Riemann problem · Particle chain model
Model reduction · Machine learning

1 Introduction

For many problems in science and engineering, microscopic properties can heavily influence the macroscopic behavior. Therefore, it is important to consider microscopic effects in the mathematical model development. The obvious possibility to account for such small-scale effects is to solve the microscopic model everywhere. However, despite advances in computing power over the last decades, it is usually still not feasible. This scenario applies to the case of compressible fluid flows with liquid–vapor phase transition. Most applications require a computational domain on

J. Magiera (✉) · C. Rohde
Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Stuttgart, Germany
e-mail: jim.magiera@mathematik.uni-stuttgart.de

C. Rohde
e-mail: crohde@mathematik.uni-stuttgart.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_23

a laboratory scale, which is many orders of magnitude apart from a truly microscopic model that considers effects on the molecular level.

One approach to this problem is to perform multiscale domain-decomposition of micro- and macroscale models, where in a part of the domain a microscale particle model is solved instead of the macroscale model, and both models are coupled via suitable boundary conditions. This coupling approach has been investigated, for example, in [13] for the incompressible Navier–Stokes equations on the macroscale and a Lennard–Jones particle model as the microscale model. In [11], multiscale domain-decomposition is applied for crack propagation in brittle materials, where a set of conservation laws is used in the continuum domain and near the crack a microscale particle model is applied. Furthermore, the phase change of a liquid on a hot plate has been examined in [5].

In this work, however, we propose a multiscale model for the description of single liquid droplets, based on the heterogeneous multiscale method (HMM) [6, 7], which is a general framework for developing multiscale models. The main idea behind it is to compute solutions of a microscopic model for some given macroscopic constraints and propagate hereby obtained parameters to the macroscopic model. Consequently, instead of performing multiscale domain-decomposition coupling of the scales, a data-based approach is promoted.

2 The Macroscale Model: Compressible, Isothermal Euler Equations

On the macroscopic scale, we consider the behavior of a single liquid droplet in a vapor atmosphere. For such two-phase flows, it is possible to consider either a diffuse interface approach [2], where the phase boundary has a finite thickness, or a sharp interface approach, as in [14, 19], where a discontinuous transition between the phases is present. In this work, we follow the second approach and assume that the interface between the phases is represented as a discontinuous shock wave.

Furthermore, we assume that the fluid flow is compressible, inviscid, and isothermal at reference temperature T_{ref}, such that the dynamics are described by the isothermal Euler equations

$$\partial_t \rho + \nabla \cdot (\rho v) = 0, \qquad \partial_t (\rho v) + \nabla \cdot (\rho v \otimes v) + \nabla p(1/\rho) = 0, \tag{1}$$

for the density ρ and velocity v in the space–time domain $\Omega \times (0, T)$, with $T > 0$ and $\Omega \subset \mathbb{R}^d$ an open set.

To describe the two separate phases, we distinguish at each point of time $t \in [0, T]$, between the two distinct bulk phases $\Omega_{\mathrm{vap}}(t)$ and $\Omega_{\mathrm{liq}}(t)$ with common boundary/interface $\Gamma(t)$, such that $\Omega_{\mathrm{vap}}(t) \cup \Omega_{\mathrm{liq}}(t) \cup \Gamma(t) = \Omega$. Figure 1 shows a sketch of this setting. To close the system (1), the pressure p has to be specified. For describing a generic two-phase system, we consider the van der Waals pressure function, in

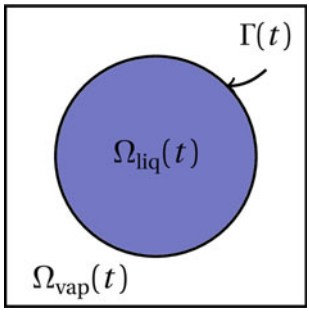


Fig. 1 Sketch of the two-phase flow domains

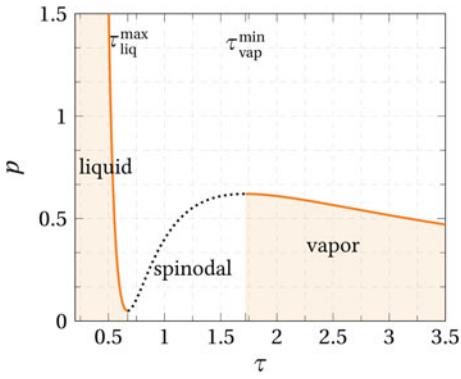


Fig. 2 The van der Waals pressure function for $T_{\text{ref}} < T_{\text{c}}$

terms of the specific volume $\tau = \frac{1}{\rho}$, as in [14],

$$p(\tau) = \frac{RT_{\text{ref}}}{\tau - b} - \frac{1}{\tau^2}, \tag{2}$$

with some constants $R, b, a > 0$. If the temperature T_{ref} is greater than the critical temperature $T_{\text{c}} = \frac{8a}{27Rb}$, the van der Waals pressure function is monotone and the system (1) is hyperbolic. However, if $T_{\text{ref}} < T_{\text{c}}$, the pressure is non-monotone, and the system becomes elliptic for $\tau \in (\tau_{\text{liq}}^{\max}, \tau_{\text{vap}}^{\min})$ which is called spinodal region. Thus, we define the admissible set of densities as $\mathscr{A}_{\text{vdw}} := (b, \infty) \setminus (\tau_{\text{liq}}^{\max}, \tau_{\text{vap}}^{\min})$ and distinguish between the liquid phase for $\tau \in (b, \tau_{\text{liq}}^{\max})$ and the vapor phase for $\tau \in (\tau_{\text{vap}}^{\min}, \infty)$ (Fig. 2).

In order to complete the two-phase model, we have to formulate, besides initial and boundary conditions, some additional coupling conditions at the interface $\Gamma(t)$. Therefore, let $\xi \in \Gamma(t)$ and $t \in [0, T)$ be fixed. The speed of the interface $\Gamma(t)$ in normal direction $\mathbf{n}(\xi, t) \in \mathbb{S}^{d-1}$ (always pointing into the vapor phase) is denoted by $s(\xi, t) \in \mathbb{R}$. Then the mass and momentum balance at the interface, neglecting surface tension, take the following form

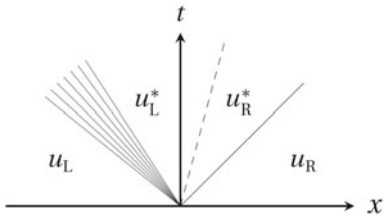


Fig. 3 Sketch of a wave pattern for two-phase flow. The dashed line indicates the phase transition, which is sharp as an additional discontinuous wave

$$\begin{aligned} [\![\rho(v \cdot \mathbf{n} - s)]\!] &= 0, \\ [\![\rho(v \cdot \mathbf{n} - s)v \cdot \mathbf{n} + p(1/\rho)]\!] &= 0, \\ [\![v \cdot \mathbf{t}]\!] &= 0, \quad \forall \mathbf{t} \perp \mathbf{n}, \end{aligned} \tag{3}$$

where $[\![\cdot]\!]$ denotes the difference between liquid and vapor phase values. The well-posedness of the free boundary value problem requires still another coupling condition. For the relevant subsonic case, one assumes that this condition can be written down as an algebraic equation, called kinetic relation. It describes the entropy dissipation at the interface [16].

For given initial Riemann data $u_L = (\rho, \rho v)_L$ for $x \leq 0$, and $u_R = (\rho, \rho v)_R$ for $x > 0$, the solution of the initial value problem (1) evolves (in contrast to the one-phase case) as a 3-wave pattern—a sketch of such a wave pattern is depicted in Fig. 3.

Two-phase models with kinetic relations have been investigated in detail; see, for example, [1, 3, 12].

However, it can be seen that for certain settings, the wrong choice of the kinetic relation can lead to a behavior of the model that is not observed by physical experiments, see, e.g., [19]. For that reason, we want to return to a more elementary notion of the physical properties and regard the flow at the interface on a molecular level. This has the advantage that no kinetic relation is needed. Furthermore, most physical parameters on the molecular level can be determined accurately by experiments. These advantages become even more apparent if one considers non-isothermal multiphase flow and mixtures, where the physically correct choice of the kinetic relation is usually not clear.

3 The Microscale Model: Particle Chain Model

For the description of the liquid–vapor interaction of droplets on a microscopic scale, we apply an atomistic one-dimensional particle chain model, which has been investigated for example in [8]. More precisely that means that we consider a one-dimensional system of N particles with position $x_i = x_i(t)$, velocity $v_i = v_i(t)$, and mass m_i, for $i = 1, \ldots, N$. The distance between the ith and $(i + 1)$th particle is given by $r_{i,i+1} = |x_{i+1} - x_i|$; see Fig. 4. The particles are assumed to interact only

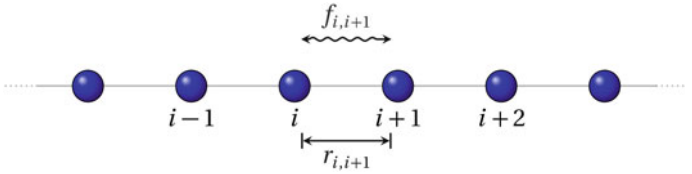


Fig. 4 Sketch of the particle chain model

with direct neighbors via a potential $\phi\colon \mathbb{R}^+ \to \mathbb{R} : r \mapsto \phi(r)$, where r denotes the distance between the particles. The ith particle is subject to the forces $f_{i-1,i}$, $f_{i,i+1}$ originating from the potentials of the neighboring particles, and the resulting force f_i is therefore given by

$$f_i = f_{i-1,i} + f_{i,i+1} = \phi'(|x_{i-1} - x_i|) - \phi'(|x_{i+1} - x_i|).$$

Consequently, the acceleration $a_i = a_i(t)$ of the ith particle is given by $a_i = f_i/m_i$. For the boundary conditions, we assume that f_0 and f_N are zero. This gives us the following ordinary initial value problem for the particle motion

$$\tfrac{\mathrm{d}^2}{\mathrm{d}t^2} x_i(t) = \tfrac{1}{m_i} f_i(t), \qquad x_i(0) = x_i^0, \qquad v_i(0) = v_i^0, \tag{4}$$

with initial positions x_i^0 and velocities v_i^0 for $i = 1, \ldots, N$.

3.1 Micro-/Macroscale Conversion: Irving–Kirkwood Formulas

To design a multiscale scheme that accounts for microscopic properties, it is essential to convert the key quantities from the macroscopic to the microscopic scale and vice versa. In case of a particle model, this can be achieved via the Irving–Kirkwood formulas [9]. The microscopic instantaneous density $\widehat{\rho}(x,t)$ and momentum $(\widehat{\rho v})(x,t)$ distributions are realized by

$$\widehat{\rho}(x,t) = \sum_{i=1}^{N} m_i\, \delta(x - x_i(t)), \quad (\widehat{\rho v})(x,t) = \sum_{i=1}^{N} m_i\, v_i(t)\, \delta(x - x_i(t)), \tag{5}$$

where m_i, x_i, v_i are the mass, position, and velocity of the ith particle and δ denotes the Dirac distribution. Employing momentum balance for the integrated quantities, the instantaneous pressure distribution $\widehat{p}(x,t)$ computes as

$$\widehat{p}(x,t) = \frac{1}{d}\left(\sum_{i=1}^{N}(m_i\overline{v}_i \cdot \overline{v}_i)\,\delta(x - x_i(t)) + \sum_{\substack{i=1,\dots,N\\ j<i}}(f_{ij}\cdot r_{ij})\,\lambda_{ij}(x,t)\right). \quad (6)$$

Here, $\overline{v}_i$ denotes the relative velocity with respect to a local mean value, $r_{ij}(t) := (x_i(t) - x_j(t))$. The pressure distribution consists of a local kinetic part and a non-local contribution originating from the pair-interactions between the particles. Following [9], the interaction term is localized by averaging along the straight path between the ith and jth particles, i.e., $\lambda_{ij}(x,t)$ is defined by

$$\lambda_{ij}(x,t) := \int_0^1 \delta\big(x - (x_j(t) + \lambda\,(x_i(t) - x_j(t)))\big)\mathrm{d}\lambda.$$

To get averaged quantities that can be passed to the macroscopic model, we have to average the distributions $\widehat{\rho}$, $\widehat{v}$ and $\widehat{p}$ over a sampling domain. Consequently, we obtain the spatially averaged, microscopic quantities ρ, v, and p. In the following, we will only consider these averaged quantities.

For a homogeneous particle chain with constant particle masses $m = m_i$, the averaged microscopic pressure is given by $p(\tau) = -\phi'(\tau)$, as a function of the specific volume $\tau = m/\rho$, if the local microscopic temperature is zero, which is the case in our setting, as the particles are initialized without any random fluctuations. Using this relation, the macroscopic pressure function can be determined directly from the microscale model. This means that for the consistency of both models we have to set $\phi(\tau) = \psi(\tau)$, where ψ denotes the specific Helmholtz free energy of the macroscopic system, satisfying $p(\tau) = -\psi'(\tau)$. In the following, we consider the potential

$$\phi(r) = -\frac{a}{r} - R\theta\,\ln(b-r), \qquad \phi'(r) = \frac{a}{r^2} + \frac{R\theta}{b-r}, \quad (7)$$

which is consistent with the van der Waals pressure (2). However, we stress that the choice of the potential is arbitrary and implies the macroscopic pressure, not the other way round. Here, the explicit choice of ϕ is done to compare the multiscale scheme with already existing solvers for van der Waals fluids.

3.2 The Microscopic Riemann Problem

Our main goal is to describe the dynamics of the fluid at the liquid–vapor interface, which we interpret as a Riemann problem. To incorporate microscopic properties, we define a Riemann problem on the microscopic scale and solve it in order to extract the wave pattern, which will be used to compute the fluxes at the interface on the macroscopic scale.

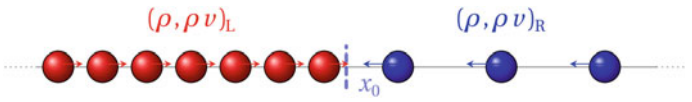


Fig. 5 Schematic representation of Riemann data at the microscopic scale

Therefore, we have to convert the macroscopic quantities to the microscale quantities and vice versa using the Irving–Kirkwood formulas (5). To be more precise, for macroscopic Riemann problem data $u_{\mathrm{L}} = (\rho, \rho v)_{\mathrm{L}}$ and $u_{\mathrm{R}} = (\rho, \rho v)_{\mathrm{R}}$ we set the initial particle configuration uniformly, such that for both $\alpha = \mathrm{L}$ and $\alpha = \mathrm{R}$

$$x_i^0 - x_{i-1}^0 = m_i \rho_\alpha^{-1}, \qquad v_i^0 = v_\alpha, \qquad \text{for all } i \in I_\alpha,$$

holds, where $I_{\mathrm{L}} = \{i \,|\, i = 1, \ldots, N \text{ with } x_i \leq 0\}$, $I_{\mathrm{R}} = \{i \,|\, i = 1, \ldots, N \text{ with } x_i > 0\}$ are the index sets for the left-/right-hand particles. A schematic depiction of such a configuration can be seen in Fig. 5. This gives us the microscopic Riemann problem for

$$(\rho, \rho v)(x,\, t = 0) = \begin{cases} (\rho, \rho v)_{\mathrm{L}} & : \; x \leq 0, \\ (\rho, \rho v)_{\mathrm{R}} & : \; x > 0, \end{cases} \tag{8}$$

with a left state $(\rho, \rho v)_{\mathrm{L}}$ and a right state $(\rho, \rho v)_{\mathrm{R}}$, defined by local averages of (5), with the jump at zero.

After running the microscale simulation, the evolving wave pattern has to be transferred to the macroscopic model. For that we perform some local averaging over the particles states using the Irving–Kirkwood formulas (5). The interface speed is obtained by tracking the interface position on the microscopic scale.

Remark: Via (6), a macroscopic pressure $p = \widehat{p}(\rho^{-1})$ is defined, which is by $p(\tau) = -\phi'(\tau)$ consistent with the microscopic interaction potential ϕ and exactly leads to (2). This function is monotone increasing in the spinodal region $\mathscr{A}_{\mathrm{vdw}}$. States in the spinodal region lead to instabilities in the numerical scheme and must be avoided. In our numerical examples, we never experienced such problems, but we cannot prove that the proposed averaging process excludes spinodal states. Neither we can guarantee in general that the overall numerical method will not lead to spinodal values, even if the microscale Riemann solver avoids them (except for some special cases, see [4]).

3.2.1 Extracting Key Quantities

For a given solution to the microscopic Riemann problem, it is still the question how we extract the key quantities from the microscopic solution.

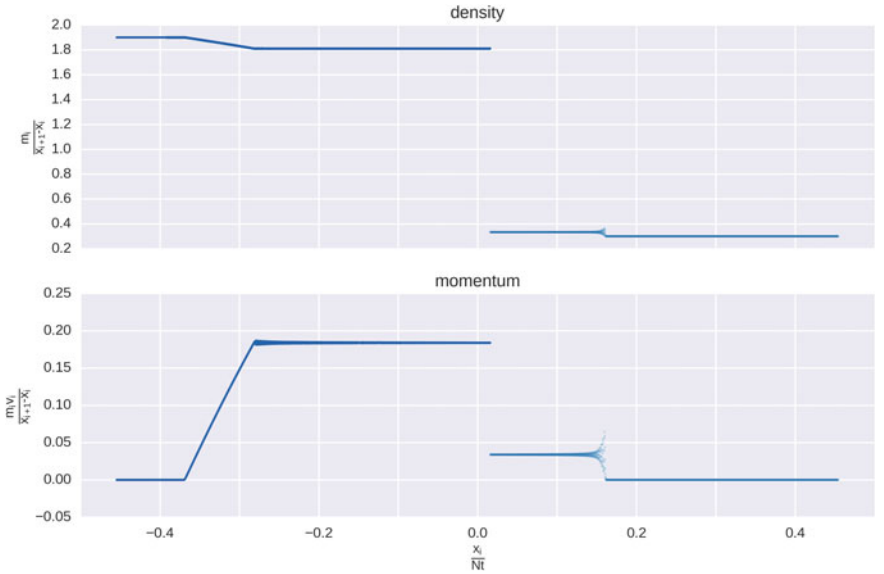


Fig. 6 Example of a solution of the microscopic particle model with van der Waals potential (7) for the initial values $(\rho, \rho v)_{\mathrm{L}} = (1.9, 0)$ and $(\rho, \rho v)_{\mathrm{R}} = (0.3, 0)$ for 16000 particles at $t = 2500$. The phase boundary is located at the density jump near the origin

In Fig. 6, an example of a solution of the particle model is depicted. It can be seen that, similar to wave patterns in the continuum case, a 3-wave pattern evolves—see Fig. 3. We apply this analogy to construct a numerical flux for the interface dynamics. To this end, similar to the numerical flux in [4], we need to extract the states adjacent to the interface from the wave pattern, and also the interface propagation speed. To obtain these values, the interface is tracked by considering the biggest local change in density and then the neighboring states can be computed easily by local averaging left and right of the interface.

3.3 Discretization of the Particle System

For the time-discretization of the particle system, we apply the velocity Verlet algorithm [17]. It is an explicit scheme with microscale time step $\Delta t > 0$ of the following form:

$$\begin{aligned} x(t + \Delta t) &= x(t) + \Delta t\, v(t) + \tfrac{1}{2}\Delta t^2\, a(t), \\ v(t + \Delta t) &= v(t) + \tfrac{1}{2}\Delta t \left(a(t) + a(t + \Delta t)\right), \end{aligned} \tag{9}$$

where $v = \frac{\mathrm{d}x}{\mathrm{d}t}$ is the particle velocity, and $a = \frac{\mathrm{d}v}{\mathrm{d}t}$ the particle acceleration, computed from the forces between the particles at each time step. It is of second order and has the advantage that no intermediate values of x, v, or a have to be stored. Furthermore, we see that all steps can be run in parallel. This enables us to run the particle simulations on a graphics processing unit (GPU) which gives a major speedup, as opposed to conventional hardware.

4 The Multiscale Model

To design the multiscale model, we consider the continuum model (1) with the interface conditions (3) as our macroscopic model. The bulk phases of the continuum model are solved by a standard finite volume scheme, and we focus on the description of the interface dynamics. We refrain from formulating a kinetic relation, and instead include data from the microscopic Riemann solutions of the particle model presented in Sect. 3.2. Hereby, the communication between the macroscale continuum model and microscale particle model is solely data-driven. Only the macroscopic constraints $(\rho, \rho v)_{\mathrm{L}}$ and $(\rho, \rho v)_{\mathrm{R}}$ are needed for setting up the microscale Riemann problem, and in return, for the computation of the macroscale interface flux just the response values $(s, u_{\mathrm{L}}^*, u_{\mathrm{R}}^*)$ from the wave pattern are needed; see Sect. 4.2. Consequently, for the continuum model only the input-output relation $(u_{\mathrm{L}}, u_{\mathrm{R}}) \mapsto (s, u_{\mathrm{L}}^*, u_{\mathrm{R}}^*)$ from the microscopic Riemann problem is important.

4.1 Model Reduction Algorithm

The evaluation of the microscale model is computationally relatively expensive, and if it is evaluated at each interface edge and time step of the continuum model, the coupled micro-/macroscale model becomes computationally unfeasible—see Sect. 5 for more details. To counter this problem, we exploit the fact that the coupling is solely data-driven, and apply a reduced, kernel-based surrogate model for the particle model input-response relation $f_{\mathrm{micro}} \colon (u_{\mathrm{L}}, u_{\mathrm{R}}) \mapsto (s, u_{\mathrm{L}}^*, u_{\mathrm{R}}^*)$, where $u := (\rho, \rho v)$. More abstractly, we apply the microscale model as a black box and put the reduced model into the framework of machine learning. For that $x \in \mathbb{R}^{d_1}$ denotes the d_1-dimensional input data, which is in our case $x = (u_{\mathrm{L}}, u_{\mathrm{R}})$, and $y \in \mathbb{R}^{d_2}$ is the d_2-dimensional response of our model, in our case the measured data $(s, u_{\mathrm{L}}^*, u_{\mathrm{R}}^*)$. The aim is now, to train a regression function from samples $D_n = \{(x_i, y_i) : i = 1, \ldots, n\}$, obtained from observations $y_i = f_{\mathrm{micro}}(x_i) + \varepsilon_{\mathrm{s}}$ that describes f_{micro} in an optimal sense. Here ε_{s} accounts for possible normal distributed measurement noise. To get the regression function from the sample set D_n, we apply a support vector regression scheme; see, e.g., [15]. Therefore, we have to train the reduced model function $f(x) = \sum_{i=1}^{n} \alpha_i \, k_\gamma(x_i, x)$, on the trainings data set D_n, where k_γ is the radial basis kernel function $k_\gamma(x_i, x) = \exp(-\gamma \|x - x_i\|^2)$. In this context, that means that we

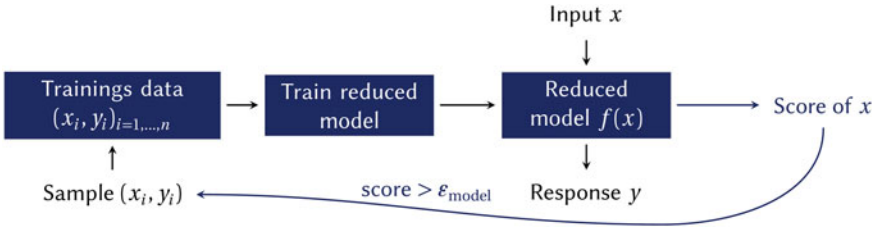


Fig. 7 Sketch of the model reduction scheme with dynamic sampling

have to determine the coefficients $\alpha_i \in \mathbb{R}$ such that f describes f_{micro} optimally under the observations in D_n. Consequently, an optimization problem has to be solved each time the reduced model is trained.

More details on kernel-based surrogate modeling can be found in, e.g., [10, 18].

4.1.1 Dynamic Sampling Scheme

In our case, the input values that are needed cannot be prescribed a priori. Therefore, we apply a dynamic sampling strategy, which is described in this section.

The sampling set D_n is updated dynamically at each time step of the continuum model. To this end, we assign each input value $x \in \mathbb{R}^{d_1}$ a score $\gamma(x;\, D_n)$ that describes the quality of the surrogate model at the point x. This score is computed at each evaluation of the surrogate model. If the score is below a certain threshold $\varepsilon_{\text{model}} > 0$, we simply evaluate the point x by the surrogate model. On the other hand, if it is above the threshold, we draw a new sample by evaluating the microscale model and add it to the training set $D_{n+1} = D_n \cup \{(x_{n+1},\, y_{n+1})\}$. A sketch of the complete model reduction scheme is shown in Fig. 7.

In the following, we use the distance from an input value x to the nearest point of the sample data set D_n, i.e., $\gamma(x\,;\, D_n) = \min_{i \le n} \|x - x_i\|$. One drawback of this simple choice is that we only consider the input values and ignore the output values, which could give an indication whether the (local) variance of the underlying model equation is higher or lower in certain areas of the input space.

4.2 Numerical Discretization of the Multiscale Model

To discretize the macroscale model, we apply the time-explicit front tracking finite volume scheme for systems from [4]. It has the advantage that the sharp interface is resolved within the mesh, i.e., the discretized phase boundary always coincides with a (moving) mesh edge. At the interface, we have to solve a special Riemann problem including the phase dynamics. From its solution, we have to extract the interface propagation speed s and the adjacent fluid states u_{R}^* and u_{L}^*; see Fig. 3.

However, instead of solving the microscale Riemann problem each time, we insert the model reduction scheme from Sect. 4.1. The wave pattern values are inserted in the numerical flux at the interface $g(u_\mathrm{L}, u_\mathrm{R}) = \frac{1}{2}\left(f(u_\mathrm{L}^*) + f(u_\mathrm{R}^*) - s(u_\mathrm{L}^* + u_\mathrm{R}^*)\right)$. In the bulk phases, we apply a standard Lax–Friedrichs flux scheme.

5 Numerical Simulations

In this section, we present some numerical simulation results to show that the multiscale scheme is viable and applicable to (two-dimensional) droplet dynamics.

A Multiscale Simulation of the Riemann Problem: The first simulation results show the consistency between the particle model and the multiscale model in one spatial dimension. Therefore, we run both the particle model and the multiscale model for the same set of Riemann data and compare the averaged particle solution with the multiscale solution. For the initial conditions, we have $\rho_\mathrm{L} = 2.0$, $v_\mathrm{L} = 0$ for $x < 0$ in the liquid phase, and on the right side the vapor-phase Maxwell equilibrium state $\rho_\mathrm{R} \approx 0.317$, $v_\mathrm{R} = 0$. In Fig. 8, both solutions are superimposed and we can see that they fit well, and in particular the wave speeds of the phase boundary coincide.

Multiscale Simulations of a Droplet in 2D: Next, we solve the multiscale model on the continuum scale in two spatial dimensions.

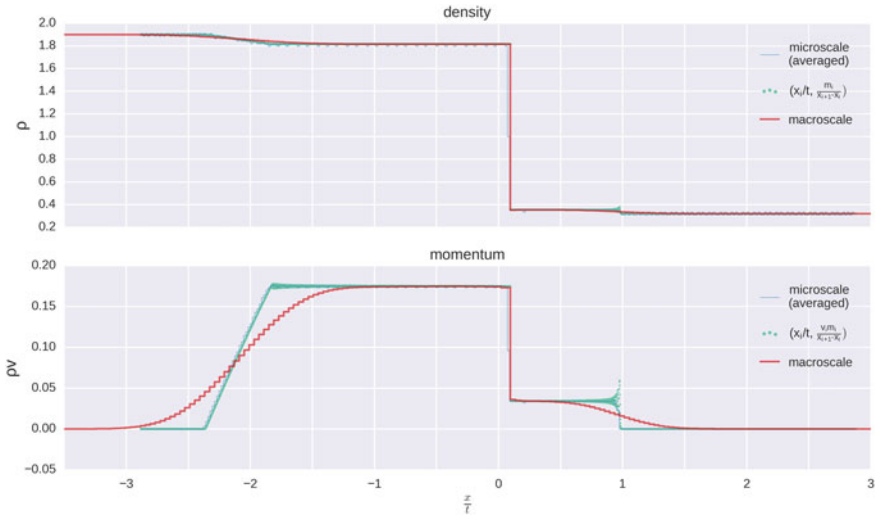


Fig. 8 One-dimensional solution of the multiscale model and the particle model for the Riemann problem, where the phase boundary is located at the density jump near $x/t = 0.2$

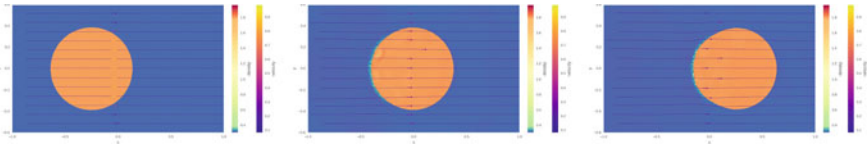


Fig. 9 Multiscale simulation of a moving droplet at $t = 0$, $t = 1.25$ and $t = 2.5$ (from left to right)

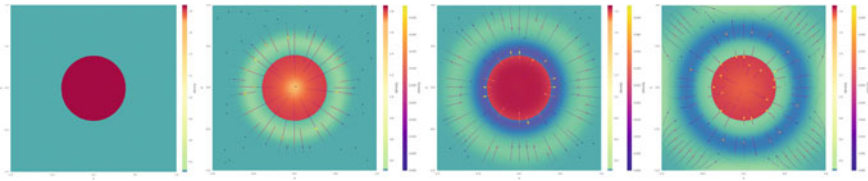


Fig. 10 Multiscale simulation of an oscillating droplet at $t = 0, 0.25, 0.5, 0.75$ (from left to right)

Droplet transport: In the first simulation, we present the performance of the front tracking scheme in two spatial dimensions. The initial conditions for the density are the Maxwell equilibrium states, which are $\rho_{\text{liq}} \approx 1.804$ for the liquid phase and $\rho_{\text{vap}} \approx 0.317$ for the vapor phase. The initial velocity in the domain and on the boundary is set to $v = (0.2, 0)^\top$. In Fig. 9, we see that the droplet is transported through the domain and mostly keeps its shape. Furthermore, it remains in equilibrium, and the increased density at the interface in the vapor on the left side and the small oscillations are due to the local averaging if the triangulation is restructured.

Oscillating droplet: In the next simulation, we consider a droplet that is perturbed from the liquid phase equilibrium, i.e., $\rho_{\text{liq}} = 1.85$, and measure the effect of the model tolerance $\varepsilon_{\text{model}}$ on the computational time. The simulation results for $\varepsilon_{\text{model}} = 0.5$ are presented in Fig. 10. We consider reflecting boundary conditions and thus, the droplet oscillates slightly. The computational time[1] for this simulation is depicted in Fig. 11. It can be seen that the time for computing new samples is of the same order of the finite volume computations, which underlines the performance of the model reduction scheme. If we would not apply model reduction, we would have to run microscale simulation (around 20 s per sample) for all 8000 time steps at each of the ∼160 interface edges. This would lead to a computational time that amounts to roughly one year. Comparing to that the runtime with the model reduction scheme adds up to only several minutes. This gives us huge speedups (with/without model reduction) as shown in Table 1.

[1] All simulations were performed on a single workstation equipped with an Intel® i7-6700 CPU at 3.4 GHz, 16GB RAM, and a Nvidia® GTX980 Ti GPU.

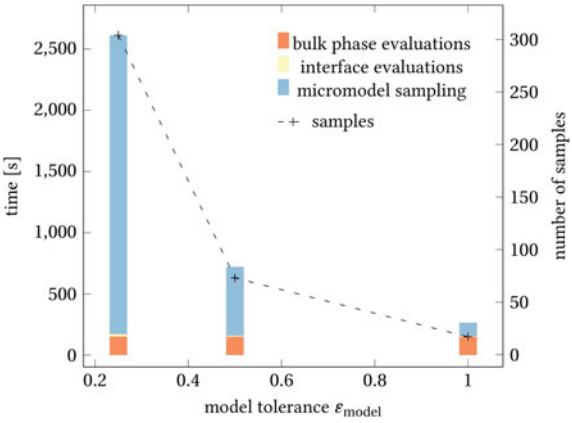


Fig. 11 Computational time in seconds with respect to the model tolerance $\varepsilon_{\text{model}}$. The dashed line indicates the number of samples that are drawn from the microscale model

Table 1 Speedup, with/without model reduction

$\varepsilon_{\text{model}}$	Speedup
1.0	222112
0.5	47000
0.25	10836

6 Conclusions

We have presented a multiscale model for the description of two-phase flows with a sharp interface that incorporates microscale features originating from an atomistic particle model. We have exploited the fact that the coupling of the micro- and macroscale model is solely data-based and developed a model reduction scheme that dynamically samples new data from the microscale model and makes the whole multiscale scheme computationally feasible. Numerical simulation results are presented, that not only show the consistency of the scheme, but also that the applicability in more complex situations without prescribing some (ad hoc) kinetic relations.

Acknowledgements The work was supported by the German Research Foundation (DFG) through SFB TRR 75 "Droplet dynamics under extreme ambient conditions."

References

1. R. Abeyaratne, J.K. Knowles, Kinetic relations and the propagation of phase boundaries in solids. Arch. Ration. Mech. Anal. **114**(2) (1991)
2. D.M. Anderson, G.B. McFadden, A.A. Wheeler, Diffuse-interface methods in fluid mechanics. Annu. Rev. Fluid Mech. **30**(1) (1998)
3. N. Bedjaoui, C. Chalons, F. Coquel, P.G. Lefloch, Non-monotonic traveling waves in van der waals fluids. Anal. Appl. **03**(04) (2005)

4. C. Chalons, C. Rohde, M. Wiebe, A finite volume method for undercompressive shock waves in two space dimensions. ESAIM: M2AN **51**(5), 1987–2015 (2017). https://doi.org/10.1051/m2an/2017027
5. I.A. Cosden, A hybrid atomistic-continuum model for liquid-vapor phase change. Ph.D. thesis, University of Pennsylvania, 2013
6. W. E, *Principles of Multiscale Modeling* (Cambridge University Press, 2011)
7. W. E, B. Engquist, X. Li, W. Ren, E. Vanden-Eijnden, Heterogeneous multiscale methods: a review. Commun. Comput. Phys. **2**(3) (2007)
8. M. Herrmann, J.D.M. Rademacher, Riemann solvers and undercompressive shocks of convex FPU chains. Nonlinearity **23**(2) (2010)
9. J.H. Irving, J.G. Kirkwood, The statistical mechanical theory of transport processes. iv. the equations of hydrodynamics. J. Chem. Phys. **18**(6) (1950)
10. F. Kissling, C. Rohde, The computation of nonclassical shock waves in porous media with a heterogeneous multiscale method: the multidimensional case. Multiscale Model. Simul. **13**(4) (2015)
11. X. Li, J.Z. Yang, W. E, A multiscale coupling method for the modeling of dynamics of solids with application to brittle cracks. J. Comput. Phys. **229**(10) (2010)
12. C. Merkle, C. Rohde, The sharp-interface approach for fluids with phase change: Riemann problems and ghost fluid techniques. ESAIM: M2AN **41**(6) (2007)
13. W. Ren, Analytical and numerical study of coupled atomistic-continuum methods for fluids. J. Comput. Phys. **227**(2) (2007)
14. C. Rohde, C. Zeiler, A relaxation Riemann solver for compressible two-phase flow with phase transition and surface tension. Appl. Numer. Math. **95** (2015)
15. I. Steinwart, A. Christmann, *Support Vector Machines* (Springer, 2008)
16. L. Truskinovsky, Kinks versus shocks, in *Shock Induced Transitions and Phase Structures in General Media*, IMA Vol. Math. Appl. vol. 52 (Springer, New York, 1993)
17. L. Verlet, Computer "experiments" on classical fluids. i. thermodynamical properties of Lennard-Jones molecules. Phys. Rev. **159**, 98–103 (1967)
18. D. Wirtz, N. Karajan, B. Haasdonk, Surrogate modeling of multiscale models using kernel methods. Int. J. Numer. Methods Eng. **101**(1) (2015)
19. C. Zeiler, Liquid vapor phase transitions: modeling, Riemann solvers and computation. Ph.D. thesis, Universität Stuttgart (2015)

L^p-L^q Decay Estimates for Dissipative Linear Hyperbolic Systems in 1D

Corrado Mascia and Thinh Tien Nguyen

Abstract Given $A,\ B \in \mathbb{R}^{n\times n}$, we consider the Cauchy problem for partially dissipative hyperbolic systems having the form

$$\partial_t u + A\partial_x u + Bu = 0,$$

with the aim of providing a detailed description of the large-time behavior. Sharp L^p-L^q estimates are established for the distance between the solution to the system and a time-asymptotic profile, where the profile includes a solution to a parabolic system and a solution of a hyperbolic system. The key tools for the proof are the Fourier transform together with the Young inequality and the interpolation inequality.

Keywords Large-time behavior · Dissipative linear hyperbolic systems Asymptotic expansions

1 Introduction

In this framework, we consider the Cauchy problem for partially dissipative linear hyperbolic systems in one-dimensional space, namely

$$\partial_t u + A\partial_x u + Bu = 0, \qquad u(x,0) = u_0(x), \tag{1}$$

where $A,\ B \in \mathbb{R}^{n\times n}$, and $u \in \mathbb{R}^n$ under general and reasonable assumptions on the coefficients. Decay estimates for (1) have been established for years as in [1, 5].

C. Mascia
Dipartimento di Matematica "G. Castelnuovo", Sapienza—Università di Roma,
P.le Aldo Moro, 2, 00185 Roma, Italy
e-mail: mascia@mat.uniroma1.it

T. T. Nguyen (✉)
Department of Mathematics, Gran Sasso Science Institute—Istituto Nazionale
di Fisica Nucleare, Viale Francesco Crispi, 7, 67100 L'Aquila, Italy
e-mail: nguyen.tienthinh@gssi.infn.it

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_24

Consider as an example the Goldstein–Kac model for chemotaxis

$$\begin{cases} \partial_t u_1 - \partial_x u_1 = -\frac{1}{2}u_1 + \frac{1}{2}u_2, \\ \partial_t u_2 + \partial_x u_2 = \frac{1}{2}u_1 - \frac{1}{2}u_2. \end{cases} \tag{2}$$

It can be easily checked that $w := u_1 + u_2$ solves the linear damped wave equation

$$\partial_{tt} w + \partial_t w - \partial_{xx} w = 0. \tag{3}$$

A detailed description of the asymptotic behavior of (3) was explored in [3] via L^p-L^q *estimates*. More precisely, for any initial datum $(w, \partial_t w)|_{t=0} = (w_0, w_1)$, there are associated time-dependent functions ϕ solution to a heat equation and $\psi(x,t) := e^{-t/2}[w_0(x+t) + w_0(x-t)]/2$ such that for any $1 \le q \le p \le \infty$, the error estimate

$$\|w - \phi - \psi\|_{L^p} \le C\, t^{-\frac{1}{2}(\frac{1}{q}-\frac{1}{p})-1} \|(w_0, w_1)\|_{L^q} \quad \forall t \ge 1$$

holds. This result is remarkable in that the asymptotic profile is more descriptive (compared to [1, 5]), the decay rate is sharp, and p, q are arbitrary in $[1, \infty]$.

Our aim in this paper is to establish this L^p-L^q-type estimate for the general system (1). Applications include, for example, linearized systems arising in the Broadwell model for the Boltzmann equation, the Goldstein–Kac system (2), and its generalization considered in [4]. Having in mind these applications, we impose the following structural assumptions on (1):

A. *[Hyperbolicity] A is diagonalizable with real eigenvalues;*

B. *[Partial dissipativity] The spectrum of B is $\sigma(B) = \{0\} \cup \sigma_0$, where 0 is semi-simple with algebraic multiplicity $m \ge 1$ and $\sigma_0 \subset \{\lambda \in \mathbb{C} : \mathrm{Re}\,(\lambda) > 0\}$.*

Let $P_0^{(0)}$ be the eigenprojection associated with $0 \in \sigma(B)$. Let $C := P_0^{(0)} A P_0^{(0)}$, the dynamics on the equilibrium manifold is then approximately described by the reduced system $\partial_t \omega + C\partial_x \omega \approx 0$, where $\omega := P_0^{(0)} u$. Thus, C describes the large-time transport mechanism. We assume

C. *[Reduced hyperbolicity] C has m real semi-simple eigenvalues in* $\ker(B)$.

In addition, we assume the presence of the dissipativity described in terms of spectral properties of the Fourier symbol

$$E(\kappa) := B + \kappa A, \qquad \kappa \in \mathbb{C}. \tag{4}$$

D. *[Uniform dissipativity] There is a positive constant $\theta \in \mathbb{R}$ such that*

$$\mathrm{Re}\,(\lambda)(\kappa) \ge \theta |\kappa|^2/(1+|\kappa|^2), \qquad \kappa \in \mathbb{C}\backslash\{0\},$$

where λ are the eigenvalues of E.

Consider the kernel $G_t(x) := G(x,t) \in \mathbb{R}^{n\times n}$ associated with the system (1) satisfying

$$\partial_t G + A\partial_x G + BG = 0, \qquad G(x,0) = \delta(x)I, \tag{5}$$

where δ is the delta distribution and I is the $n \times n$ identity matrix. In order to study the large-time behavior of the solution u to (1), one studies the kernel G_t since $u = G_t * u_0$, and in this paper, by studying G_t in the frequency domain under assumptions **A**, **B**, **C**, and **D**, one can decompose G_t into

$$G_t(x) = K_t(x) + W_t(x) + R_t(x), \tag{6}$$

where $K_t(x) := K(x,t)$ arising in the low frequency includes kernels of diffusion waves propagating along the characteristics governed by the matrix C in $\ker(B)$, $W_t(x) := W(x,t)$ arising in the high frequency contains delta distributions decaying exponentially and it describes the propagation of signals along the hyperbolic characteristics governed by A, and $R_t(x) := R(x,t)$ is the remainder decaying faster than $K_t(x)$. It follows that

$$u(x,t) = U(x,t) + V(x,t) + L(x,t), \tag{7}$$

where $U(x,t) := K_t * u_0(x)$, $V(x,t) := W_t * u_0(x)$ and $L(x,t) := R_t * u_0(x)$.

Let $s \le m$ be the cardinality of the spectrum of C in $\ker(B)$. By the derivation of K_t, $U = \sum_{j=1}^{s} U_j$ where $U_j := P_j^{(0)} U$ and $P_j^{(0)}$ is the eigenprojection of the eigenvalue c_j of C in $\ker(B)$ for $j \in \{1, \dots, s\}$. Furthermore, let Γ_0 be an oriented closed curve in the resolvent set $\rho(B)$ of B such that it encloses 0 except for the other eigenvalues of B, one sets

$$D := -[P_0^{(1)} B P_0^{(1)} + P_0^{(0)} A P_0^{(1)} + P_0^{(1)} A P_0^{(0)}],$$

where $S_0^{(0)} := \frac{1}{2\pi i} \int_{\Gamma_0} z^{-1} (B - zI)^{-1}\, dz$ and $P_0^{(1)} := -[P_0^{(0)} A S_0^{(0)} + S_0^{(0)} A P_0^{(0)}]$. Then, U_j satisfies the system

$$\partial_t \widetilde{U} + c_j \partial_x \widetilde{U} - D_j \partial_{xx} \widetilde{U} = 0, \qquad \widetilde{U}(x,0) = \widetilde{U}_0(x), \tag{8}$$

where $\widetilde{U} \in \mathbb{R}^n$, $D_j := P_j^{(0)} D P_j^{(0)}$ is positive definite and $\widetilde{U}_0$ will be chosen later.

Let $r \le n$ be the cardinality of the spectrum $\sigma(A) := \{\alpha_1, \dots, \alpha_r\}$ of A, where $\alpha_j \in \mathbb{R}$. By the derivation of W_t, $V = Q \sum_{j=1}^{r} V_j$, where Q is the invertible matrix diagonalizing A and QV_j is the image of V in the range of the eigenprojection of α_j for $j \in \{1, \dots, r\}$. Moreover, let $\Lambda := Q^{-1} A Q = \operatorname{diag}(a_1, \dots, a_n)$ where $a_j \in \mathbb{R}$, then the system (1) in the diagonal coordinate $v := Q^{-1} u$ becomes

$$\partial_t v + \Lambda \partial_x v + Q^{-1} B Q v = 0, \qquad v(x,0) = Q^{-1} u_0(x). \tag{9}$$

For $j \in \{1, \dots, r\}$, one sets $\mathscr{S}_j := \{i \in \{1, \dots, n\} : a_i = \alpha_j\}$ and the projection

$$\left(\Pi_j^{(0)}\right)_{hk} := \begin{cases} 1 & \text{if } h = k \in \mathscr{S}_j, \\ 0 & \text{if otherwise.} \end{cases} \tag{10}$$

Then, V_j satisfies the following system, with respect to $\widetilde{V} \in \mathbb{R}^n$, obtained by projecting the system (9) onto $\mathrm{ran}(\Pi_j^{(0)})$; i.e., V_j satisfies the equation

$$\partial_t \widetilde{V} + \alpha_j \partial_x \widetilde{V} + \Pi_j^{(0)} Q^{-1} B Q \widetilde{V} = 0, \qquad \widetilde{V}(x,0) = \Pi_j^{(0)} Q^{-1} u_0(x). \tag{11}$$

Theorem 1 **(L^p-L^q decay estimates).** *Let u be the solution to (1), if* **A**, **B**, **C**, *and* **D** *hold, then for $1 \le q \le p \le \infty$, there are $C := C(p,q) > 0$ and $\delta > 0$ such that*

$$\|u - U - V\|_{L^p} \le C\, t^{-\frac{1}{2}(\frac{1}{q}-\frac{1}{p})-\frac{1}{2}} \|u_0\|_{L^q}, \tag{12}$$

where $\|U\|_{L^p} \le C t^{-\frac{1}{2}(\frac{1}{q}-\frac{1}{p})} \|u_0\|_{L^q}$ and $\|V\|_{L^2} \le C e^{-\delta t} \|u_0\|_{L^2}$, for $t \ge 1$.

Furthermore, the Goldstein–Kac system (2) possesses a symmetry property that allows us to increase the decay rate in the estimate (12). Thus, we are also interested in the following assumptions on the system (1).

C'. *[Reduced strictly hyperbolicity] C has m real distinct eigenvalues in* $\ker(B)$.

S. *[Symmetry] There is an invertible symmetric matrix S such that $AS = -SA$ and $BS = SB$.*

Theorem 2 **(Increased decay rates).** *With the same hypotheses in Theorem 1, if* **C'** *and* **S** *hold in addition, then for $1 \le q \le p \le \infty$, there is $C := C(p,q) > 0$ and*

$$\|u - U - V\|_{L^p} \le C\, t^{-\frac{1}{2}(\frac{1}{q}-\frac{1}{p})-1} \|u_0\|_{L^q} \qquad \text{for } t \ge 1. \tag{13}$$

2 Proofs of Main Results

The aim of this section is to give proofs of Theorems 1 and 2 by studying the Fourier transform $\hat{G}$ of the solution G to (5), which satisfies

$$\partial_t \hat{G} + E(i\xi)\hat{G} = 0, \qquad \hat{G}(\xi,0) = I, \tag{14}$$

where E is given by (4) with $\kappa := i\xi$ for $\xi \in \mathbb{R}$ and I is the $n \times n$ identity matrix.

2.1 Spectral Analysis

Following [2], since E in (4) can be expanded asymptotically near some specific points in the frequency domain, $\hat{G}(\xi,t) = e^{-E(i\xi)t}$ satisfying (14) can be expanded

based on expansions of E. Thus, $G(x,t) = \mathscr{F}^{-1}(\hat{G}(\xi,t))(x)$ satisfying (5) can be expanded and a time-asymptotic profile of G arises in the expansions of G. Hence, this subsection is dealt with a study of asymptotic expansions of E. Precisely, we consider the eigenvalue problem for E with low frequency as $|\xi| \to 0$, high frequency as $|\xi| \to +\infty$, and intermediate frequency as $|\xi|$ far from $0, +\infty$.

As $|\xi| \to 0$, since the eigenvalues of $E(i\xi) = B + i\xi A$ converge to the eigenvalues of B, they are divided into distinct groups represented by the distinct eigenvalues of B to which they converge. Without loss of generality, one considers the 0-*group* of E, where the 0-group contains the eigenvalues converging to 0 as $|\xi| \to 0$. Recall Γ_0 the oriented closed curve in the resolvent set $\rho(B)$ of B such that it encloses 0 except for the other eigenvalues of B, then following [2], $P_0 := -\frac{1}{2\pi i}\int_{\Gamma_0} R(z,\cdot)\,dz$ is the *total projection* of the 0-group of E; i.e., P_0 is the sum of the eigenprojections of the elements of the group, where $R(z,i\xi) := (E(i\xi) - zI)^{-1}$ is the resolvent of E. Moreover, following [2], on the compact set Γ_0, R is expanded as

$$R(z,i\xi) = R^{(0)}(z) + i\xi R^{(1)}(z) + (i\xi)^2 R^{(2)}(z) + \mathscr{O}(|\xi|^3), \qquad |\xi| \to 0, \tag{15}$$

where $R^{(0)}(z) := (B - zI)^{-1}$, $R^{(1)} := -R^{(0)}AR^{(0)}$ and $R^{(2)} := R^{(0)}AR^{(0)}AR^{(0)}$. Thus,

$$P_0(i\xi) = P_0^{(0)} + i\xi P_0^{(1)} + (i\xi)^2 P_0^{(2)} + \mathscr{O}(|\xi|^3), \qquad |\xi| \to 0, \tag{16}$$

where $P_0^{(j)} := -\frac{1}{2\pi i}\int_{\Gamma_0} R^{(j)}(z)\,dz$ for $j = 0, 1, 2$. Noting that $P_0^{(0)}$ is the eigenprojection of the eigenvalue 0 of B. Moreover, one has $E = EP_0 + E(I - P_0)$ corresponding to $\mathbb{C}^n = \mathrm{ran}(P_0) \oplus \mathrm{ran}(I - P_0)$. Thus, the eigenvalues in the 0-group of E are the eigenvalues of EP_0 in $\mathrm{ran}(P_0)$. Furthermore, if 0 is a semi-simple eigenvalue of B, then one obtains from (4) and (16) that

$$E(i\xi)P_0(i\xi) = (i\xi)[C - i\xi D + (i\xi)^2 H + \mathscr{O}(|\xi|^3)], \qquad |\xi| \to 0, \tag{17}$$

where $C := P_0^{(0)}AP_0^{(0)}$, $D := -(P_0^{(1)}BP_0^{(1)} + P_0^{(0)}AP_0^{(1)} + P_0^{(1)}AP_0^{(0)})$ and

$$H := P_0^{(1)}BP_0^{(2)} + P_0^{(2)}BP_0^{(1)} + P_0^{(0)}AP_0^{(2)} + P_0^{(2)}AP_0^{(0)} + P_0^{(1)}AP_0^{(1)}. \tag{18}$$

Hence, λ is an eigenvalue of EP_0 in $\mathrm{ran}(P_0)$ if and only if $\tilde{\lambda} := (i\xi)^{-1}\lambda$ is an eigenvalue of $E_0(i\xi) := (i\xi)^{-1}E(i\xi)P_0(i\xi) = C - i\xi D + \mathscr{O}(|\xi|^2)$ in $\mathrm{ran}(P_0)$. Therefore, the study of the 0-group of E is reduced to the study of the eigenvalue problem for E_0 in $\mathrm{ran}(P_0)$, i.e., the *reduction process* introduced in [2].

Based on the formula of E_0 and by the same argument as before, let $c_j \in \sigma(C)$, where $\sigma(C)$ is the spectrum of C, the total projection of the c_j-group of E_0 satisfies

$$P_j(i\xi) = P_j^{(0)} + i\xi P_j^{(1)} + \mathscr{O}(|\xi|^2), \qquad |\xi| \to 0, \tag{19}$$

where $P_j^{(0)}$ is the eigenprojection of c_j for $j \in \{1, \ldots, s\}$ and s the cardinality of $\sigma(C)$. Nonetheless, since we reduced $\mathbb{C}^n$ to $\mathrm{ran}(P_0)$, we consider only the c_j-groups

of E_0 whose total projections are *subprojections* of P_0, namely $P_j P_{j'} = \delta_{jj'} P_j$, $P_0 = \sum_{j=1}^{s} P_j$ and $\mathrm{ran}(P_0) = \bigoplus_{j=1}^{s} \mathrm{ran}(P_j)$, where $\delta_{jj'}$ is the Kronecker delta. It follows from [2] that we consider the c_j belonging to the set

$$\sigma[C, \ker(B)] := \{\text{the eigenvalues of } C \text{ in } \ker(B)\}. \tag{20}$$

Moreover, one has $E_0 = \sum_{j=1}^{s} E_0 P_j$ corresponding to $\mathrm{ran}(P_0) = \bigoplus_{j}^{s} \mathrm{ran}(P_j)$. Thus, if the semi-simplicity of the eigenvalues of C considered in $\ker(B)$ is given, then the process can be continued by considering, in $\mathrm{ran}(P_j)$, the operator

$$E_j(i\xi) := (i\xi)^{-1}(E_0(i\xi) - c_j I)P_j(i\xi) = -D_j + (i\xi)H_j + \mathscr{O}(|\xi|^2), \quad |\xi| \to 0, \tag{21}$$

where $D_j := P_j^{(0)} D P_j^{(0)}$ and

$$H_j := P_j^{(1)}(C - c_j I)P_j^{(1)} + P_j^{(0)} D P_j^{(1)} + P_j^{(1)} D P_j^{(0)} + P_j^{(0)} H P_j^{(0)}, \tag{22}$$

where D and H are in (17), for $j \in \{1, \dots, s\}$. Thus, for each $j \in \{1, \dots, s\}$, let

$$\sigma[D_j, \ker(C - c_j I)|_{\ker(B)}] := \{\text{the eigenvalues of } D_j \text{in} \\ \ker(C - c_j I) \text{restricted to} \ker(B)\}. \tag{23}$$

If $d_{jk} \in \sigma[D_j, \ker(C - c_j I)|_{\ker(B)}]$ for $k \in \{1, \dots, s_j\}$ and s_j the cardinality of $\sigma[D_j, \ker(C - c_j I)|_{\ker(B)}]$, the total projection of the d_{jk}-group of E_j satisfies

$$P_{jk}(i\xi) = P_{jk}^{(0)} + i\xi P_{jk}^{(1)} + \mathscr{O}(|\xi|^2), \qquad |\xi| \to 0, \tag{24}$$

where $P_{jk}^{(0)}$ is the eigenprojection of d_{jk}. Moreover, P_{jk} are subprojections of P_j.

Recall E in (4), $\sigma[C, \ker(B)]$ in (20), and $\sigma[D_j, \ker(C - c_j I)|_{\ker(B)}]$ in (23).

Proposition 1 (Low frequency). *If* **B** *and* **C** *hold, then for small* $|\xi|$*, we have*

$$E(i\xi) = \sum_{j,k=1}^{s,s_j} \big[(ic_j\xi + d_{jk}\xi^2)I + \xi^2 N_{jk}^{(0)} + \mathscr{O}(|\xi|^3)\big]\big[P_{jk}^{(0)} + i\xi P_{jk}^{(1)} + \mathscr{O}(|\xi|^2)\big] \\ + \sum_{j=1}^{h} \big[b_j I + M_j^{(0)} + \mathscr{O}(|\xi|)\big]\big[F_j^{(0)} + \mathscr{O}(|\xi|)\big], \tag{25}$$

where $c_j \in \sigma[C, \ker(B)]$, $d_{jk} \in \sigma[D_j, \ker(C - c_j I)|_{\ker(B)}]$, $P_{jk}^{(\ell)}$ *for* $\ell = 0, 1$ *are in* (24) *and* $N_{jk}^{(0)}$ *is the nilpotent matrix associated with* d_{jk} *for* $j \in \{1, \dots, s\}$ *and* $k \in \{1, \dots, s_j\}$*;* $b_j \in \sigma(B)\backslash\{0\}$*, where* $\sigma(B)$ *is the spectrum of* B*,* $F_j^{(0)}$ *and* $M_j^{(0)}$ *are the eigenprojection and the nilpotent matrix associated with* b_j*, respectively, for* $j \in \{1, \dots, h\}$ *and* h *is the cardinality of* $\sigma(B)\backslash\{0\}$*. If* **D** *holds,* $\mathrm{Re}\,(d_{jk}) > 0$ *for all* j *and* k*. Additionally, if* **C'** *and* **S** *hold,* $\mathscr{O}(|\xi|^3)$ *is increased to* $\mathscr{O}(|\xi|^4)$*.*

Proof. The expansion of E is obtained by the reduction process introduced as before under assumptions **B** and **C**.

We prove $\mathrm{Re}\,(d_{jk}) > 0$ if **D** holds for $j \in \{1, \dots, s\}$ and $k \in \{1, \dots, s_j\}$. From (25), the representation of E in $\mathrm{ran}(P_{jk})$ is $E_{jk}(i\xi) := ic_j\xi I + \xi^2[d_{jk}I + N_{jk}^{(0)} + \mathscr{O}(|\xi|)]$, where $c_j \in \sigma[C, \ker(B)]$, $d_{jk} \in \sigma[D_j, \ker(C - c_jI)|_{\ker(B)}]$, $N_{jk}^{(0)}$ is the nilpotent matrix associated with d_{jk} and P_{jk} is in (24). Thus, the eigenvalues of E in $\mathrm{ran}(P_{jk})$ are $\lambda = ic_j\xi + \xi^2\tilde{\lambda}$ where $\tilde{\lambda}$ are the eigenvalues of $d_{jk}I + N_{jk}^{(0)} + \mathscr{O}(|\xi|)$. Moreover, since $N_{jk}^{(0)}$ is nilpotent, $\tilde{\lambda} \to d_{jk}$ as $|\xi| \to 0$; i.e., $\tilde{\lambda} = d_{jk} + \mathscr{o}(1)$. Furthermore, by the construction of P_{jk}, the total projection P_0 in (16) of the 0-group of E satisfies $P_0 = \sum_{j,k=1}^{s,s_j} P_{jk}$. Thus, the 0-group of E includes the eigenvalues of E_{jk} in $\mathrm{ran}(P_{jk})$ for all j and k, namely the eigenvalues $\lambda_{jk}(i\xi) = ic_j\xi + \xi^2[d_{jk} + \mathscr{o}(1)]$. Thus, since $c_j \in \mathbb{R}$ under assumption **C**, if **D** holds, for $0 < |\xi| < \varepsilon$, there is $\theta > 0$ such that

$$\theta|\xi|^2/(1 + |\xi|^2) \le \mathrm{Re}\,[\lambda_{jk}(i\xi)] = \mathrm{Re}\,[ic_j\xi + d_{jk}\xi^2 + \mathscr{O}(|\xi|^2)] \le \mathrm{Re}\,(d_{jk})|\xi|^2 + \varepsilon|\xi|^2.$$

Dividing by $|\xi|^2$ in both sides and letting $\varepsilon \to 0$, one has $\mathrm{Re}\,(d_{jk}) \ge \theta > 0$.

Moreover, if **C'** holds, $c_j \in \sigma[C, \ker(B)]$ is simple for all $j \in \{1, \dots, s\}$. Hence, $d_{jk} \in \sigma[D_j, \ker(C - c_jI)|_{\ker(B)}]$ is simple for all $k \in \{1, \dots, s_j\}$ since they are considered in $\mathrm{ran}(P_j^{(0)})$, where $P_j^{(0)}$ is the eigenprojection of c_j for $j \in \{1, \dots, s\}$. Thus, the reduction process is continued, and following [2], the eigenvalues in the 0-group of E are approximated analytically by

$$\lambda_{jk\ell}(i\xi) := ic_j\xi - d_{jk}(i\xi)^2 + e_{jk\ell}(i\xi)^3 + \mathscr{O}(|\xi|^4), \quad |\xi| \to 0,$$

where $c_j \in \sigma[C, \ker(B)]$, $d_{jk} \in \sigma[D_j, \ker(C - c_jI)|_{\ker(B)}]$ and $e_{jk\ell}$ are the ℓth eigenvalue of $K := P_{jk}^{(0)} H_j P_{jk}^{(0)}$ in $\mathrm{ran}(P_{jk}^{(0)})$ where H_j is given by (22) and $P_{jk}^{(0)}$ in (24) is the eigenprojection of d_{jk}. On the other hand, **S** implies that the spectra of $E(\kappa)$ and $E^*(\kappa) := E(-\kappa)$ are the same for $\kappa \in \mathbb{C}$. Moreover, it follows from the reduction process for E^* that $e_{jk\ell}$ are also the eigenvalues of $-K$ in $\mathrm{ran}(P_{jk}^{(0)})$. Hence, since d_{jk} are simple, $\dim \mathrm{ran}(P_{jk}^{(0)}) = 1$ and thus $\ell = 1$ and $e_{jk\ell} = 0$. It also implies that the leading coefficients $K + N_{jk}^{(0)} P_{jk}^{(1)} + P_{jk}^{(1)} N_{jk}^{(0)}$ associated with $(i\xi)^3$ in $\mathscr{O}(|\xi|^3)$ are the null matrices, where $P_{jk}^{(1)}$ is in the expansion (24) and $N_{jk}^{(0)}$ is the nilpotent matrix associated with d_{jk}. Hence, $\mathscr{O}(|\xi|^3)$ is increased to $\mathscr{O}(|\xi|^4)$. The proof is done.

For high frequency, recall $\Lambda = Q^{-1}AQ$ where Q is the matrix diagonalizing A. E in (4) is written as $E(\zeta) = \zeta^{-1}QT(\zeta)Q^{-1}$ where $T(\zeta) := \Lambda + \zeta Q^{-1}BQ$ and $\zeta = (i\xi)^{-1}$. Moreover, $|\zeta| \to 0$ as $|\xi| \to +\infty$ and the eigenvalue problem for E can be treated as before.

Proposition 2 (High frequency). *If* **A** *holds, then for large* $|\xi|$*, we have*

$$E(i\xi) = Q \sum_{j,k=1}^{r,r_j} \big[(i\alpha_j\xi + \beta_{jk})I + \Theta_{jk}^{(0)} + \mathscr{O}(|\xi|^{-1})\big]\big[\Pi_{jk}^{(0)} + \mathscr{O}(|\xi|^{-1})\big]Q^{-1}, \quad (26)$$

where $\alpha_j \in \sigma(\Lambda)$, where $\sigma(\Lambda)$ is the spectrum of Λ, and $\Pi_j^{(0)}$ is in (10), β_{jk} is the kth eigenvalue of $\Pi_j^{(0)} Q^{-1} B Q \Pi_j^{(0)}$ in $\ker(\Lambda - \alpha_j I)$, $\Pi_{jk}^{(0)}$ and $\Theta_{jk}^{(0)}$ are the eigenprojection and the nilpotent matrix associated with β_{jk}, respectively, for $k \in \{1, \dots, r_j\}$ with r_j the cardinality of the set of β_{jk} and $j \in \{1, \dots, r\}$ with r the cardinality of the spectrum of A. Additionally, if **D** *holds, then* $\mathrm{Re}(\beta_{jk}) > 0$ *for all j and k.*

Proof. Similarly to before, one considers the reduction process for T and one deduces the expansion of E by substituting $\zeta = (i\xi)^{-1}$ into $E(\zeta) = \zeta^{-1} Q^{-1} T(\zeta) Q$. On the other hand, the eigenvalues of E are $\lambda_{jk}(i\xi) = i\alpha_j\xi + \beta_{jk} + \mathscr{O}(1)$ as $|\xi| \to +\infty$. Thus, since $\alpha_j \in \mathbb{R}$ under **A**, if **D** holds, for $|\xi|^{-1} < \varepsilon$, there is $\theta > 0$ such that $\theta|\xi|^2/(1 + |\xi|^2) \le \mathrm{Re}\,(\beta_{jk}) + \varepsilon$. As $\varepsilon \to 0$, $\mathrm{Re}\,(\beta_{jk}) \ge \theta > 0$. The proof is done.

Remark 1. (Intermediate frequency). For $\varepsilon \le |\xi| \le R$, except a finite number of exceptional points, the operator E in (4) has p (independent from ξ) distinct holomorphic eigenvalues together with p holomorphic eigenprojections and p holomorphic eigennilpotents associated with them. A more detailed discussion is in [2].

Let $P_{jk}^{(0)}$ and $P_{jk}^{(1)}$ be in the expansion (24). The initial data for (8) are chosen as

$$\widetilde{U}(x,0) := P_{jk}^{(0)} u_0(x), \quad (27)$$

$$\widetilde{U}(x,0) := P_{jk}^{(0)} u_0(x) + P_{jk}^{(1)} \partial_x u_0(x). \quad (28)$$

The choice (27) is for Theorem 1, and the choice (28) is for Theorem 2.

2.2 Fundamental Solution

We primarily introduce a useful lemma for which a proof can be found in [1].

Lemma 1. *If N is a constant complex nilpotent matrix, then for $\varepsilon' > 0$, there is $C := C(\varepsilon') > 0$ such that $|e^{\alpha N + M} - e^{\alpha N}| \le C e^{\varepsilon'|\alpha| + C|M|} |M|$ for every complex constant $\alpha := \alpha(t)$ and matrix $M := M(t)$ for $t > 0$.*

Considering the coefficients introduced in Propositions 1 and 2, let

$$\hat{K}(\xi, t) := \sum_{j,k=1}^{s,s_j} e^{-(ic_j\xi + d_{jk}\xi^2)t} e^{-\xi^2 t N_{jk}^{(0)}} P_{jk}^{(0)},$$

$$\hat{K}^*(\xi,t) := \sum_{j,k=1}^{s,s_j} e^{-(ic_j\xi+d_{jk}\xi^2)t} e^{-\xi^2 t N_{jk}^{(0)}} \left(P_{jk}^{(0)} + i\xi P_{jk}^{(1)}\right)$$

and one also sets $\hat{W}(\xi,t) := Q \sum_{j,k=1}^{r,r_j} e^{-(i\alpha_j\xi+\beta_{jk})t} e^{-\Theta_{jk}^{(0)} t} \Pi_{jk}^{(0)} Q^{-1}$.

Recall the solution $\hat{G}$ to the system (14), we have the following estimates.

Proposition 3 (Fundamental solution estimates). *Let $0 < \varepsilon < R$, $r \in [1,\infty]$ and $t \geq 1$, if* **A**, **B**, **C**, *and* **D** *hold, there are $C := C(r) > 0$ and $\delta > 0$ such that*
1. For $|\xi| < \varepsilon$, we have

$$\|\hat{G} - \hat{K}\|_{L^r} \leq Ct^{-\frac{1}{2}\frac{1}{r}-\frac{1}{2}} \quad \text{and} \quad \|\hat{W}\|_{L^r} \leq Ce^{-\delta t}. \tag{29}$$

In addition, if **C'** *and* **S** *hold, then*

$$\|\hat{G} - \hat{K}^*\|_{L^r} \leq Ct^{-\frac{1}{2}\frac{1}{r}-1}. \tag{30}$$

2. For $\varepsilon \leq |\xi| \leq R$, we have

$$\|\hat{G}\|_{L^r},\ \|\hat{K}\|_{L^r},\ \|\hat{K}^*\|_{L^r},\ \|\hat{W}\|_{L^r} \leq Ce^{-\delta t}. \tag{31}$$

3. For $|\xi| > R$, we have

$$\|\hat{G} - \hat{W}\|_{L^r} \leq Ce^{-\delta t} \text{ for } r > 1 \quad \text{and} \quad \|\hat{K}\|_{L^r},\ \|\hat{K}^*\|_{L^r} \leq Ce^{-\delta t}. \tag{32}$$

In particular, one has

$$\|\mathscr{F}^{-1}(\hat{G} - \hat{W})\|_{L^\infty} \leq Ce^{-\delta t}. \tag{33}$$

Proof. Noting that if $\{\Phi_{jk}\}$ is a sequence of projections satisfying that $\Phi_{jk}\Phi_{j'k'} = \delta_{jj'}\delta_{kk'}\Phi_{jk}$ for $j, j' \in \{1,\dots,m\}$ and $k, k' \in \{1,\dots,n\}$ for some integers $m, n \geq 1$, where $\delta_{jj'}$ is the Kronecker delta, and $\sum_{j,k=1}^{m,n} \Phi_{jk} = I$ the identity matrix, then if $X = \sum_{j,k=1}^{m,n} X_{jk}\Phi_{jk}$, where X_{jk} commutes with Φ_{jk}, one can prove that $e^X = \sum_{j,k=1}^{m,n} e^{X_{jk}}\Phi_{jk}$ since $e^X = \sum_{\ell=0}^{+\infty}(X)^\ell/\ell!$ and $(X)^\ell = \sum_{j,k=1}^{m,n}(X_{jk})^\ell\Phi_{jk}$ for any $\ell \geq 0$.

For $|\xi| < \varepsilon$, if **B** and **C** hold, then from Proposition 1, the expansion (25) of $E(i\xi) = B + i\xi A$ and the sequence including P_{jk} in (24) for $j \in \{1,\dots,s\}$ and $k \in \{1,\dots,s_j\}$ and F_j for $j \in \{1,\dots,h\}$, where $F_j(i\xi) := F_j^{(0)} + \mathscr{O}(|\xi|)$, satisfy the above properties of X and Φ_{jk} respectively. Hence, since the solution to (14) is $\hat{G}(\xi,t) = e^{-E(i\xi)t}$, one obtains from (25) that $\hat{G}(\xi,t) = \hat{G}_1 + \hat{G}_2$, where

$$\hat{G}_1(\xi,t) := \sum_{j,k=1}^{s,s_j} e^{-(ic_j\xi+d_{jk}\xi^2)t} e^{-\xi^2 t N_{jk}^{(0)}+\mathscr{O}(|\xi|^3)t}\big[P_{jk}^{(0)}+\mathscr{O}(|\xi|)\big], \tag{34}$$

$$\hat{G}_2(\xi,t) := \sum_{j=1}^{h} e^{-b_j t} e^{-M_j^{(0)}t+\mathscr{O}(|\xi|)t}\big[F_j^{(0)}+\mathscr{O}(|\xi|)\big]. \tag{35}$$

Furthermore, if, in addition, **C'** and **S** hold, from Proposition 1, for $j \in \{1,\dots,s\}$ and $k \in \{1,\dots,s_j\}$, the representation of E in $\mathrm{ran}(P_{jk})$ is $(ic_j\xi + d_{jk}\xi^2)I + \xi^2 N_{jk}^{(0)} + \mathscr{O}(|\xi|^4)$. Thus, if, in addition, **C'** and **S** hold, we have

$$\hat{G}_1(\xi,t) := \sum_{j,k=1}^{s,s_j} e^{-(ic_j\xi+d_{jk}\xi^2)t} e^{-\xi^2 t N_{jk}^{(0)}+\mathscr{O}(|\xi|^4)t}\big[P_{jk}^{(0)}+i\xi P_{jk}^{(1)}+\mathscr{O}(|\xi|^2)\big]. \tag{36}$$

Noting that if X is an $n\times n$ nilpotent matrix, there is an integer $m \le n$ such that $e^X = \sum_{\ell=1}^{m}[(\ell-1)!]^{-1}X^{\ell-1}$. In particular, if X is a nilpotent matrix associated with an eigenvalue, then m is exactly the algebraic multiplicity of this eigenvalue. Thus, from here, we always have such a sum of X if X is a nilpotent matrix.

Then, by Proposition 1, if **D** holds, since $\mathrm{Re}\,(d_{jk}) > 0$ and $N_{jk}^{(0)}$ is nilpotent for all j and k, by applying Lemma 1, if $m_{jk} \ge 1$ is the algebraic multiplicity of d_{jk} then

$$|\hat{G}_1 - \hat{K}| \le C\sum_{j,k=1}^{s,s_j} e^{-\frac{1}{2}\mathrm{Re}(d_{jk})|\xi|^2 t}|\xi|^3 t + C\sum_{j,k,\ell=1}^{s,s_j,m_{jk}} e^{-\frac{1}{2}\mathrm{Re}(d_{jk})|\xi|^2 t}(|\xi|^2 t)^{\ell-1}|\xi|.$$

Similarly, if **C'**, **S**, and **D** hold, one has

$$|\hat{G}_1 - \hat{K}^*| \le C\sum_{j,k=1}^{s,s_j} e^{-\frac{1}{2}\mathrm{Re}(d_{jk})|\xi|^2 t}|\xi|^4 t + C\sum_{j,k,\ell=1}^{s,s_j,m_{jk}} e^{-\frac{1}{2}\mathrm{Re}(d_{jk})|\xi|^2 t}(|\xi|^2 t)^{\ell-1}|\xi|^2.$$

Moreover, by Proposition 1, since $\mathrm{Re}\,(b_j) > 0$ and $M_j^{(0)}$ is nilpotent for all j, by Lemma 1, there is $C > 0$ such that $|\hat{G}_2| \le C\sum_{j=1}^{h} e^{-\frac{1}{2}\mathrm{Re}\,(b_j)t}(|\xi| + \sum_{\ell=1}^{m_j} t^{\ell-1})$, where $m_j \ge 1$ is the algebraic multiplicity associated with b_j for all j.

On the other hand, by changing of variables, for $c,\ \gamma > 0$, and $\delta \ge 0$, one has

$$\big\||\xi|^\gamma t^\delta e^{-c|\xi|^2 t}\big\|_{L^r} \le C(r)t^{-\frac{1}{2}\frac{1}{r}-\frac{\gamma}{2}+\delta} \quad \text{for } r\in[1,\infty] \text{ and } t \ge 1.$$

Thus, if **B**, **C**, and **D** hold, $\|\hat{G}-\hat{K}\|_{L^r} \le \|\hat{G}_1-\hat{K}\|_{L^r} + \|\hat{G}_2\|_{L^r} \le Ct^{-1/2r-1/2}$; and similarly if **B**, **C'**, **S**, and **D** hold, $\|\hat{G}-\hat{K}^*\|_{L^r} \le Ct^{-1/2r-1}$ for $r\in[1,\infty]$ and $t \ge 1$.

Moreover, by Proposition 2, if **A** and **D** hold, $\mathrm{Re}\,(\beta_{jk}) > 0$ and $\Theta_{jk}^{(0)}$ is nilpotent for all j and k. Thus, there are $C > 0$ and $\delta > 0$ such that we have

$$|\hat{W}| \le C \sum_{j,k,\ell=1}^{r,r_j,n_{jk}} e^{-\mathrm{Re}\,(\beta_{jk})t} t^{\ell-1} \le C e^{-\delta t}, \tag{37}$$

where $n_{jk} \ge 1$ is the algebraic multiplicity associated with β_{jk} for all j and k, and thus, $\|\hat{W}\|_{L^r} \le Ce^{-\delta t}$ for $|\xi| < \varepsilon$, $r \in [1, \infty]$ and $t \ge 1$.

For $\varepsilon \le |\xi| \le R$, by Remark 1, if **D** holds, there are $C > 0$ and $\delta > 0$ such that

$$\|\hat{G}\|_{L^r} \le \left\| e^{-\theta|\xi|^2 t/(1+|\xi|^2)} \right\|_{L^r} \le Ce^{-\delta t} \qquad \text{for } r \in [1, \infty] \text{ and } t \ge 1.$$

Moreover, for $\varepsilon \le |\xi| \le R$, the following hold

$$\|\hat{K}\|_{L^r}, \|\hat{K}^*\|_{L^r} \le C \sum_{j,k,\ell=1}^{s,s_j,m_{jk}} e^{-\mathrm{Re}(d_{jk})\varepsilon^2 t} t^{\ell-1} \le Ce^{-\delta t} \qquad \text{for } r \in [1, \infty] \text{ and } t \ge 1.$$

Furthermore, (37) implies $\|\hat{W}\|_{L^r} \le Ce^{-\delta t}$ for $\varepsilon \le |\xi| \le R$, $r \in [1, \infty]$ and $t \ge 1$.

Following Proposition 2, for $|\xi| > R$ large, from (26), the solution to (14) is

$$\hat{G}(\xi, t) := Q \sum_{j,k=1}^{r,r_j} e^{-(i\alpha_j\xi+\beta_{jk})t} e^{-\Theta_{jk}^{(0)} t + \mathscr{O}(|\xi|^{-1})t} \big[\Pi_{jk}^{(0)} + \mathscr{O}(|\xi|^{-1})\big] Q^{-1}. \tag{38}$$

Thus, since $\mathrm{Re}\,(\beta_{jk}) > 0$ and $\Theta_{jk}^{(0)}$ is nilpotent for all j and k due to Proposition 2, by applying Lemma 1, one has $|\hat{G} - \hat{W}| \le C \sum_{j,k=1}^{r,r_j} e^{-\frac{1}{2}\mathrm{Re}(\beta_{jk})t} |\xi|^{-1} t$. Hence, for $r > 1$ and $t \ge 1$, we have $\|\hat{G} - \hat{W}\|_{L^r} \le Ce^{-\delta t}$ for a $\delta > 0$. We are also interested in the behavior of $\hat{K}$ and $\hat{K}^*$ for $|\xi|$ large. In fact, since $|\xi| > R$ for R large, there is $\delta > 0$ such that for $r \in [1, \infty]$ and $t \ge 1$, one has

$$\|\hat{K}\|_{L^r}, \ \|\hat{K}^*\|_{L^r} \le C \sum_{j,k,\ell=1}^{s,s_j,m_{jk}} e^{-\frac{1}{2}\mathrm{Re}(d_{jk})R^2 t} \left\| e^{-\frac{1}{2}\mathrm{Re}(d_{jk})|\xi|^2 t} (|\xi|^2 t)^{\ell-1} \right\|_{L^r} \le Ce^{-\delta t}.$$

Finally, by applying the Taylor expansion for $X \mapsto e^X$, we can decompose

$$\hat{G} - \hat{W} = Q \sum_{j,k,\ell=1}^{r,r_j,n_{jk}} \frac{e^{-i\alpha_j\xi t}}{i\xi} \frac{e^{-\beta_{jk}t}}{(\ell-1)!} \big(-\Theta_{jk}^{(0)} t\big)^{\ell-1} M \Pi_{jk}^{(0)} Q^{-1} + e^{-(i\alpha_j\xi+\beta_{jk})t} H(|\xi|^{-2}),$$

where M is the leading coefficient matrix with respect to the term $(i\xi)^{-1}$ in $\mathscr{O}(|\xi|^{-1})$ and by Lemma 1, for $\varepsilon' > 0$, $H(|\xi|^{-2})$ satisfies $|H(|\xi|^{-2})| \le C(\varepsilon') e^{\varepsilon' t + C(\varepsilon')|\xi|^{-1} t} (1 + t)|\xi|^{-2}$. Taking the inverse Fourier transform, one obtains

$$\|\mathscr{F}^{-1}(\hat{G}-\hat{W})\|_{L^\infty} \le C\sum_{j,k,\ell=1}^{r,r_j,n_{jk}} e^{-\mathrm{Re}(\beta_{jk})t}t^{\ell-1}\|\mathscr{F}^{-1}[e^{-i\alpha_j xit}(i\xi)^{-1}]\|_{L^\infty} + \|\mathscr{F}^{-1}[e^{-i\alpha_j\xi t-\beta_{jk}t}H(|\xi|^{-2})]\|_{L^\infty}.$$

On the other hand, since $\omega(x,t) := \mathscr{F}^{-1}[e^{-i\alpha_j\xi t}(i\xi)^{-1}]$ is a solution to the wave equation $\partial_{tt}^2\omega - \partial_{xx}^2\omega = 0$, it implies $\|\mathscr{F}^{-1}[e^{-i\alpha_j\xi t}(i\xi)^{-1}]\|_{L^\infty}$ is bounded. Furthermore, by choosing $\varepsilon' = \mathrm{Re}\,(\beta_{jk})/4$, since $|\xi|^{-1} < \varepsilon$, we have

$$\|\mathscr{F}^{-1}[e^{-i\alpha_j\xi t-\beta_{jk}t}H(|\xi|^{-2})]\|_{L^\infty} \le C\sum_{j,k=1}^{r,r_j} e^{-\mathrm{Re}(\beta_{jk})t}\|H(|\xi|^{-2})\|_{L^1}.$$

Hence, there is $\delta > 0$ such that $\|\mathscr{F}^{-1}(\hat{G}-\hat{W})\|_{L^\infty} \le Ce^{-\delta t}$ for all $t \ge 1$.

2.3 Multiplier Estimates

One has the following estimates.

Proposition 4. *If* **A**, **B**, **C**, *and* **D** *hold, then for* $t \ge 1$, *one has*

$$\|\mathscr{F}^{-1}(\hat{G}-\hat{K}-\hat{W})\|_{L^1} \le Ct^{-\frac{1}{2}}. \tag{39}$$

If, in addition, **C'** *and* **S** *hold, one has*

$$\|\mathscr{F}^{-1}(\hat{G}-\hat{K}^*-\hat{W})\|_{L^1} \le Ct^{-1}. \tag{40}$$

Proof. Let $\chi_{1,3}$ be cut-off functions defined on $|\xi| \le \varepsilon$ and $|\xi| \ge R$, respectively, one sets $\chi_2 := 1-\chi_1-\chi_3$. We begin with the case $|x| \le Ct$ for a $C > 0$.

For $|\xi| \le \varepsilon$, one has $\hat{G}-\hat{K} = \hat{G}_1-\hat{K}+\hat{G}_2$, where $\hat{G}_1$ and $\hat{G}_2$ are given by (34) and (35), respectively. On the other hand, $\hat{G}_1-\hat{K} = I_1+I_2$, where

$$I_1 := \sum_{j,k=1}^{s,s_j} e^{(-ic_j\xi-d_{jk}\xi^2)t}\left[e^{-N_{jk}^{(0)}\xi^2t+\mathscr{O}(|\xi|^3)t} - e^{-N_{jk}^{(0)}\xi^2t}\right]P_{jk}^{(0)},$$

$$I_2 := \sum_{j,k=1}^{s,s_j} e^{(-ic_j\xi-d_{jk}\xi^2)t}e^{-N_{jk}^{(0)}\xi^2t+\mathscr{O}(|\xi|^3)t}\mathscr{O}(|\xi|).$$

We primarily estimate $\mathscr{F}^{-1}(\chi_1 I_1)$. For each j and each k, let $z = e^{i\phi/2}\xi$ where $\phi = \arg(d_{jk}) \in (-\pi/2, \pi/2)$ since $\mathrm{Re}\,(d_{jk}) > 0$, one obtains

$$\mathscr{F}^{-1}(\chi_1 I_1)(x,t) = \sum_{j,k=1}^{s,s_j} \int_\gamma \chi_1(e^{-i\phi/2}z) e^{i(x-c_j t)e^{-i\phi/2}z - |d_{jk}|z^2 t} \cdot \left(e^{-N_{jk}^{(0)} e^{-i\phi} z^2 t + \mathscr{O}(|e^{-i\phi/2}z|^3)t} - e^{-N_{jk}^{(0)} e^{-i\phi} z^2 t}\right) P_{jk}^{(0)} e^{-i\phi/2} dz,$$

where $\gamma := \{z \in \mathbb{C} : z = e^{i\phi/2}\xi, \xi \in [-\varepsilon, \varepsilon]\}$.

Then, we estimate each summand by letting $\eta := \min\{|x - c_j t|/2|d_{jk}|t, \varepsilon/2\}$. Since the integrand is holomorphic, we can change γ to $\widetilde{\gamma} := \gamma_1 \cup \gamma_2 \cup \gamma_3$, where

$$\gamma_1 := \{-\varepsilon e^{i\phi/2} + i\,\mathrm{sgn}(x - c_j t)\eta e^{-i\phi/2} s : s \in [0,1]\},$$
$$\gamma_2 := \{\zeta e^{i\phi/2} + i\,\mathrm{sgn}(x - c_j t)\eta e^{-i\phi/2} : \zeta \in [-\varepsilon, \varepsilon]\}$$

and $\gamma_3 := \{\varepsilon e^{i\phi/2} + i\,\mathrm{sgn}(x - c_j t)\eta e^{-i\phi/2}(1-s) : s \in [0,1]\}$. Then, there is $\delta > 0$ such that for $t \geq 1$ and $\varepsilon' = |d_{jk}|\cos(\phi)/8$, by Lemma 1, we have

$$\left|\int_{\gamma_1}\right| \leq C \int_0^1 e^{-|x-c_j t|\eta\cos(\phi)s} e^{-|d_{jk}|\cos(\phi)(\varepsilon^2 - \eta^2 s^2)t} e^{2\varepsilon'\varepsilon^2 t} \varepsilon^4 t ds \leq C e^{-\delta t}.$$

Similarly, $\left|\int_{\gamma_3}\right| \leq C e^{-\delta t}$ and if $\eta = |x - c_j t|/(2|d_{jk}|t)$, there is $c > 0$ such that

$$\left|\int_{\gamma_2}\right| \leq C \sum_{\ell=0}^{3} e^{-|x-c_j t|^2\cos(\phi)/8|d_{jk}|t} (|x - c_j t|/\sqrt{t})^\ell \int_{-\varepsilon}^{\varepsilon} e^{-\frac{1}{2}|d_{jk}|\cos(\phi)\zeta^2 t} |\zeta|^{3-\ell} t^{1-\frac{\ell}{2}} d\zeta$$
$$\leq C t^{-1} e^{-|x-c_j t|^2/c|d_{jk}|t},$$

and if $\eta = \varepsilon/2$, then for $t \geq 1$ and $\varepsilon' = |d_{jk}|\cos(\phi)/16$, there is $\delta > 0$ such that

$$\left|\int_{\gamma_2}\right| \leq C \int_{-\varepsilon}^{\varepsilon} e^{-|x-c_j t|\eta\cos(\phi)} e^{-|d_{jk}|\cos(\phi)(\zeta^2 - \eta^2)t} e^{2\varepsilon'\varepsilon^2 t} \varepsilon^3 t d\zeta \leq C e^{-\delta t}.$$

On the other hand, since $|x| \leq Ct$, $e^{-\delta t}$ is absorbed by $t^{-1}e^{-|x-c_j t|^2/c|d_{jk}|t}$ and it follows that $\|\mathscr{F}^{-1}(\chi_1 I_1)\|_{L^1} \leq Ct^{-1/2}$ for $t \geq 1$. Similarly, $\|\mathscr{F}^{-1}(\chi_1 I_2)\|_{L^1} \leq Ct^{-1/2}$. Thus $\|\mathscr{F}^{-1}[\chi_1(\hat{G} - \hat{K})]\|_{L^1} \leq Ct^{-1/2}$.

Moreover, since $\mathscr{F}^{-1} : L^1 \to L^\infty$, all remaining terms are bounded by $e^{-\delta t}$ by Proposition 3 and thus are absorbed by $t^{-1}e^{-|x-c_j t|^2/c|d_{jk}|t}$ except for $\mathscr{F}^{-1}[\chi_3(\hat{G} - \hat{W})]$. Hence, we estimate $\mathscr{F}^{-1}\left[\chi_3(\hat{G} - \hat{W})\right]$. For $|\xi| \geq R$, $\hat{G} - \hat{W} = J_1 + J_2$, where

$$J_1 := Q \sum_{j,k,\ell=1}^{r,r_j,n_{jk}} \frac{e^{-i\alpha_j \xi t}}{i\xi} \frac{e^{-\beta_{jk} t}}{(\ell-1)!} \left(-\Theta_{jk}^{(0)} t\right)^{\ell-1} M \Pi_{jk}^{(0)} Q^{-1},$$

where M is the leading coefficient matrix with respect to the term $(i\xi)^{-1}$ in $\mathscr{O}(|\xi|^{-1})$ and for all $\varepsilon' > 0$, one has $J_2 := \sum_{j,k=1}^{r,r_j} e^{-(i\alpha_j\xi+\beta_{jk})t} H(|\xi|^{-2})$, where $H(|\xi|^{-2})$ satisfies $|H(|\xi|^{-2})| \le C(\varepsilon')e^{\varepsilon' t + C(\varepsilon')|\xi|^{-1}t}(1+t)|\xi|^{-2}$.

For $t \ge 1$, since $|\mathscr{F}^{-1}[e^{-i\alpha_j\xi t}(i\xi)^{-1}]| \le C|x-\alpha_j t|$ for all j, we have

$$|\mathscr{F}^{-1}(\chi_3 J_1)(x,t)| \le C\sum_{j=1}^{r} te^{-\delta t}|x-\alpha_j t| + Ce^{-\delta t} \le Ct^{-1}e^{-|x-c_j t|^2/c|d_{jk}|t}.$$

Moreover, one has $|\mathscr{F}^{-1}(\chi_3 J_2)(x,t)| \le \|J_2\|_{L^1} \le Ce^{-\delta t} \le Ct^{-1}e^{-|x-c_j t|^2/c|d_{jk}|t}$. Thus,

$$\|\mathscr{F}^{-1}[\chi_3(\hat{G}-\hat{W})]\|_{L^1} \le \|\mathscr{F}^{-1}(\chi_3 J_1)\|_{L^1} + \|\mathscr{F}^{-1}(\chi_3 J_2)\|_{L^1} \le Ct^{-\frac{1}{2}}.$$

The proof is done for (39), where $|x| \le Ct$ for a $C > 0$. By similar computations, we can also prove (40), where $|x| \le Ct$ for a $C > 0$ if **C'** and **S** hold in addition, where $t^{-1}e^{-|x-c_j t|^2/c|d_{jk}|t}$ is substituted by $t^{-3/2}e^{-|x-c_j t|^2/c|d_{jk}|t}$.

We consider the case $|x| > Ct$ for $C > 0$. We estimate the L^1-norm of $\mathscr{F}^{-1}(\hat{G}-\hat{W})$ if **A** and **D** hold. The other estimates are similar. We have

$$\mathscr{F}^{-1}(\hat{G}-\hat{W})(x,t) = \lim_{R\to+\infty}\int_{-R}^{R} e^{ix\xi}\big(\hat{G}(\xi,t)-\hat{W}(\xi,t)\big)d\xi. \tag{41}$$

On the other hand, since $E(i\xi) = B + i\xi A$, $\hat{G}(\xi,t) = e^{-E(i\xi)t}$ is an entire function. Moreover, $\hat{W}$ is also holomorphic; by considering $\xi = \zeta + i\eta \in \mathbb{C}$, one changes the path of the integral in (41) from $\{(\zeta,0) : \zeta \text{ from } -R \text{ to } R\}$ to $\gamma := \gamma_1 \cup \gamma_2 \cup \gamma_3$, where $\gamma_1 := \{(\zeta,\eta) : \zeta = -R, \eta \text{ from } 0 \text{ to } x/t\}$, $\gamma_2 := \{(\zeta,\eta) : \zeta \text{ from } -R \text{ to } R, \eta = x/t\}$ and $\gamma_3 := \{(\zeta,\eta) : \zeta = R, \eta \text{ from } x/t \text{ to } 0\}$. Then, since R and $|x|/t$ are large, along γ, $\hat{G}$ has the representation of the high frequency case (38), and thus, one has the estimate

$$|\mathscr{F}^{-1}(\hat{G}-\hat{W})(x,t)| \le Ce^{-|x|^2/ct} \le Ct^{-1}e^{-|x|^2/2ct},$$

for some $c, C > 0$ since $e^{-|x|^2/2ct} \le e^{-C^2/2ct} \le t^{-1}$ due to the fact that $|x| > Ct$ for C large enough. Hence, we obtain $\|\mathscr{F}^{-1}(\hat{G}-\hat{W})\|_{L^1} \le Ct^{-1}$, and the proof is done.

2.4 *Proofs of Theorems 1 and 2*

If **A**, **B**, **C**, and **D** hold, we have $u - U - V = \mathscr{F}^{-1}(\hat{G}-\hat{K}-\hat{W}) * u_0$. On the other hand, let χ be the characteristic function, we have

$$\mathscr{F}^{-1}(\hat{G}-\hat{K}-\hat{W}) = \mathscr{F}^{-1}[(\hat{G}-\hat{K}-\hat{W})(\chi_{[0,\varepsilon)}+\chi_{[\varepsilon,R]})(|\xi|)] + \mathscr{F}^{-1}[(\hat{G}-\hat{W})\chi_{(R,\infty)}(|\xi|)] - \mathscr{F}^{-1}[\hat{K}\chi_{(R,\infty)}(|\xi|)].$$

Thus, since $\mathscr{F}^{-1} : L^1 \to L^\infty$, we have

$$\|\mathscr{F}^{-1}(\hat{G}-\hat{K}-\hat{W})\|_{L^\infty} \le \|(\hat{G}-\hat{K}-\hat{W})(\chi_{[0,\varepsilon)}+\chi_{[\varepsilon,R]})\|_{L^1} + \|\hat{K}\chi_{(R,\infty)}\|_{L^1} + \|\mathscr{F}^{-1}[(\hat{G}-\hat{W})\chi_{(R,\infty)}(|\xi|)]\|_{L^\infty}.$$

By (29), (31), (32), and (33) in Proposition 3 and the Young inequality, we obtain

$$\|u-U-V\|_{L^\infty} \le \|\mathscr{F}^{-1}(\hat{G}-\hat{K}-\hat{W})\|_{L^\infty}\|u_0\|_{L^1} \le Ct^{-1}\|u_0\|_{L^1}, \qquad t \ge 1.$$

Furthermore, by (39) in Proposition 4 and the Young inequality, one has

$$\|u-U-V\|_{L^r} \le \|\mathscr{F}^{-1}(\hat{G}-\hat{K}-\hat{W})\|_{L^1}\|u_0\|_{L^r} \le Ct^{-\frac{1}{2}}\|u_0\|_{L^r}, \quad r \in [1,\infty],\ t \ge 1.$$

Therefore, by interpolation, we obtain (12) in Theorem 1.

Similarly, if **C'** and **S** hold in addition, then substituting $\hat{K}$ by $\hat{K}^*$ and using (30), (31), (32), and (33) in Proposition 3 and (40) in Proposition 4, we obtain (13) in Theorem 2. We finish the proofs.

Acknowledgements The authors thank the conference organizers for giving us an opportunity to write this contribution. We are grateful to the referees for their helpful comments and suggestions.

References

1. S. Bianchini, B. Hanouzet, R. Natalini, Asymptotic behavior of smooth solutions for partially dissipative hyperbolic systems with a convex entropy. Commun. Pure Appl. Math. **60**(11), 1559–1622 (2007)
2. T. Kato, *Perturbation Theory for Linear Operators*, vol. 132. (Springer Science & Business Media, 1995)
3. P. Marcati, K. Nishihara, The L^p-L^q estimates of solutions to one-dimensional damped wave equations and their application to the compressible flow through porous media. J. Differ. Equ. **191**(2), 445–469 (2003)
4. C. Mascia, Exact representation of the asymptotic drift speed and diffusion matrix for a class of velocity-jump processes. J. Differ. Equ. **260**(1), 401–426 (2016)
5. Y. Ueda, R. Duan, S. Kawashima, Decay structure for symmetric hyperbolic systems with non-symmetric relaxation and its application. Arch. Ration. Mech. Anal. **205**(1), 239–266 (2012)

A Numerical Approach of Friedrichs' Systems Under Constraints in Bounded Domains

Clément Mifsud and Bruno Després

Abstract We present here an explicit finite volume scheme on unstructured meshes adapted to first-order hyperbolic systems under constraints in bounded domains. This scheme is based on the work (Coudière, Vila, Villedieu in C R Acad Sci Paris Sér I Math 331:95–100, 2000, [3]) in the unconstrained case and the splitting strategy of Després, Lagoutière, Seguin (Nonlinearity 24:3055–3081, 2011, [4]). We show that this scheme is stable under a Courant–Friedrichs–Lewy condition (and convergent for problems posed in the whole space), and we illustrate the solution constructed by this scheme on the example of the simplified model of perfect plasticity. From the theoretical point of view, the interaction between the constraint and the boundary of the domain in the model of perfect plasticity is encoded by a nonlinear boundary condition. With this numerical approach, we will show that, even if this scheme uses the underlying linear boundary condition, the results are consistent with the nonlinear model (and in particular with the nonlinear boundary condition).

Keywords Finite volume schemes · Friedrichs' systems · Constrained problems

Mathematics Subject Classification 2010 65M08 · 65M12 · 35L50 · 35L60 74C05

1 Introduction

The aim of this article is to examine the numerical approximation of Friedrichs' equations under constraints (posed in the whole space or in bounded domains). To do so, we use a popular method for hyperbolic problems: the method of finite

C. Mifsud (✉) · B. Després
Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie – Paris 6,
CNRS, UMR 7598, 75005 Paris, France
e-mail: mifsud@ljll.math.upmc.fr

B. Després
Institut Universitaire de France, Paris, France
e-mail: despres@ann.jussieu.fr

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_25

volumes (for a detailed presentation of this method, we refer to [5, 6]). Although there is an important number of schemes that have been developed, the analysis of the convergence and its rate of schemes on unstructured meshes for multidimensional problems (i.e., the domain is a subset of $\mathbb{R}^n$ with $n > 1$, and the solution belongs to $\mathbb{R}^m$ with $m > 1$) are still in its infancy.

However, the article [9] has established a rate of convergence for the RKDG scheme (see [2]), using P0 finite elements in space and the RK1 scheme in time, on unstructured meshes for generic Friedrichs' systems of the following form

$$\begin{cases} \partial_t U + \sum_{j=1}^n \partial_j (A_i U) + BU = f, & \text{in } (0,T) \times \mathbb{R}^n, \\ U(0,x) = U_0(x), & \text{in } \mathbb{R}^n, \end{cases} \tag{1}$$

where $U : (t,x) \in (0,T) \times \mathbb{R}^n \to \mathbb{R}^m$, $A_i : (t,x) \in (0,T) \times \mathbb{R}^n \to \mathbb{M}^{m\times m}_{\text{sym}}$, $B : (t,x) \in (0,T) \times \mathbb{R}^n \to \mathbb{M}^{m\times m}$, $f : (t,x) \in (0,T) \times \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbb{M}^{m\times m}$ (resp. $\mathbb{M}^{m\times m}_{\text{sym}}$) is the space of $m \times m$ (resp. symmetric) matrices with real coefficients. A similar analysis has been performed in the note [3] on bounded domains.

In addition, the study of the convergence of a scheme based on the Rusanov scheme on Cartesian meshes has been performed in [4] for constrained Friedrichs' systems. In fact, to show the existence of a weak solution (in the sense of Definition 1) to the constrained Friedrichs' system

$$\begin{cases} \partial_t U + \sum_{j=1}^n A_j \partial_j U = 0 \quad \text{in } (0,T] \times \mathbb{R}^n; \quad U(0,x) = U^0(x) \quad \text{if } x \in \mathbb{R}^n, \\ U(t,x) \in \mathscr{C} \quad \text{if } (t,x) \in [0,T] \times \mathbb{R}^n, \end{cases} \tag{2}$$

where $\mathscr{C}$ is a fixed closed and convex subset of $\mathbb{R}^m$ (with $0 \in \overset{\circ}{\mathscr{C}}$), the authors construct a numerical solution with a two-step scheme such that a subsequence converges to a weak solution of (2). In this paper, we extend the strategy of [4] to schemes on unstructured meshes and to problems posed in bounded domains.

In Sect. 2, we recall some notations and define our finite volume scheme on unstructured meshes for constrained Friedrichs' systems in bounded domains.

In Sect. 3, we recall some results of [4] on constrained Friedrichs' systems in the whole space and state a convergence result in the whole space on a similar scheme (to the one presented in Sect. 2 on bounded domains). This result tells us that the finite volume scheme on unstructured meshes, based on the work [9], associated with a projection step has the same rate of convergence (in the space $L^2((0,T) \times \mathbb{R}^n; \mathbb{R}^m)$) as in the unconstrained case (obtained in [9]).

In Sect. 4, we show that the scheme presented in Sect. 2 is stable (under a Courant–Friedrichs–Lewy condition) in the space $L^\infty(0,T; L^2(\Omega, \mathbb{R}^m))$.

Then in Sect. 5, we briefly recall the equations of the simplified model of the dynamical perfect plasticity problem (described in [1]) and how this problem is related to the constrained Friedrichs' systems.

Finally, in Sect. 6, we illustrate the solution constructed by this scheme on the example of the simplified model of the dynamical perfect plasticity problem and show that the interaction between the constraint and the boundary condition that

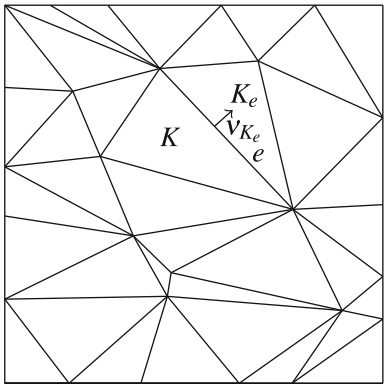


Fig. 1 An unstructured meshes of the square $[0, 1] \times [0, 1]$. Here the polytopes are triangles

has been underlined theoretically by the nonlinear boundary condition can also be observed numerically.

2 Description of the Scheme

In this section, we present the general framework of this work and the scheme we are interested in. Let $\mathscr{T}_h$ be a triangulation of $\Omega \subset \mathbb{R}^n$ (a n-dimensional polytope); i.e., $\mathscr{T}_h = (K_i)_{i\in\mathscr{I}}$, with $\mathscr{I} \subset \mathbb{N}$, is a family of open nonempty convex polytope such that $\cup_{i\in\mathscr{I}} \overline{K_i} = \overline{\Omega}$, for all $i \neq j$, $K_i \cap K_j = \emptyset$ and $h = \sup_{i\in\mathscr{I}} (\mathrm{diam} K_i) < +\infty$. The set of edges of a polytope K is denoted $\mathscr{E}_K$. We introduce the following notations (see also Fig. 1),

$m_K, m_{\partial K}$: $\mathscr{L}^n$-measure of K, $\mathscr{H}^{n-1}$-measure ∂K,
$e \in \mathscr{E}_K$: an edge ($(n-1)$-dimensional polytope) of K with $\mathscr{H}^{n-1}$-measure m_e,
$\mathscr{E}_{K\mathrm{i}}, \mathscr{E}_{K\mathrm{b}}$: the set of interior edges e of K, the set of boundary edges e of K,
ν_{K_e} : the unit exterior normal of K on the edge e with $\nu_{K_e}=(\nu^1_{K_e}, \nu^2_{K_e}, \ldots, \nu^n_{K_e})$,
K_e : neighboring cell of K with $\overline{K} \cap \overline{K_e} = e$.

We also suppose that the triangulation is regular in the sense that there exists a constant $C_1 > 0$ (independent of the triangulation $\mathscr{T}_h$) such that

$$\forall K \in \mathscr{T}_h, \qquad C_1 h^n \leq m_K, \quad \text{and} \quad \forall K \in \mathscr{T}_h, \quad \forall e \in \mathscr{E}_K \qquad C_1 h^{n-1} \leq m_e.$$

We want to investigate the numerical approximation (using finite volume schemes) of the following constrained Friedrichs system

$$\begin{cases} \partial_t U + \sum_{i=1}^n A_i \partial_i U = f, \quad \text{on } (0,T)\times\Omega; \quad U(0,x) = U_0(x), \quad \text{on } \Omega, \\ (A_\nu - M_\nu)U = 0, \quad \text{on } (0,T)\times\partial\Omega; \quad U(t,x) \in \mathscr{C}, \quad \text{a.e in } (0,T)\times\Omega \end{cases} \tag{3}$$

where $\mathscr{C} \subset \mathbb{R}^m$ is a closed convex (independent of t and x) with $0 \in \overset{\circ}{\mathscr{C}}$, $A_\nu = \sum_{i=1}^n A_i \nu^i$ with $\nu = \left(\nu^1, \ldots, \nu^n\right)$ is the unit exterior normal to Ω, and M_ν is a non-negative symmetric matrix that encodes the boundary condition and has to satisfy some algebraic conditions (see [8, Sect. 2.1]).

Remark 1. In particular, due to the hypotheses on A_ν and M_ν, we have

1. For all $k \in \mathbb{R}^m$, there exists a unique triple (k^0, k_-, k_+) such that $k = k^0 + k_- + k_+$ and $k^0 \in \ker A_\nu, k^- \in (\ker(A_\nu - M_\nu)) \cap \mathrm{Im} A_\nu$, and $k^+ \in (\ker(A_\nu + M_\nu)) \cap \mathrm{Im} A_\nu$.
2. For all $k, \kappa \in \mathbb{R}^m$, $\langle k | A_\nu \kappa \rangle = \langle k_- | A_\nu \kappa_- \rangle + \langle k_+ | A_\nu \kappa_+ \rangle$.

The equations of (3) have to be understood in a weak sense (see Definition 1 for the case $\Omega = \mathbb{R}^n$ and Sect. 5 for the general case). To approximate the solutions of this kind of problem, we first forget about the constraint and use a finite volume scheme (explicit in time) based on the note [3]. More precisely, we use a piecewise constant approximation of U, denoted by V_h, such that

$$\forall (t, x) \in [t^p, t^{p+1}) \times K, \qquad V_h(t, x) = v_K^p, \quad \text{with } v_K^0 = \frac{1}{m_K} \int_K U_0(x)\, \mathrm{d}x,$$

where $0 = t^0 < t^1 < \cdots < t^{N+1} = T$ ($t^{p+1} - t^p = \Delta t$), and in a first step, we construct

$$\frac{m_K}{\Delta t} \left(v_K^{p+1,*} - v_K^p \right) + \sum_{e \in \mathscr{E}_K} g_{K_e} m_e = f_K^p := \frac{1}{m_K \Delta t} \int_{t^p}^{t^{p+1}} \int_K f(t, x)\, \mathrm{d}x\, \mathrm{d}t,$$

where $A_{K_e} = \sum_{i=1}^n A_i \nu_{K_e}^i$ and we define the interior fluxes ($e \cap \partial\Omega = \emptyset$),

$$g_{K_e} = \underbrace{(A_{K_e})^+ v_K^p}_{\text{Outcoming flow from } K \text{ to } K_e} + \underbrace{(A_{K_e})^- v_{K_e}^p}_{\text{Incoming flow in } K \text{ from } K_e}, \tag{4}$$

where we denote $(A_{K_e})^-$ (resp. $(A_{K_e})^+$) the negative (resp. positive) part of A_{K_e}, and the (centered) boundary fluxes,

$$g_{K_e} = \frac{A_{K_e} + M_{K_e}}{2} v_K^p, \tag{5}$$

with $M_{K_e} = M_{\nu_{K_e}}$ a matrix satisfying the conditions of [8, Sect. 2.1] (see also Remark 1). In order to take account of the constraint, we simplify project on each cell K the value $v_K^{p+1,*}$ onto the set $\mathscr{C}$. Hence, the second step is

$$v_K^{p+1} = P_{\mathscr{C}} \left(v_K^{p+1,*} \right).$$

where $P_{\mathscr{C}}$ is the projection onto $\mathscr{C}$. It leads us to the following scheme for $U_0 \in L^2(\mathbb{R}^n;\mathscr{C})$,

$$\boxed{\begin{cases} \forall K \in \mathscr{T}_h, \quad v_K^0 = \frac{1}{m_K}\int_K U_0(x)\,\mathrm{d}x, \\ \forall K \in \mathscr{T}_h, \forall 0 \le p \le N,\ v_K^{p+1,*} = v_K^p - \frac{\Delta t}{m_K}\sum_{e\in\mathscr{E}_K} g_{K_e} m_e + \Delta t f_K^p, \\ \forall K \in \mathscr{T}_h, \forall 0 \le p \le N, \quad v_K^{p+1} = P_{\mathscr{C}}\left(v_K^{p+1,*}\right). \end{cases}} \tag{6}$$

Thanks to the following discrete Green formula

$$\sum_{e\in\mathscr{E}_K} A_{K_e} m_e = 0 \quad \Leftrightarrow \quad \sum_{e\in\mathscr{E}_{K\mathrm{b}}} A_{K_e} m_e + \sum_{e\in\mathscr{E}_{K\mathrm{i}}} (A_{K_e})^+ m_e = \sum_{e\in\mathscr{E}_{K\mathrm{i}}} -(A_{K_e})^- m_e, \tag{7}$$

one can rewrite the first step of the scheme (6) in a nonconservative form

$$\frac{v_K^{p+1,*} - v_K^p}{\Delta t} = \sum_{e\in\mathscr{E}_{K\mathrm{i}}} \frac{m_e}{m_K}(A_{K_e})^-(v_K^p - v_{K_e}^p) - \sum_{e\in\mathscr{E}_{K\mathrm{b}}} \frac{m_e}{m_K}\frac{M_{K_e} - A_{K_e}}{2} v_K^p + f_K^p. \tag{8}$$

Remark 2. We denote by $\langle ; \rangle$ the canonical scalar product of $\mathbb{R}^m$ and $|.|$ the associated norm. By abuse of notation, we also use the notation $|.|$ for the (matrix) operator norm associated with the canonical norm of $\mathbb{R}^m$.

Remark 3. When $\Omega = \mathbb{R}^n$, one can use the scheme (6) to approximate the solution of the problem (2). In that case, all the sums over $\mathscr{E}_{K\mathrm{b}}$ are empty sums.

3 Previous Results on Constrained Friedrichs' Systems in the Whole Space

The aim of this section is to recall the definition of weak solutions to Friedrichs' systems under convex constraints in the whole space and to state some numerical results about these systems. We consider the following Cauchy problem: find $U : [0,T]\times\mathbb{R}^n \to \mathbb{R}^m$ such that

$$\begin{cases} \partial_t U + \sum_{j=1}^n A_j \partial_j U = 0 \ \text{ in } (0,T)\times\mathbb{R}^n; \quad U(0,x) = U^0(x) \ \text{ if } x \in \mathbb{R}^n, \\ U(t,x) \in \mathscr{C} \ \text{ if } (t,x) \in [0,T]\times\mathbb{R}^n, \end{cases} \tag{9}$$

where $\mathscr{C}$ is a fixed (*i.e.,* independent of the time and space variables) nonempty closed and convex subset of $\mathbb{R}^m$ containing 0 in its interior, the matrices A_j are $m \times m$ symmetric matrices independent of time and space, and $T > 0$. This type of nonlinear hyperbolic problems has been introduced in [4] where a notion of weak solutions to problem (9) has been defined.

Definition 1. Let $U^0 \in L^2(\mathbb{R}^n, \mathscr{C})$ and $T > 0$. A function $U \in L^2([0,T] \times \mathbb{R}^n, \mathscr{C})$ is a weak constrained solution of (9) if we have for all $\kappa \in \mathscr{C}$ and $\phi \in \mathscr{C}_c^\infty([0,T[\times\mathbb{R}^n)$ with $\phi \geq 0$,

$$\int_0^T \int_{\mathbb{R}^n} \left(|U-\kappa|^2 \partial_t \phi + \sum_{j=1}^n \big\langle U-\kappa;\, A_j(U-\kappa) \big\rangle \, \partial_j \phi \right) \mathrm{d}x \, \mathrm{d}t + \int_{\mathbb{R}^n} |U^0(x) - \kappa|^2 \phi(0,x) \, \mathrm{d}x \geq 0. \tag{10}$$

We recall here the existence and uniqueness result of [4].

Theorem 1. *Assume that $U^0 \in L^2(\mathbb{R}^n, \mathscr{C})$. There exists a unique weak constrained solution $U \in L^2([0,T] \times \mathbb{R}^n, \mathscr{C})$ to (9) in the sense of Definition 1.*

The existence of a solution has been obtained in [4] thanks to a finite volume scheme on Cartesian grids. At each time step, the scheme first let the solution evolve according to the Rusanov scheme without taking care about the constraint. Then, on each mesh they project the solution onto the set of constraints.

Thanks to this splitting strategy and to a compactness argument (which relies on the fact that the mesh is Cartesian), they show that the numerical solution admits a convergent subsequence and they prove that the limit of this subsequence has to be a solution of (9) in the sense of Definition 1.

In this paper, we use this splitting strategy for schemes defined on unstructured meshes. One can show that the scheme (6) (see Remark 3) enjoys the same rate of convergence as in the unconstrained case (for the complete proof, see [7]).

Theorem 2. *Let $U \in H^1((0,T) \times \mathbb{R}^n; \mathscr{C})$ be a dissipative solution associated with the initial condition $U_0 \in H^1(\mathbb{R}^n; \mathscr{C})$. Let V_h be the solution constructed from U_0 thanks to the scheme (6) (see Remark 3). Then we have,*

$$\|U - V_h\|_{L^2((0,T)\times\mathbb{R}^n;\mathbb{R}^m)} \leq C\sqrt{h},$$

for some constant C depending on ε, n, T, U_0 and the matrices A_i.

4 Stability in Time of Schemes

Once we know that the strategy of [4] combined with the scheme, analyzed in [9], leads to a convergent scheme (on unstructured meshes) for constrained Friedrichs' systems in $(0,T) \times \mathbb{R}^n$, one can analyze this splitting strategy on bounded domains (i.e., for Problem (3)). In this section, we prove that the scheme (6) enjoys a stability property under a Courant–Friedrichs–Lewy condition. For simplicity, we decide to derive this stability property in the case where the source term is null. In that case, the $L^2(\mathbb{R}^n)$-norm of the solution does not increase in time.

Proposition 1. *Suppose that the following CFL condition holds:*

$$\max\left(\sup_{K,e\in\mathscr{E}_K}\frac{\Delta t m_{\partial K}}{m_K}\left|(A_{K_e})^-\right|,\ \sup_{K,e\in\mathscr{E}_{K\mathrm{b}}}\frac{\Delta t m_{\partial K}}{m_K}\left|(M_{K_e}-A_{K_e})/2\right|\right)\leq 1, \quad (11)$$

the scheme (6) is stable; i.e., the approximate solution V_h satisfies (here $f\equiv 0$)

$$\forall t\in[0,T],\qquad \|V_h(t,\cdot)\|_{L^2(\mathbb{R}^n;\mathbb{R}^m)}\leq\|U_0\|_{L^2(\mathbb{R}^n;\mathbb{R}^m)}.$$

Proof. From the nonconservative form (8), we have

$$v_K^{p+1,*}=\sum_{e\in\mathscr{E}_K}\frac{m_e}{m_{\partial K}}v_K^{p+1,*}(e),$$

where we set

$$v_K^{p+1,*}(e)=\begin{cases}v_K^p+\frac{\Delta t m_{\partial K}}{m_K}(A_{K_e})^-(v_K^p-v_{K_e}^p), & \text{if } e\in\mathscr{E}_{K\mathrm{i}},\\ v_K^p-\frac{\Delta t m_{\partial K}}{m_K}\frac{M_{K_e}-A_{K_e}}{2}v_K^p, & \text{if } e\in\mathscr{E}_{K\mathrm{b}}.\end{cases}$$

Observe that we have for all $e\in\mathscr{E}_{K\mathrm{i}}$, since $(A_{K_e})^-\in\mathbb{M}^{m\times m}_{\mathrm{sym}}$,

$$\begin{aligned}|v_K^{p,*}(e)|^2=|v_K^p|^2-\tfrac{\Delta t m_{\partial K}}{m_K}\left(-\left\langle v_K^p;(A_{K_e})^-v_K^p\right\rangle+\left\langle v_{K_e}^p;(A_{K_e})^-v_{K_e}^p\right\rangle\right)\\ +\tfrac{\Delta t m_{\partial K}}{m_K}\left\langle v_K^p-v_{K_e}^p;\left(\mathrm{Id}+\tfrac{\Delta t m_{\partial K}}{m_K}(A_{K_e})^-\right)(A_{K_e})^-(v_K^p-v_{K_e}^p)\right\rangle\end{aligned}$$

Using the CFL condition, we obtain that

$$\forall y\in\mathbb{R}^m,\left\langle\left(\mathrm{Id}+\frac{\Delta t m_{\partial K}}{m_K}(A_{K_e})^-\right)y;y\right\rangle\geq 0. \quad (12)$$

In particular, if we apply (12) to $y=\left(-(A_{K_e})^-\right)^{1/2}(v_K^p-v_{K_e}^p)$, it yields

$$|v_K^{p,*}(e)|^2\leq|v_K^p|^2-\frac{\Delta t m_{\partial K}}{m_K}\left(-\left\langle v_K^p;(A_{K_e})^-v_K^p\right\rangle+\left\langle v_{K_e}^p;(A_{K_e})^-v_{K_e}^p\right\rangle\right). \quad (13)$$

Now, if $e\in\mathscr{E}_{K\mathrm{b}}$, we have, again since A_{K_e} and M_{K_e} belong to $\mathbb{M}^{m\times m}_{\mathrm{sym}}$,

$$\begin{aligned}|v_K^{p+1,*}(e)|^2=|v_K^p|^2-\frac{\Delta t m_{\partial K}}{m_K}\left\langle v_K^p;\frac{M_{K_e}-A_{K_e}}{2}v_K^p\right\rangle\\ -\frac{\Delta t m_{\partial K}}{m_K}\left\langle\frac{M_{K_e}-A_{K_e}}{2}\left(\mathrm{Id}-\frac{\Delta t m_{\partial K}}{m_K}\left(\frac{M_{K_e}-A_{K_e}}{2}\right)\right)v_K^p;v_K^p\right\rangle. \quad (14)\end{aligned}$$

Similarly, the CFL condition (11) implies that for all $y \in \mathbb{R}^m$, we have

$$\left\langle \mathrm{Id} - \frac{\Delta t m_{\partial K}}{m_K} \left(\frac{M_{K_e} - A_{K_e}}{2} \right) y; y \right\rangle \geq 0,$$

and algebraic manipulations (see Remark 1) tell us that

$$\begin{aligned} &\left\langle \frac{M_{K_e} - A_{K_e}}{2} \left(\mathrm{Id} - \frac{\Delta t m_{\partial K}}{m_K} \left(\frac{M_{K_e} - A_{K_e}}{2} \right) \right) v_K^p; v_K^p \right\rangle \\ &\qquad = \left\langle \left(\mathrm{Id} - \frac{\Delta t m_{\partial K}}{m_K} \left(\frac{M_{K_e} - A_{K_e}}{2} \right) \right) M_{K_e}^{1/2} (v_K^p)_+; M_{K_e}^{1/2} (v_K^p)_+ \right\rangle \geq 0, \end{aligned}$$

which implies that (14) becomes

$$|v_K^{p+1,*}(e)|^2 \leq |v_K^p|^2 - \frac{\Delta t m_{\partial K}}{m_K} \left\langle v_K^p; \frac{M_{K_e} - A_{K_e}}{2} v_K^p \right\rangle.$$

Using convexity, it yields

$$\begin{aligned} |v_K^{p+1,*}|^2 \leq |v_K^p|^2 - \frac{\Delta t}{m_K} \sum_{e \in \mathscr{E}_{K\mathrm{i}}} \left(- \left\langle v_K^p; (A_{K_e})^- v_K^p \right\rangle + \left\langle v_{K_e}^p; (A_{K_e})^- v_{K_e}^p \right\rangle \right) m_e \\ - \frac{\Delta t}{m_K} \sum_{e \in \mathscr{E}_{K\mathrm{b}}} \left\langle v_K^p; \frac{M_{K_e} - A_{K_e}}{2} v_K^p \right\rangle m_e. \end{aligned}$$

Furthermore, if we use the relation (7), we obtain

$$\begin{aligned} |v_K^{p+1,*}|^2 \leq |v_K^p|^2 - \frac{\Delta t}{m_K} \sum_{e \in \mathscr{E}_{K\mathrm{i}}} \left(\left\langle v_K^p; (A_{K_e})^+ v_K^p \right\rangle + \left\langle v_{K_e}^p; (A_{K_e})^- v_{K_e}^p \right\rangle \right) m_e \\ - \frac{\Delta t}{m_K} \sum_{e \in \mathscr{E}_{K\mathrm{b}}} \left\langle v_K^p; \frac{A_{K_e} + M_{K_e}}{2} v_K^p \right\rangle m_e. \end{aligned} \tag{15}$$

Remark that, thanks to Remark 1, we have for all $e \in \mathscr{E}_{K\mathrm{b}}$

$$\left\langle v_K^p; \frac{A_{K_e} + M_{K_e}}{2} v_K^p \right\rangle = \left\langle (v_K^p)_-; M_{K_e} (v_K^p)_- \right\rangle \geq 0.$$

Consequently, from (15) and since for all $y \in \mathbb{R}^m$, $|P_{\mathscr{C}}(y)| \leq |y|$, we obtain

$$|v_K^{p+1}|^2 \leq |v_K^p|^2 - \frac{\Delta t}{m_K} \sum_{e \in \mathscr{E}_{K\mathrm{i}}} \left(\left\langle v_K^p; (A_{K_e})^+ v_K^p \right\rangle + \left\langle v_{K_e}^p; (A_{K_e})^- v_{K_e}^p \right\rangle \right) m_e. \tag{16}$$

Then, we remark

$$\sum_{K\in\mathscr{T}_h}\sum_{e\in\mathscr{E}_{K\mathrm{i}}}\left(\left\langle v_K^p;(A_{K_e})^+v_K^p\right\rangle+\left\langle v_{K_e}^p;(A_{K_e})^-v_{K_e}^p\right\rangle\right)m_e=0.$$

Consequently, summing the inequality (16) over $K\in\mathscr{T}_h$ and from $p=0$ to $q-1$, where $t\in[0,T]$ and q an integer such that $t\in[t^q,t^{q+1})$ (or $q=N+1$ if $t=T$), leads to the stability property.

5 The Simplified Model of the Dynamical Perfect Plasticity

Let us briefly recall the equations of this model and the two points of view that one can use to describe its (theoretical) solution. First, the equations, derived from the physics of solids (see [1, Sects. 3.1 and 3.2]), of this simplified model of dynamical perfect plasticity are

$$\begin{cases}\partial_t v-\mathrm{div}\sigma=f, & \nabla v=\partial_t\sigma+\partial_t p,\\ |\sigma|\le 1, \quad \text{and } \langle\sigma;\partial_t p\rangle=|\partial_t p|.\end{cases}\tag{17}$$

where $v:\Omega\times[0,T]\to\mathbb{R}$ is the velocity of the material, $\sigma:\Omega\times[0,T]\to\mathbb{R}^2$ the Cauchy stress tensor, and $p:\Omega\times[0,T]\to\mathbb{R}^2$ the plastic deformation tensor and Ω is a open bounded subset of $\mathbb{R}^2$. The tensor σ is constrained to stay in the unit closed Euclidean ball of $\mathbb{R}^2$, denoted $\overline{B}$. To these equations, we add initial and boundary conditions. The boundary condition, that comes from the hyperbolic point of view, is the following nonlinear one

$$\langle\sigma;\nu\rangle+T(v)=0,\quad \text{on } (0,T)\times\partial\Omega,\tag{18}$$

where $T(z)=\min(-1,\max(z,1))$. It shows a threshold on the velocity (due to the constraint) in the boundary condition. We also need an initial condition

$$(v,\sigma)(t=0)=(v_0,\sigma_0)\tag{19}$$

that has to satisfy two hypotheses

$$\langle\sigma_0;\nu\rangle+v_0=0\quad\mathscr{H}^1\text{ on }\partial\Omega,\tag{20}$$

$$|\sigma_0|\le 1\text{ a.e. in }\Omega.\tag{21}$$

The first condition asserts that the initial condition has to satisfy the hyperbolic boundary condition that one could use in the unconstrained case, and the second condition states that the initial condition satisfies the constraint. In fact, one can show (see [1, Proposition 7.1]) that the solution of this simplified model satisfies the

following inequality for all $(k,\tau)\in\mathbb{R}\times\overline{B}$ and all $\varphi\in W^{1,\infty}(\mathbb{R}\times\mathbb{R}^2)$ (with $\varphi\geq 0$ and compactly supported in $\mathbb{R}\times\mathbb{R}^2$)

$$\int_0^T\int_\Omega \left((v-k)^2+|\sigma-\tau|^2\right)\partial_t\varphi\,\mathrm{d}x\,\mathrm{d}t+\int_\Omega\left((v_0-k)^2+|\sigma_0-\tau|^2\right)\varphi(0)\,\mathrm{d}x$$
$$-2\int_0^T\int_\Omega(\sigma-\tau)\cdot\nabla\varphi(v-k)\,\mathrm{d}x\,\mathrm{d}t+2\int_0^T\int_\Omega f(v-k)\varphi\,\mathrm{d}x\,\mathrm{d}t$$
$$+2\int_0^T\int_{\partial\Omega}(\sigma\cdot\nu-\tau\cdot\nu)(T(v)-k)\varphi\,\mathrm{d}\mathscr{H}^{n-1}\,\mathrm{d}t\geq 0. \quad (22)$$

Thanks to (18) and algebraic manipulations, one has

$$(\sigma\cdot\nu-\tau\cdot\nu)(T(v)-k)$$
$$=\frac{1}{4}\left((k+\tau\cdot\nu)^2-(T(v)-k-(\sigma\cdot\nu-\tau\cdot\nu))^2\right)\geq\frac{1}{4}(k+\tau\cdot\nu)^2, \quad (23)$$

Equation (23) allows us to rewrite (22), using the hyperbolic variable $U=^t(v,\sigma)$ as

$$\int_0^T\int_\Omega|U-\kappa|^2\,\partial_t\varphi+\sum_{i=1}^2\langle U-\kappa;A_i(U-\kappa)\rangle\,\partial_i\varphi+2\langle F;U-\kappa\rangle\,\varphi\,\mathrm{d}x\,\mathrm{d}t$$
$$+\int_\Omega|U_0-\kappa|^2\,\varphi(t=0)\,\mathrm{d}x+\int_0^T\int_{\partial\Omega}\langle\kappa_+;M_\nu\kappa_+\rangle\,\varphi\,\mathrm{d}\mathscr{H}^{n-1}(x)\,\mathrm{d}t\geq 0, \quad (24)$$

where $F=^t(f,0,0)$, $U_0=^t(v_0,\sigma_0)$, $\kappa=^t(k,\tau)$

$$A_1=\begin{pmatrix}0&-1&0\\-1&0&0\\0&0&0\end{pmatrix},\quad A_2=\begin{pmatrix}0&0&-1\\0&0&0\\-1&0&0\end{pmatrix}\quad\text{and}\quad M_\nu=\begin{pmatrix}1&0&0\\0&(\nu^1)^2&\nu^1\nu^2\\0&\nu^1\nu^2&(\nu^2)^2\end{pmatrix}, \quad (25)$$

and κ_+ stands for the projection onto $(\ker(A_\nu+M_\nu))\cap\mathrm{Im}A_\nu$. The fact that Eq. (24) is satisfied for all κ and all φ is the definition of a solution to Problem (3) (see also [8]). In addition, when the solution U is in $W^{1,\infty}([0,T];L^2(\Omega;\mathscr{C})$, one can show (see [1, Sect. 7]) that Eqs. (17)–(19) are equivalent to this definition of a weak constrained solution to Problem (3).

6 A Numerical Test on the Simplified Model of the Dynamical Perfect Plasticity

Now that this mechanical problem has been put into the hyperbolic framework (3), the simplified model of dynamical perfect plasticity can be approached thanks to the scheme described in Sect. 2. One important point to notice first is that this scheme

does not include a special treatment at the boundary to model the nonlinear boundary condition (18). Indeed, we only take into account the constraint thanks to a projection step on every mesh and the first step of this scheme uses the linear boundary condition

$$(A_v - M_v)U = 0 \qquad \Leftrightarrow \qquad \langle \sigma ; \nu \rangle + v = 0. \tag{26}$$

Our goal now is to test numerically the interactions between the boundary condition and the constraint for this particular hyperbolic system under constraint and to see if the nonlinear boundary condition is obtained with this scheme. The major point that allows us to bring to light these facts is the velocity threshold overrun in the boundary condition (18). To observe this overrun, we present here one test case (for more test cases, see [7, Sect. 4.4]).

The test is based on the following formal motivation: We want to observe large velocities near the boundary. But if we look at the equation of motion

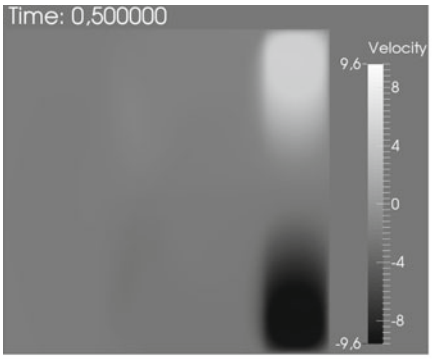
$$\partial_t v - \mathrm{div}\sigma = f,$$

we see that if f is positive (for example) near the boundary (for each time), then the velocity is going to increase over time near the boundary. Hence, we present a test case when the source term f varies from -50 to 50 near the boundary and is equal to zero elsewhere.

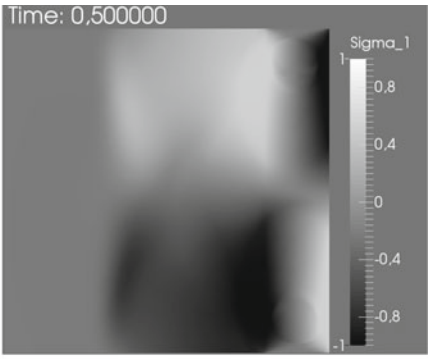
This test allows us to obtain large velocity near the boundary (i.e., $|v| \gg 1$ near $\partial\Omega$) and to bring to light that the nonlinear boundary is taken into account by our scheme. For this test case, we use the following data

- Spatial domain : $\Omega = [0, 1] \times [0, 1]$. Our mesh is regular and contains 80000 triangles.
- Final time : $T = 1$. We use 800 time steps, and consequently, the CFL condition (11) is approximately equal to 0.71.
- Initial data : In this test, we use data that touch the boundary $x = 1$. The initial velocity v_0 is null outside the open ball B_1 of radius 0.3 and center (1, 0.5), and v_0 is equal to -1 on the open ball B_2 of radius 0.25 and center (1, 0.5). In the strip between these two balls, we join these two constants using a $\mathscr{C}^1$ connection. It is important to notice that $-1 \leq v_0 \leq 0$. In order to satisfy the (linear) boundary condition at $x = 1$, the first component of σ is equal to $-v_0$. The second component of σ is null on Ω. Consequently, we have $v_0 + \langle \sigma ; \nu \rangle = 0$ on $\partial\Omega$. Remark also that the initial data belong to the convex set of constraints.
- The term source f is equal to $100y - 50$ for all $t \in [0, T]$, for all $y \in [0, 1]$ and $x > 0.8$ and to 0 elsewhere.

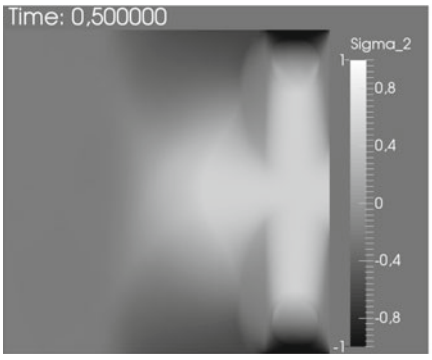
We decide to highlight the interaction between the constraint and the boundary at time $t = 0.5$ in Fig. 2. In this figure, we display the velocity (top left of the figure), the first component, denoted σ_1 in the following, of σ (top right), the second component (bottom left), denoted σ_2, and the term $\sigma_1 + T(v)$ (which is involved in the boundary condition at $x = 1$: $\sigma_1 + T(v) = 0$).



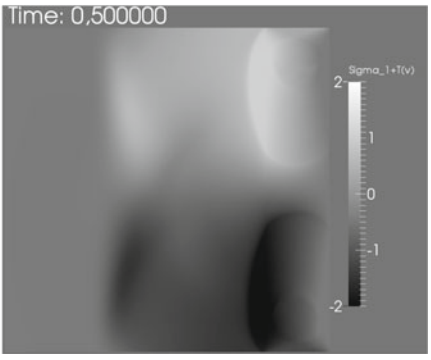
(a) Velocity



(b) First component σ_1 of σ



(c) Second component σ_2 of σ



(d) $\sigma_1 + T(v)$ (*i.e.* the boundary term on the right of the domain)

Fig. 2 Test case at time $t = 0.5$

We observe that the introduction of our term source in the strip $[0.8, 1] \times [0, 1]$ allows us to get a large velocity (i.e., $|v| \gg 1$) near the boundary $x = 1$ (see Fig. 2a). The theoretical boundary condition implies that in this situation we should see that $\sigma_1 = -1$ at the upper end of the boundary $x = 1$ (and, consequently, $\sigma_2 = 0$ due to the constraint) and $\sigma_1 = 1$ at the lower end of the boundary $x = 1$ (and $\sigma_2 = 0$ due to the constraint). Numerically, the scheme produces a solution that matches the mathematical model (see Fig. 2b, c). Consequently, the nonlinear boundary condition is satisfied by the numerical approximation (see Fig. 2d) despite the fact that we have not implemented any particular treatment at the boundary to get this nonlinear boundary condition. This fact may be seen as a first validation of our scheme.

References

1. J.-F. Babadjian, C. Mifsud, Hyperbolic structure for a simplified model of dynamical perfect plasticity. Arch. Ration. Mech. Anal. **223**(2), 761–815 (2017)
2. B. Cockburn, C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. Math. Comp. **52**(186), 411–435 (1989)
3. Y. Coudière, J.-P. Vila, P. Villedieu, Convergence d'un schéma volumes finis explicite en temps pour les systèmes hyperboliques linéaires symétriques en domaines bornés. C. R. Acad. Sci. Paris Sér. I Math. **331**(1), 95–100 (2000)
4. B. Després, F. Lagoutière, N. Seguin, Weak solutions to Friedrichs systems with convex constraints. Nonlinearity **24**(11), 3055–3081 (2011)
5. R. Eymard, T. Gallouët, R. Herbin, Finite volume methods, in *Handbook of Numerical Analysis*, vol. VII (North-Holland, Amsterdam, 2000), pp. 713–1020
6. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics (Cambridge University Press, Cambridge, 2002)
7. C. Mifsud, Variational and hyperbolic methods applied to constrained mechanical systems. Ph.D. thesis, Université Pierre et Marie Curie (2016)
8. C. Mifsud, B. Després, N. Seguin, Dissipative formulation of initial boundary value problems for Friedrichs' systems. Commun. Partial Differ. Equ. **41**(1), 51–78 (2016)
9. J.-P. Vila, P. Villedieu, Convergence of an explicit finite volume scheme for first order symmetric systems. Numer. Math. **94**(3), 573–602 (2003)

Lagrangian Representation for Systems of Conservation Laws: An Overview

Stefano Modena

Abstract We present an overview on some recent works in collaboration with S. Bianchini (see Bianchini and Modena in Lagrangian representation for solution to general systems of conservation laws [9] and the Ph.D. thesis Modena in Interaction functionals, Glimm approximations and Lagrangian structure of BV solutions for hyperbolic systems of conservation laws [15]), in which we propose a way to describe BV solutions to hyperbolic systems of conservation laws in one space dimension from a *Lagrangian* point of view.

Keywords Conservation laws · Hyperbolic systems · Interaction functional Lagrangian representation

1 Introduction

One of the key observations in fluid dynamics is that the fluid flow can be described from two different (and in some sense complementary) points of view: the Lagrangian points of view (in which the trajectory in space–time of each single fluid particle is tracked) and the Eulerian point of view (in which one looks at fluid motion focusing on fixed locations in the space through which the fluid flows as time passes).

From a mathematical perspective, such duality between the Lagrangian and the Eulerian approach can be seen, for instance, in the framework of the continuity equation:

$$\begin{cases} \partial_t v(t,x) + \mathrm{div}_x\big(v(t,x)b(t,x)\big) & = 0, \\ v(0,x) & = \bar{v}(x), \end{cases} \tag{1}$$

where $v : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}$ is the unknown and $b : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^d$ is a given vector field. It is well known that, under suitable regularity assumptions, the solution

S. Modena (✉)
Max Planck Institute for Mathematics in the Sciences,
Inselstrasse 22, 04103 Leipzig, Germany
e-mail: Stefano.Modena@math.uni-leipzig.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_26

to (1) can be written, for any time $t \in [0, \infty)$, as

$$v(t, \cdot)\mathscr{L}^1 = \mathtt{X}(t)_\sharp\big(\bar{v}\mathscr{L}^1\big), \tag{2}$$

where $\mathtt{X} : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}^d$ is the flow generated by the ODE

$$\begin{cases} \dfrac{\partial \mathtt{X}}{\partial t}(t, y) &= b(t, \mathtt{X}(t, y)), \\ \mathtt{X}(0, y) &= y, \end{cases} \tag{3}$$

$\mathscr{L}^d$ is the Lebesgue measure on $\mathbb{R}^d$ and $\sharp$ denotes the push-forward in the sense of measures.[1]

In the framework of the continuity equation, the Lagrangian and the Eulerian approach help each other: For instance, in the smooth setting, one can use the ODE (3) (Lagrangian approach) to solve the PDE (1) (Eulerian approach), while in the non-smooth setting one can use the PDE to solve the ODE (see [13]). The duality between the two approaches can be used not only to prove the existence of solutions, but also to prove their uniqueness and their stability and to investigate further properties of them, like their fine structure, their regularity, and so on. In few words, we could say that *two is better than one*: what cannot be done using the Lagrangian approach could be hopefully done using the Eulerian one, and vice versa.

For these reasons, it is an interesting question whether systems of conservation laws

$$\begin{cases} \partial_t u + \partial_x F(u) &= 0, \\ u(0, x) &= \bar{u}(x), \end{cases} \qquad u = u(t, x) \in \mathbb{R}^n, \ t \geq 0, \ x \in \mathbb{R}, \tag{4}$$

can be analyzed from a Lagrangian point of view. Here, $F : \mathbb{R}^n \to \mathbb{R}^n$ is a generic smooth function, which is only assumed to be *strictly hyperbolic*; i.e. its differential $DF(u)$ has n distinct real eigenvalues in each point of its domain. We restrict our analysis to one space dimension, since this is the setting where a satisfying well-posedness theory for entropic solutions is available.[2]

For a scalar conservation law with a smooth initial datum, the method of characteristics provides a reasonable Lagrangian approach to the problem. Such method was extended by C. Dafermos (through the notion of *generalized characteristics* in [12]) to systems whose characteristic fields are either genuinely nonlinear or linearly degenerate,[3] and to initial data which are just BV. However, Dafermos' approach can not be further generalized to systems where the flux F has no convexity properties.

[1] If A, B are sets, $\mathscr{A}, \mathscr{B}$ are σ-algebras on A, B, respectively, and $f : A \to B$ is a measurable function, then for any measure μ on $(A, \mathscr{A})$, the push-forward $f_\sharp\mu$ is the measure on $(B, \mathscr{B})$, defined by $f_\sharp\mu(E) = \mu(f^{-1}(E))$ for any $E \in \mathscr{B}$.

[2] By *entropic solution*, we mean a solution obtained as limit of vanishing viscosity approximations; see [3].

[3] See [11] for the definition of genuinely nonlinear or linearly degenerate characteristic fields. Roughly speaking, it amounts to say that the flux F has some strong *convexity* property.

Another Lagrangian approach in the analysis of conservation laws was proposed by T.-P. Liu in [14], where he introduced the notion of *wave tracing* for the waves present in an approximate solution to the system (4), constructed by means of the Glimm scheme. However, in [14], only approximate solutions (which in some sense are just piecewise constant functions) are considered.

Recently, some papers appeared in which a Lagrangian analysis is developed for the exact (and not approximate) entropic solution to conservation laws with a flux F which does not satisfy any convexity assumption. In particular,

- in [4, 5] (see also for a previous, slightly different approach [10]) S. Bianchini and E. Marconi develop a Lagrangian approach for the solution to the Cauchy problem associated to a scalar conservation law ($n = 1$), whose flux $F : \mathbb{R} \to \mathbb{R}$ is any smooth function and whose initial datum $\bar{u} \in L^\infty(\mathbb{R})$ is any bounded function;
- in [9, 15] S. Bianchini and the author develop a Lagrangian approach for the solution to the Cauchy problem associated to a system of conservation laws ($n \geq 1$), whose flux is any smooth strictly hyperbolic function and whose initial datum $\bar{u} \in BV(\mathbb{R})$ is a function of bounded variation.

In both cases, the starting point is the analysis of BV entropic solutions to scalar conservation laws. The extension to L^∞ initial data (for the scalar equation) [4, 5] or to systems [9, 15] requires, however, several new ideas. The goal of this notes is to present the notion of *Lagrangian representation* for BV entropic solutions to systems of conservation laws (4), proposed in [9, 15], and to present the main ideas behind the construction of such Lagrangian representation, focusing in particular on the difficulties in extending the scalar BV analysis to the system case.

As a final remark, we would like to stress that both in the scalar case and in the system one, the Lagrangian analysis is done in the same setting in which the well-posedness of the Cauchy problem is already know. We do not want to use Lagrangian methods to prove such well-posedness again. Rather, the aim of our new *Lagrangian tools* is to analyze in a more precise way the solution u to the Cauchy problem (4), in order to prove further properties of it. As an example, in the scalar case, the Lagrangian approach can be used to prove the *concentration of entropies* (see the papers by Bianchini and Marconi [4, 5]); in the system case, the Lagrangian tools can be used to study the fine structure of the solution (see the paper by Bianchini and the author [9] and the Ph.D. thesis of the author [15]).

2 Analysis of BV Solutions to Scalar Conservation Laws

The starting point of our analysis is the study of entropic BV solutions to scalar conservation laws

$$\begin{cases} \partial_t u + \partial_x F(u) = 0, \\ u(0, x) = \bar{u}(x), \end{cases} \qquad \bar{u} \in BV(\mathbb{R}) \text{ with compact support, } F : \mathbb{R} \to \mathbb{R} \text{ smooth.} \tag{5}$$

The first question we have to answer is: What is a good notion of *Lagrangian representation* of the solution to (5)? A hint in this direction is given by the following observation: By Vol'pert's rule[4] the distributional derivative $\partial_x u$ (which is a measure, being $u \in BV$) satisfies the $1D$ continuity equation

$$\partial_t(\partial_x u) + \partial_x(\hat{\lambda}(t,x)\partial_x u) = 0 \text{ in a distributional sense,} \tag{6}$$

where

$$\hat{\lambda}(\bar{t},\bar{x}) := \begin{cases} F'(u(\bar{t},\bar{x})) & \text{if } x \mapsto u(\bar{t},x) \text{ is continuous at } \bar{x}, \\ \frac{F(u(\bar{t},\bar{x}+))-F(u(\bar{t},\bar{x}-))}{u(\bar{t},\bar{x}+)-u(\bar{t},\bar{x}-)} & \text{if } x \mapsto u(\bar{t},x) \text{ has a jump at } \bar{x}. \end{cases} \tag{7}$$

Mimicking (1)–(2)–(3), we give the following definition.

Definition 1. A *Lagrangian representation* for the entropic solution u to (5) is a triple $(W, \mathtt{X}, \rho)$, where

1. $W \subseteq \mathbb{R}$ is a bounded interval; its elements are denoted by w and are called *waves*;
2. $\mathtt{X} : [0,\infty) \times W \to \mathbb{R}$ is $\|F'\|_{L^\infty}$-Lipschitz in t for fixed w and increasing in w for fixed t, and it is called *flow* or *position function*;
3. $\rho : W \to [-1,1]$ is called *density function*,

such that for a.e. time $t \in [0,\infty)$

$$\partial_x u(t,\cdot) = \mathtt{X}(t,\cdot)_\sharp\big(\rho \mathscr{L}^1|_W\big) \text{ in the sense of measures} \tag{8}$$

and

$$\frac{\partial \mathtt{X}}{\partial t}(t,w) = \hat{\lambda}(t,\mathtt{X}(t,w)) \text{ for } |\rho|\mathscr{L}^1 - \text{a.e.} w \in W. \tag{9}$$

Remark 1. Notice that (8) is the analog of (2) and (9) is the analog of (3); only two differences must be observed:

- In (8), the term which is transported is an absolute continuous measure w.r.t. $\mathscr{L}^1$, even if the initial datum $\partial_x \bar{u}$ has a jump part or a Cantor part;
- In (9), in general $\mathtt{X}(0,w) \neq w$; i.e., w is just the label of a particle with no relationship with its starting point.

Definition 1 provides a (hopefully) good notion of Lagrangian representation. How can we now explicitly construct the objects W, $\mathtt{X}$, ρ satisfying the properties above?

As usual in the theory of conservation laws, the idea is to consider a sequence of approximate solutions $(u^q)_{q\in\mathbb{N}}$ solving the approximate Cauchy problem

$$\begin{cases} \partial_t u^q + \partial_x F^q(u^q) = 0, \\ u^q(0,x) = \bar{u}^q(x), \end{cases} \tag{10}$$

[4]The Vol'pert's rule (see, for instance, [1, Theorem 3.96]) is the chain rule for the derivative of the composition $F(u(x))$ of a Lipschitz function F with a BV function u.

where F^q is the piecewise affine interpolation of F with grid size 2^{-q} and $\bar{u}^q$ is a piecewise constant function taking values in $2^{-q}\mathbb{Z}$ such that $\|\bar{u}^q - \bar{u}\|_{L^1} \to 0$ as $q \to \infty$. The solution u^q to (10) can be constructed by means of the wavefront tracking algorithm (see [11, Chap. 4]), and it is a piecewise constant function with values in $2^{-q}\mathbb{Z}$ for any time t which converges strongly in L^1 to the entropic solution of (5), as $q \to \infty$.

Since $u^q(t,\cdot)$ is piecewise constant, it is not difficult to construct by hand a Lagrangian representation of it.[5] Now, the family $\{\mathtt{X}^q\}_q$ is pre-compact in $L^1([0,\infty) \times \mathbb{R})$, since, by Definition 1, each $\mathtt{X}^q$ is $\|F'\|$-Lipschitz in t for fixed w and increasing in w for fixed t; the family $\{\rho^q\}$ is weakly* pre-compact in $L^\infty(W)$. Therefore, up to subsequences, $\mathtt{X}^q \to \mathtt{X}$ strongly in L^1 and $\rho^q \to \rho$ weakly* in L^∞.

Equation (8) is then easily obtained passing to the limit in the corresponding equation for approximations

$$\partial_x u^q(t) = \mathtt{X}^q(t)_\sharp\big(\rho^q \mathscr{L}^1|_W\big) \tag{11}$$

and using that $u^q \to u$ in L^1.

On the contrary, Eq. (9) cannot be deduced directly from the corresponding equation for the approximations

$$\partial_t \mathtt{X}^q(t,w) = \hat{\lambda}^q(t, \mathtt{X}^q(t,w)), \tag{12}$$

since, in general, for fixed t, $\hat{\lambda}^q(t) \circ \mathtt{X}^q(t) \nrightarrow \hat{\lambda}(t) \circ \mathtt{X}(t)$, as the following example shows.

Example 1. Assume that u is a solution of the scalar conservation law $\partial_t u + \partial_x F(u) = 0$, taking values in the finite set $\{u^L, u^M, u^R\}$, with $u^L, u^M, u^R \in \mathbb{R}$ and $u^L < u^M < u^R$, as described in Fig. 1.

Assume that the sequence of approximations $(u^q)_q$ is given by $u^q(t,x) := u(t - 1/q, x)$. Notice now that at time $\bar{t}$, $u^q(\bar{t},\cdot)$ is made by two consecutive jumps, while

[5] This can be done, for instance, in the following way. Assume for simplicity $u^q(t,\cdot)$ is right continuous. Set $\bar{U}^q(x) := \text{Tot.Var.}(\bar{u}^q; (-\infty, x])$. Set $W^q := (0,\ \text{Tot.Var.}(\bar{u}^q))$,

$$\mathtt{X}^q(0,w) := (\bar{U}^q)^{-1}(w), \quad \rho^q(w) := \begin{cases} 1 & \text{if } u^q \text{ has a positive jump at } \mathtt{X}^q(0,w), \\ -1 & \text{if } u^q \text{ has a negative jump at } \mathtt{X}^q(0,w). \end{cases}$$

Set also for simplicity $\mathtt{u}^q(w) := \int_0^w \rho^q(w')dw'$. Denote by $\{(t_j, x_j)\}_j$ the points in the (t,x)-plane where two wavefronts in u^q collide (the discontinuity points at $t = 0$ are treated as collision points). By recursion, assume $\mathtt{X}^q(t,\cdot)$ is defined on $[0, t_j]$ and let us define it on $(t_j, t_{j+1}]$. Assume that at (t_j, x_j) the outgoing Riemann problem is (u^L, u^R) with $u^L < u^R$ (the case $u^R < u^L$ is completely similar). Set $\mathtt{A}(w) := \min\{\max\{\mathtt{u}^q(w') \mid w' \le w\}, u^R\}$ for any $w \in \mathtt{X}^q(t_j)^{-1}(x_j)$ and then

$$\mathtt{X}^q(t,w) := x_j + \left[\frac{d\,\text{conv}_{[u^L,u^R]}F^q}{du}(\mathtt{A}(w))\right](t - t_j) \quad \text{for any } w \in \mathtt{X}^q(t_j)^{-1}(x_j) \text{ and any } t \in (t_j, t_{j+1}].$$

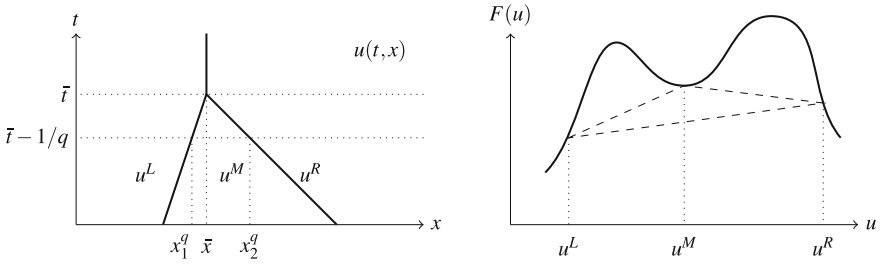


Fig. 1 The solution $u(t,x)$ to the scalar conservation law $\partial_t u + \partial_x F(u) = 0$ and the function $F(u)$

$u(\bar t, \cdot)$ is made by a single jump, which is given, roughly speaking, by the juxtaposition of the two jumps in the approximations.

By (9), all the waves w located in x_1^q have speed

$$\sigma_1 := \partial_t \mathrm{X}^q(\bar t, w) = \hat\lambda^q(\bar t, x_1^q) = \frac{F(u^M) - F(u^L)}{u^M - u^L},$$

all the waves w located in x_2^q have speed

$$\sigma_2 := \partial_t \mathrm{X}^q(\bar t, w) = \hat\lambda^q(\bar t, x_2^q) = \frac{F(u^R) - F(u^M)}{u^R - u^M},$$

while in the exact solution all the waves w should have speed

$$\sigma := \partial_t \mathrm{X}(\bar t, w) = \hat\lambda(\bar t, \bar x) = \frac{F(u^R) - F(u^L)}{u^R - u^L}.$$

Unfortunately, in general $\sigma_1, \sigma_2 \neq \sigma$ and thus $\hat\lambda^q(\bar t) \circ \mathrm{X}^q(\bar t) \nrightarrow \hat\lambda(\bar t) \circ \mathrm{X}(\bar t)$ as $q \to \infty$.

To overcome this problem and recover (9), we can proceed as follows (the argument is taken from [4]). From (6) and (8), we get for every $\varphi \in C_c^\infty((0,\infty) \times \mathbb{R})$

$$\iint u \partial_t \partial_x \varphi dx dt = \iint \partial_x \varphi(t,x) \hat\lambda(t,x) \partial_x u(t, dx) dt = \iint \partial_x \varphi(t, \mathrm{X}(t,w)) \hat\lambda(t, \mathrm{X}(t,w)) \rho(w) dw dt.$$

On the other side, testing (8) against $\partial_t \varphi$, we get

$$\iint u \partial_t \partial_x \varphi dx dt = -\iint \partial_t \varphi(t, \mathrm{X}(t,w)) \rho(w) dw dt = \iint \partial_x \varphi(t, \mathrm{X}(t,w)) \partial_t \mathrm{X}(t,w) \rho(w) dw dt.$$

Therefore,

$$\partial_x \Big[\mathrm{X}(t)_\sharp \Big(\rho(\hat\lambda(t, \mathrm{X}(t,\cdot)) - \partial_t \mathrm{X}(t,\cdot)) \mathscr{L}^1|_W \Big) \Big] = 0.$$

Equation (9) follows, just observing that $\mathtt{X}(t,\cdot)$ takes values in a compact set and that $\partial_t \mathtt{X}(t,w)$ is constant on waves having the same position (since $w \mapsto \mathtt{X}(t,w)$ is increasing).

3 Analysis of Linear Systems of Conservation Laws

We wish now to extend the scalar analysis done in the previous section to the system case. As a first step in this direction, let us study the linear system of conservation laws

$$\partial_t u + A\partial_x u = 0, \quad \text{where } A \text{ is a } n \times n \text{ strictly hyperbolic matrix}, \tag{13}$$

together with an initial datum $u(0,\cdot) = \bar{u} \in BV(\mathbb{R})$.

Let $\lambda_1, \dots, \lambda_n$ be the n distinct real eigenvalues of A, $r_1, \dots, r_n$ be the right eigenvectors (i.e., $Ar_k = \lambda_k r_k$) normalized such that $|r_k| = 1$, $l_1, \dots, l_n$ be the left eigenvectors (i.e., $l_k A = \lambda_k l_k$), normalized such that $l_k \cdot r_h = \delta_{kh}$.

Our aim is to find a good definition of *Lagrangian representation* for the solution to the linear system (13) and to explicitly construct such Lagrangian representation. This is easily done, observing that the scalar product of (13) with l_k gives the n scalar equations $\partial_t(l_k \cdot u) + \lambda_k \partial_x (l_k \cdot u) = 0$, with constant field λ_k.

Therefore, by the analysis in Sect. 2, for each k we can find a set W_k (called the *set of k-waves*), a flow $\mathtt{X}_k : [0,\infty) \times W_k \to \mathbb{R}$ and a density $\rho_k : W_k \to [-1,1]$, as in Definition 1, such that

$$\partial_x (l_k \cdot u) = \mathtt{X}_k(t)_\sharp \left(\rho_k \mathscr{L}^1|_{W_k}\right)$$

and

$$\partial_t \mathtt{X}_k(t,w) \equiv \lambda_k. \tag{14}$$

Definition 2. A *Lagrangian representation* of the solution to the linear system (13) is thus defined as a family of n triples $(W_k, \mathtt{X}_k, \rho_k)$, $k = 1, \dots, n$, (with the same regularity properties as the ones described in Definition 1) such that

$$\partial_x u(t) = \sum_{k=1}^{n} \partial_x (l_k \cdot u) r_k = \sum_{k=1}^{n} \mathtt{X}_k(t)_\sharp \left(\rho_k \mathscr{L}^1|_{W_k}\right) r_k \tag{15}$$

and the ODE (14) holds for every $k = 1, \dots, n$.

The existence of such a Lagrangian representation for the solution to the linear system (13) is then an immediate consequence of the scalar analysis done in Sect. 2.

4 The Riemann Problem

Before moving to the analysis of the nonlinear system (4), we need to recall some basic facts about the entropic solution to the *Riemann problem*, i.e., the Cauchy problem (4) with a piecewise constant initial datum

$$u(0,x)=\bar{u}(x)=\begin{cases}u^L & \text{if } x<0,\\ u^R & \text{if } x\geq 0,\end{cases}\quad \text{with } u^L,u^R\in\mathbb{R}^n \text{ close enough to } 0. \tag{16}$$

It is shown in [3] that for any $k=1,\dots,n$ it is possible to define a neighborhood $D_k\subsetneq\mathbb{R}^{n+2}$ of the point $(0,0,\lambda_k(0))\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}$ and two functions $\tilde{r}_k: D_k\to\mathbb{R}^n$, $\tilde{\lambda}_k: D_k\to\mathbb{R}$; $r_k(u_k,v_k,\sigma_k)$ (resp. $\lambda_k(u_k,v_k,\sigma_k)$) is called the *kth generalized eigenvector* (resp. the *kth generalized eigenvalue*) at $(u_k,v_k,\sigma_k)\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}$.

It is also shown in [3] that, given $u^L,u^R\in\mathbb{R}^n$ close enough to 0, one can find n curves $\gamma_k: I_k\to D_k\subseteq\mathbb{R}^{n+2}$, $k=1,\dots,n$, defined on the intervals

$$I_k:=\begin{cases}[0,s_k] & \text{if } s_k\geq 0,\\ [s_k,0] & \text{if } s_k\leq 0,\end{cases}$$

satisfying the fixed point problem (if $s_k>0$[6])

$$\begin{cases}u_k(\tau)=u_k^L+\int_0^\tau \tilde{r}_k(u_k(\varsigma),v_k(\varsigma),\sigma_k(\varsigma))d\varsigma, \quad u_k^L=\begin{cases}u^L & \text{if } k=1,\\ u_{k-1}(s_{k-1}) & \text{if } k>1,\end{cases}\\ v_k(\tau)=f_k(\tau)-\operatorname{conv}_{[0,s_k]} f_k(\tau),\\ \sigma_k(\tau)=\frac{d}{d\tau}\operatorname{conv}_{[0,\sigma_k]} f_k(\tau),\end{cases} \tag{17}$$

with f_k defined by $f_k(\tau):=\int_0^\tau\tilde{\lambda}_k(u_k(\varsigma),v_k(\varsigma),\sigma_k(\varsigma))d\varsigma$ and the $\operatorname{conv}_{[a.b]} g$ denotes the *convex envelope* of a function g on the interval $[a,b]$, i.e., the biggest convex function which stays below g.

The right-continuous solution to the Riemann problem (4), (16) is now given by the BV function

$$u(t,x)=\begin{cases}u^L & \text{if } x/t\leq\sigma_1(0),\\ u_k(\tau) & \text{if } x/t=\sigma_k(\tau),\\ u^R & \text{if } x/t\geq\sigma_n(\tau).\end{cases}$$

[6] If $s_k<0$ the convex envelope $\operatorname{conv}_{[0,s_k]} f_k(\tau)$ must be substituted by the concave envelope $\operatorname{conv}_{[s_k,0]} f_k(\tau)$.

5 Definition of Lagrangian Representation for Systems

We can now finally move to the analysis of the nonlinear system (4). As in the linear case, let $\lambda_1(u), \dots, \lambda_n(u)$ be the n distinct real eigenvalues of $A(u) := DF(u)$, $r_1(u), \dots, r_n(u)$ (resp. $l_1(u), \dots, l_n(u)$) be the right (resp. left) eigenvectors.

Trying to extend Definition 2 (and, in particular, Eqs. (14), (15)) from the linear to the nonlinear case, the first problem we have to face is that λ_k and r_k are not constant anymore, but they depend on u. As in the scalar case (see (7)), we have thus to find a good definition of kth eigenvalue $\hat{\lambda}_k(\bar{t}, \bar{x})$ and kth eigenvector $\hat{r}_k(\bar{t}, \bar{x})$ at a given point $(\bar{t}, \bar{x})$.

If $x \mapsto u(\bar{t}, x)$ is continuous at $\bar{x}$, the natural choice is to set $\hat{r}_k(\bar{t}, \bar{x}) := r_k(u(\bar{t}, \bar{x}))$ and $\hat{\lambda}_k(\bar{t}, \bar{x}) := \lambda_k(u(\bar{t}, \bar{x}))$.

If $x \mapsto u(\bar{t}, x)$ has a jump at $\bar{x}$ between $u^L := u(\bar{t}, x-)$ and $u^R := u(\bar{t}, x+)$, we solve the Riemann problem (u^L, u^R), defining the curves $(u_k(\cdot), v_k(\cdot), \sigma_k(\cdot))$ as in (17), and we set

$$\hat{r}_k(\bar{t}, \bar{x}) := \fint \tilde{r}_k(u_k(\varsigma), v_k(\varsigma), \sigma_k(\varsigma))d\varsigma, \quad \hat{\lambda}_k(\bar{t}, \bar{x}) := \fint \tilde{\lambda}_k(u_k(\varsigma), v_k(\varsigma), \sigma_k(\varsigma))d\varsigma.$$

Notice that, in the case of a scalar equation, the definition of $\hat{\lambda}(\bar{t}, \bar{x})$ given above coincides with (7).

After this preparation, we can now propose the following definition of Lagrangian representation for the solution to the nonlinear system (4). Compare it with Definitions 1 and 2.

Definition 3. A *Lagrangian representation* for the entropic solution u to (4) is a family of n triples $(W_k, \mathrm{X}_k, \rho_k)$, $k = 1, \dots, n$, where

1. $W_k \subseteq \mathbb{R}$ is a bounded interval, whose elements are called *waves of the kth family*; we also assume for simplicity that $W_k \cap W_h = \emptyset$ for $k \neq h$;
2. $\mathrm{X}_k : [0, \infty) \times W_k \to \mathbb{R}$ is $\|DF\|_{L^\infty}$-Lipschitz in t for fixed w and increasing in w for fixed t, and it is called *kth flow* or *kth position function*;
3. $\rho_k : [0, \infty) \times W_k \to [-1, 1]$ is uniformly BV in time for a.e. w, and it is called *kth density function*;

such that for a.e. $t \in [0, \infty)$

$$\partial_x u(t) = \sum_{k=1}^{n} \mathrm{X}_k(t)_\sharp\big(\rho_k(t)\mathscr{L}^1|_{W_k}\big)\hat{r}_k(t) \text{ in the sense of measures} \tag{18}$$

and

$$\frac{\partial \mathrm{X}_k}{\partial t}(t, w) = \hat{\lambda}_k(t, \mathrm{X}_k(t, w)) \text{ for } |\rho_k(t)|\mathscr{L}^1\text{-a.e.} w \in W_k. \tag{19}$$

Remark 2. The main difference between Definition 3 and Definitions 1 and 2 is that the density function $\rho = \rho(t, w)$ is now allowed to be a function of time. This

seems strange in comparison with Formula (2) for the continuity equation. However, this dependence on time cannot be avoided: It comes from the well-known fact that nonlinear interactions between wavefronts, taking place at times $t > 0$, can *create* new wavefronts.

Nevertheless, the total amount of created waves can be bounded a priori (see [2]): This implies that the length of the set of waves W_k can be bounded by $C(F)$ Tot.Var.$(\bar{u})$ and that ρ can be chosen uniformly BV in time for a.e. wave. Here, $C(F)$ is a constant which depends only on F.

6 Construction of a Lagrangian Representation

In Sect. 5, we proposed a possible definition of Lagrangian representation for the entropic solution u to the system (4). In this section, we state the main theorem of these notes, i.e., the existence of such a Lagrangian representation, and we present a sketch of its proof.

Theorem 1. *There exists a Lagrangian representation for the entropic solution to the system* (4)*, in the sense of Definition 3.*

Sketch of the proof. The proof follows a path similar to the one we used in the scalar case. We start by taking a sequence of piecewise constant approximate solutions $(u^q)_q$ (constructed through the wavefront tracking algorithm or the Glimm scheme) which converges in L^1 to the exact entropic solution u to (4).

For each u^q, it is not difficult to construct by hand a Lagrangian representation (as we did for the scalar conservation law in Sect. 2), i.e., for each $k = 1, \dots, n$, a set of k-waves W_k (which we assume to be independent of q, without restriction), a flow $\mathrm{X}_k^q : [0, \infty) \times W_k \to \mathbb{R}$ and a density $\rho_k^q : [0, \infty) \times W_k \to [-1, 1]$ such that:

- for a.e. time t $\partial_x u^q(t) = \sum_{k=1}^n \mathrm{X}_k^q(t)_\sharp \left(\rho_k^q(t)\mathscr{L}^1|_{W_k}\right)\hat{r}_k^q(t, \cdot)$ i.e. for any $\varphi \in C_c^\infty(\mathbb{R})$,

$$-\int \varphi'(x)u^q(t,x)dx = \sum_{k=1}^n \int_{W_k} \varphi(\mathrm{X}_k^q(t,w))\rho_k^q(t,w)\hat{r}_k^q(t,\mathrm{X}_k^q(t,w))dw; \quad (20)$$

- for a.e. time t and for $|\rho_k^q|\mathscr{L}^1$ almost every $w \in W_k$

$$\partial_t \mathrm{X}_k^q(t,w) = \lambda_k^q(t, \mathrm{X}_k^q(t,w)). \quad (21)$$

Exactly as in the scalar case, the regularity properties of X_k^q, ρ_k^q imply that there exist $\mathrm{X}_k : [0, \infty) \times W_k \to \mathbb{R}$, $\rho_k : [0, \infty) \times W_k \to [-1, 1]$ such that, up to subsequences, $\mathrm{X}_k^q(t) \to \mathrm{X}_k(t)$ strongly in $L^1(W_k)$ and $\rho_k^q(t) \to \rho_k(t)$ weakly* in $L^\infty(W_k)$, for a.e. time t. To complete the proof of Theorem 1, we have thus to pass to the limit in Formulae (20), (21) to get (18), (19), respectively.

In the scalar case, first we passed to the limit in (11) (corresponding here to (20)) to obtain (8) (corresponding here to (18)); then, we used (8) to prove (9)

(corresponding here to Eq. (19)). Example 1 showed that it is not possible to obtain (9) directly passing to the limit in its approximate version (12), because in general $\hat{\lambda}^q(t) \circ \mathtt{X}^q(t) \nrightarrow \hat{\lambda}(t) \circ \mathtt{X}(t)$.

In the system case, we cannot repeat the same argument (i.e., first passing to the limit in (20) to get (18) and then use (18) to prove (19)), because in (20) there is already a term $\hat{r}_k^q(t) \circ \mathtt{X}_k^q(t)$ which most likely does not converge in general to $\hat{r}_k(t) \circ \mathtt{X}_k(t)$, exactly as $\hat{\lambda}^q(t) \circ \mathtt{X}^q(t)$ did not converge in general to $\hat{\lambda}(t) \circ \mathtt{X}(t)$ in the scalar case. We thus need some new ideas to pass to the limit in (20), (21).

Example 1 shows that the are times (as time $\bar{t}$ in that example) for which there is no hope for $\hat{r}_k^q(t) \circ \mathtt{X}_k^q(t)$ (resp. $\hat{\lambda}_k^q(t) \circ \mathtt{X}_k^q(t)$) to converge to $\hat{r}_k(t) \circ \mathtt{X}_k(t)$ (resp. $\hat{\lambda}_k(t) \circ \mathtt{X}_k(t)$). However, the same example suggests that these times are *strong interaction times*, i.e., roughly speaking, times when many waves undergo a major change of their speed. For instance, in Example 1, $\hat{\lambda}^q(t) \circ \mathtt{X}^q(t) \to \hat{\lambda}(t) \circ \mathtt{X}(t)$ for every time, except the time $\bar{t}$ where a strong interaction between wavefronts takes place.

The strategy is thus to find a way to identify a priori those times of *strong interaction* in the solution u, to show that the set of such times has zero Lebesgue measure (or even that it is countable), and to prove that, up to those times, we can pass to the limit in (20), (21).

To identify such bad times, we introduce, for each approximate solution u^q, the Radon measure $\mu^q := \sum_{k=1}^n |\partial_t(\rho^q \partial_t \mathtt{X}_k^q)|$, which measure the change of the speed of the waves. Being u^q a piecewise constant function with a finite number of discontinuity lines, μ^q is just a finite sum of Dirac's deltas. For instance, for the configuration described in Example 1, μ^q is just a single Dirac's delta, located in the point $(\bar{t}, \bar{x})$, with size $|\sigma_1 - \sigma||u^M - u^L| + |\sigma_2 - \sigma||u^R - u^M|$.

Notice that, by construction of the Lagrangian representation in the approximations, for each u^q the times where waves can change their speed, i.e., times of strong interaction, are exactly those times t for which $\mu^q(\{t\} \times \mathbb{R}) > 0$.

Next, we prove that there is a Radon measure μ such that $\mu^q \to \mu$ weakly* in the sense of measures (see Remark 3 below for a comment about the existence of μ).

To conclude the proof, it is now enough to prove that if t is not a *time of strong interaction*; i.e., by definition, if t is a time such that

$$\mu\big(\{t\} \times \mathbb{R}\big) = 0 \tag{22}$$

(and this happens for all but a countable number of times), then we can pass to the limit in (20), (21) to get (18), (19), respectively. This would conclude the proof of Theorem 1.

Proving this last fact (i.e., passing to the limit in (20), (21)) is a major part of the proof of Theorem 1, which, however, requires the introduction of several *ad hoc* notations and contains rather technical steps. Therefore, in these notes, it is omitted. We just spend some words about the general strategy.

For each approximate solution u^q at each time t, through a fixed point procedure similar to the one described in Sect. 4 for solving the Riemann problem, we associate to each wave $w \in W_k$, a point

$$(\hat{u}_k^q(t,w), \hat{v}_k^q(t,w), \hat{\sigma}_k^q(t,w)) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R},$$

such that for each time t and each point $x \in \mathbb{R}$ for which $\mathtt{X}_k^q(t)^{-1}(x) \neq \emptyset$,

$$\hat{r}_k^q(t,x) \approx \fint_{\mathtt{X}(t)^{-1}(x)} \tilde{r}_k(\hat{u}_k^q(t,w'), \hat{v}_k^q(t,w'), \hat{\sigma}_k^q(t,w'))\rho(t,w')dw'$$

and, similarly, for the exact solution u at each time t, we associate to each $w \in W_k$ a point

$$(\hat{u}_k(t,w), \hat{v}_k(t,w), \hat{\sigma}_k(t,w)) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R},$$

such that for each time t and each point $x \in \mathbb{R}$ for which $\mathtt{X}_k(t)^{-1}(x) \neq \emptyset$,

$$\hat{r}_k(t,x) \approx \begin{cases} \tilde{r}_k(\hat{u}_k(t,w), \hat{v}_k(t,w), \hat{\sigma}_k(t,w)) & \text{if } u(t,\cdot) \text{ is continuous at } x = \mathtt{X}(t,w), \\ \fint_{\mathtt{X}(t)^{-1}(x)} \tilde{r}_k(\hat{u}_k(t,w'), \hat{v}_k(t,w'), \hat{\sigma}_k(t,w'))\rho(t,w')dw' & \text{if } u(t,\cdot) \text{ has a jump at } x. \end{cases}$$

Similar expressions hold for $\lambda_k^q(t,x)$, $\lambda_k(t,x)$. We then prove that if t is not a time of strong interaction, i.e. if (22) holds, then $\hat{u}_k^q \to \hat{u}_k$, $\hat{v}_k^q \to \hat{v}_k$, $\hat{\sigma}_k^q \to \hat{\sigma}_k$ in some appropriate topologies. Using this fact, we finally show that $\hat{r}_k^q(t) \circ \mathtt{X}^q(t) \to \hat{r}_k(t) \circ \mathtt{X}(t)$ and $\hat{\lambda}_k^q(t) \circ \mathtt{X}^q(t) \to \hat{\lambda}_k(t) \circ \mathtt{X}(t)$, thus concluding the proof of Theorem 1. □

Remark 3. Proving that the sequence $(\mu^q)_q$ is weakly* pre-compact in the sense of measure, i.e., proving that $|\mu^q| \leq C(f, \bar{u})$, where C is a constant which depends on f and the initial datum $\bar{u}$, but not on q, is not trivial at all. It amounts to prove that the total amount of change of speed of the waves present in an approximate solution u^q

$$\mu^q\big([0,\infty) \times \mathbb{R}\big) = \sum_{k=1}^{n} \int_{W_k} \text{Tot.Var.}\big(\rho^q(\cdot,w)\partial_t \mathtt{X}_k^q(\cdot,w); [0,\infty)\big)dw$$

is uniformly bounded by $C(f, \bar{u})$. Such estimate is proved in [6–8], using a *quadratic interaction potential*.

References

1. L. Ambrosio, N. Fusco, D. Pallara, *Functions of bounded variation and free discontinuity problems* Oxford Mathematical Monographs (The Clarendon Press, Oxford University Press, New York, 2000)

2. S. Bianchini, Interaction estimates and Glimm functional for general hyperbolic systems. DCDS-A **9**(1), 133–166 (2003)
3. S. Bianchini, A. Bressan, Vanishing viscosity solutions of nonlinear hyperbolic systems. Ann. Math. **161**, 223–342 (2005)
4. S. Bianchini, E. Marconi, On the concentration of entropy for scalar conservation laws. DCDS-S **9**(1), 73–88 (2016)
5. S. Bianchini, E. Marconi, On the structure of L^∞ solutions to scalar conservation laws in one-space dimension. Preprint SISSA (2016)
6. S. Bianchini, S. Modena, On a quadratic functional for scalar conservation laws. J. Hyperb. Differ. Equ. **11**(2), 355–435 (2014)
7. S. Bianchini, S. Modena, Quadratic interaction functional for systems of conservation laws: a case study. Bull. Inst. Math. Acad Sinica (New Series) **9**(3), 487–546 (2014)
8. S. Bianchini, S. Modena, Quadratic interaction functional for general systems of conservation laws. Comm. Math. Phys. **338**, 1075–1152 (2015)
9. S. Bianchini, S. Modena, Lagrangian representation for solution to general systems of conservation laws. In preparation (2016)
10. S. Bianchini, L. Yu, Structure of entropy solutions to general scalar conservation laws in one space dimension. J. Math. Anal. Appl. **428**(1), 356–386 (2015)
11. A. Bressan, *Hyperbolic systems of conservation laws* The one dimensional Cauchy problem (Oxford University Press, 2000)
12. C. Dafermos, Generalized characteristics in hyperbolic systems of conservation laws. Arch. Rational Mech. Anal. **107**, 127–155 (1989)
13. R.J. DiPerna, P.-L. Lions, Ordinary differential equations, transport theory and Sobolev spaces. Invent. Math. **130**, 312–366 (1989)
14. T.-P. Liu, The deterministic version of the Glimm scheme. Comm. Math. Phys. **57**, 135–148 (1977)
15. S. Modena, Interaction functionals, Glimm approximations and Lagrangian structure of BV solutions for hyperbolic systems of conservation laws. Ph.D. Thesis, Sissa digital library, 2015

Kinematical Conservation Laws in Inhomogeneous Media

S. Baskar, R. Murti and P. Prasad

Abstract The system of kinematical conservation laws (KCLs) in two dimensions involves a pair of first-order partial differential equations in a ray coordinate system written in the conservation form. The KCL system governs the evolution of a propagating front (a wavefront or a shock front) in 2D media, which involves four unknown variables, and therefore, we need additional equations to close the system. Such additional relation(s) can be obtained by a weakly nonlinear ray theory (WNLRT) for wavefront propagation and a shock ray theory (SRT) in the case of shock front propagation. The WNLRT and the SRT are well-studied for front propagation in homogeneous media and are successfully applied for an uniform medium filled with a polytropic gas. As these theories are shown to be applicable in the study of sonic boom propagation, it is important to develop these theories in the case of inhomogeneous media. This article summarizes the derivation and a basic numerical test of these two theories in an inhomogeneous medium. We also show that the derived systems are hyperbolic under the condition that the wave speed is greater than the sound speed in the unperturbed medium ahead of these waves.

Keywords Hyperbolic systems · Shock waves · Front propagation

S. Baskar (✉) · R. Murti
IIT Bombay, Bombay, Maharashtra, India
e-mail: baskar@math.iitb.ac.in

R. Murti
e-mail: rammurti@math.iitb.ac.in

P. Prasad
Indian Institute of Science Bangalore, Bangalore, Karnataka, India
e-mail: prasad@math.iisc.ernet.in

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_27

1 Introduction

Kinematical conservation laws (KCLs) are equations of evolution of curves and surfaces derived purely from geometrical consideration and since they are in conservation form, they can be used to study formation and evolution of special type of singularities, called *kinks*, on the curves and surfaces. KCL has been used successfully to study the propagation of wavefronts and shock fronts in two and three space dimensions ([1, 2]). As the system of KCL is written in a ray coordinate system, the KCL theory is more efficient in terms of computational cost. For instance, propagation of a weakly nonlinear shortwave in two space dimensions can be studied using the 2D system of Euler equations, whereas the propagation of a wavefront or a shock front in a 2D medium can also be studied using KCL which is a 1D problem. Since the KCL together with the closer relation(s) can be used to obtain the geometry of the propagating front along with the amplitude, the KCL-based theories can be used very efficiently in certain applications like propagation of sonic booms generated by a supersonic aircraft.

In this article, we shall like to develop KCL-based weakly nonlinear ray theory (WNLRT) and shock ray theory (SRT) for an inhomogeneous polytropic gas in 2D steady motion. General theory of WNLRT in an arbitrary hyperbolic system is available in Chap. 4 of [12] and the corresponding SRT in Sect. 9.2 of the same book. However, we need explicit expressions for the terms in WNLRT and SRT (which are quite involved and not easy to derive) for any application. As a future work, we wish to use these results to study the propagation of sonic boom in stratified media. Baskar and Prasad [3] successfully developed a KCL- based WNLRT and SRT in an uniform medium at rest to study sonic boom problem, and it is important to develop the method for an inhomogeneous medium for which the theory developed in this article will serve as a base.

Finding the geometry of the successive positions of a nonlinear wavefront and a shock front is a challenging problem because geometry of the fronts and their amplitude interact non-trivially (see Prasad 2001, Chap. 6 for more details). Linear ray theory (also called the geometric optics theory) is a subject of the study of geometry of a propagating front in a high-frequency approximation. This theory comprises a system of phase equation for geometry and a transport equation for the amplitude. In linear ray theory, the phase equation, called the eikonal equation, decouples from the transport equation, and therefore, the geometry of the wavefront can be found independent of the amplitude of the wave. The geometry is then used in the transport equation to get the amplitude. A notable feature of the linear ray theory is the focusing of a concave wavefront and the formation of caustic. Since the amplitude in linear ray theory is inversely proportional to the (square root of) ray tube area, the focusing of the wavefront leads to the blow-up of the amplitude, which makes the linear ray theory to be invalid in a vicinity of caustic.

In the case of nonlinear ray theory, the system of phase equations include the amplitude and therefore is coupled with the amplitude equation. The interaction of the amplitude and the geometry of a nonlinear front was first incorporated in the

pioneering work of [19]. This theory is widely used in many applications (see, for instance, [4, 11, 18]) to study the propagation of curved shock fronts.

The KCL in 2D was first derived by [9], whereas the 3D KCL was derived by [5] and further analyzed by [1]. Prasad [13] derived the KCL in space of arbitrary dimensions. Prasad and Sangeeta [14] derived KCL-based WNLRT in a 2D uniform medium with one closure relation and did extensive calculations. Monica and Prasad [8] used the new theory of shock dynamics (NTSD) to derive closure relations for the propagation of weak shock fronts in a 2D homogeneous medium with polytropic gas at rest. These closure relations can be put into a pair of conservation laws that includes another new variable which takes into account the gradient of flow behind the shock front. These two conservation laws together with the pair of KCLs, form a system of four conservation laws, which is the basic governing system for the KCL-based *shock ray theory* (SRT). The introduction of the new variable in the SRT, in fact, makes this theory more accurate than Whitham's GSD. Baskar and Prasad [2] validated the SRT by comparing the numerical results with the numerical simulation of Euler equations and also compared the SRT results with GSD. Their numerical study shows that SRT is more accurate than the GSD, especially for the cases where the flow behind the shock front plays an important role, for instance, the propagation of N-waves as in the case of sonic booms. Kevlahan [7] extended the NTSD to non-uniform flows and validated the theory. However, Kevlahan did not have the KCL and therefore used the differential form of the equations.

In Sect. 2, we recall the ray coordinate system and the suitable transformations between the ray coordinate system and the physical system. Further in this section, we also provide a brief discussion on the system of KCL in order to make the discussion self-content. In Sect. 3, we summarize the derivation of the transport equation in the ray coordinate system. This equation gives a closure relation in a conservation form, which when combined with KCL forms a closed system of equations for the WNLRT that governs the propagation of weakly nonlinear wavefronts in an inhomogeneous medium. In Sect. 4, we outline the derivation of the SRT equations, which involves the derivation of two compatibility conditions from transport equation in the ray coordinate system. These two relations with KCL also forms a closed system of equations which governs the propagation of a shock front in an inhomogeneous medium. Finally, in Sect. 5, we give numerical results, obtained from WNLRT and SRT.

2 Ray Coordinate System and Kinematical Conservation Laws

Let $\Omega_0 : \phi(\mathbf{x}, 0) = 0$, for $\mathbf{x} = (x, y) \in \mathbb{R}^2$ be an initial curve subject to a dynamics and $\Omega_t : \phi(\mathbf{x}, t) = 0$ be the position of the propagating curve at any time $t > 0$ in $\mathbb{R}^2$. The curve Ω_t can be interpreted as a curved wavefront or a shock front propagating in a medium. For a given initial front Ω_0 and a medium of propagation, our interest

is to obtain the front Ω_t at any time $t > 0$ as the locus of the tips of the rays emerging from different points on Ω_0 till the time t.

Let us denote the ray velocity by $\boldsymbol{\chi} = \boldsymbol{\chi}(\mathbf{x}, t, \mathbf{n})$, where $\mathbf{n} = \nabla\phi/|\nabla\phi|$ is the unit normal to the front. Since we work only in two-dimensions, we take $\boldsymbol{\chi} = (\chi_1, \chi_2)$ and $\mathbf{n} = (\cos\theta, \sin\theta)$, where θ denotes the angle between the normal and the x-axis. Then the normal and the tangential components of the ray velocity are given, respectively, by

$$\begin{aligned} C &= \chi_1 \cos\theta + \chi_2 \sin\theta, \\ T &= -\chi_1 \sin\theta + \chi_2 \cos\theta, \end{aligned} \tag{1}$$

which gives

$$\begin{aligned} \chi_1 &= C\cos\theta - T\sin\theta, \\ \chi_2 &= C\sin\theta + T\cos\theta \end{aligned} \tag{2}$$

Let us consider a ray coordinate system (ξ, t) in such a way that $\xi = \xi_0$ (constant) gives a ray as t varies and $t = t_0$ (constant) gives a wavefront Ω_{t_0} as ξ varies. Thus, for each fixed time $t > 0$, the wavefront in the ray coordinate system can be written as

$$\Omega_t : x = x(\xi, t), \quad y = y(\xi, t), \ \xi \in \mathbb{R}. \tag{3}$$

Let

$$g = \sqrt{x_\xi^2 + y_\xi^2} \tag{4}$$

be a metric associated with ξ in the ray coordinate system such that $g d\xi$ gives an element of length along the front Ω_t. The differential relation between an element of length in the physical coordinate system and that in the ray coordinate system can be obtained as

$$\begin{aligned} x_\xi &= -g\sin\theta, \\ y_\xi &= g\cos\theta \end{aligned} \tag{5}$$

along a front and

$$\begin{aligned} x_t &= \chi_1 = C\cos\theta - T\sin\theta, \\ y_t &= \chi_2 = C\sin\theta + T\cos\theta. \end{aligned} \tag{6}$$

along a ray. The transformation of derivatives between the physical and the ray coordinate systems is given by (see Prasad [12])

$$\begin{aligned} \frac{\partial}{\partial\xi} &= g\left(\cos\theta\frac{\partial}{\partial y} - \sin\theta\frac{\partial}{\partial x}\right) \\ \frac{\partial}{\partial t} &= \chi_1\frac{\partial}{\partial x} + \chi_2\frac{\partial}{\partial y}. \end{aligned} \tag{7}$$

Assuming that the curve Ω_t given by (3) is smooth and using the differential relations (5) and (6), we can obtain the pair of *kinematical conservation laws* (KCL)

in the (ξ, t)-coordinate system as (see [12])

$$\left.\begin{aligned}(g\sin\theta)_t + (C\cos\theta - T\sin\theta)_\xi &= 0\\ (g\cos\theta)_t - (C\sin\theta + T\cos\theta)_\xi &= 0,\end{aligned}\right\} \tag{8}$$

Note that in the case of propagation of shock fronts, we use the notation G and Θ instead of g and θ.

The KCL (8) is a system of two equations in four unknown variables, namely C, T, θ, and g, which is therefore an under-determined system. To make this system closed, we need at least two more relations. If the external velocity in the medium of propagation is negligible, then the tangential component T of the ray velocity may be taken to be zero. Therefore, we have only three unknowns, and we need at least one more relation in order to close the KCL system. This relation can be obtained from the transport equation derived from a system of hyperbolic equations governing motion of the gas under the weakly nonlinear and shortwave assumptions.

3 Weakly Nonlinear Ray Theory (WNLRT)

Consider a small perturbation in an inhomogeneous medium filled with a polytropic gas. Let the state variables of the undisturbed medium be denoted by $(\rho_0, \mathbf{q}_0 \approx \mathbf{0}, p_0)$, where ρ_0 and p_0 may depend on the position (x, y) of the medium. We assume that the wave belongs to the characteristic field corresponding to the forward eigenvalue $\langle \mathbf{n}, \mathbf{q}\rangle + a$ (a similar derivation can be given for the backward characteristic field). Then, the perturbed state variables under high-frequency approximation are given by

$$\rho - \rho_0 = \frac{\rho_0}{a_0} w; \tag{9}$$

$$q_\alpha - q_{\alpha 0} = n_\alpha w, \qquad \alpha = 1, 2; \tag{10}$$

$$p - p_0 = \rho_0 a_0 w, \tag{11}$$

where w is the wave amplitude, which is assumed to be of small order, say $\varepsilon > 0$, and $a_0^2 = \gamma p_0/\rho_0$ is the sound speed of the unperturbed medium.

Using the bicharacteristic lemma (see [12]) for the system of Euler equations of the polytropic gases and for the forward facing wave, we get

$$\left.\begin{aligned}\frac{d\mathbf{x}}{dt} &= \mathbf{q} + \mathbf{n}a := \chi,\\ \frac{d\mathbf{n}}{dt} &= -(\mathbf{L}a + n_\alpha \mathbf{L} q_\alpha),\end{aligned}\right\} \tag{12}$$

where $\mathbf{L} = \nabla - \mathbf{n}\langle \mathbf{n}, \nabla\rangle$ and a summation convention is used on the right-hand side of the second equation over the repeated index α. Using the relations (9)–(11), these

equations can be written as

$$\frac{d\mathbf{x}}{dt} = \mathbf{n}a_0 + \frac{\gamma+1}{2}\mathbf{n}w, \tag{13}$$

$$\frac{d\mathbf{n}}{dt} = -\mathbf{L}a_0 - \frac{(\gamma+1)}{2}\mathbf{L}w. \tag{14}$$

The transport equation corresponding to the forward characteristic field, which after using (9)–(11) gives

$$\frac{dw}{dt} = \left(k + a_0\Omega\right)w, \tag{15}$$

where

$$k = -\sum_{\alpha=1}^{2}\frac{1}{2}\frac{n_\alpha}{\rho_0}\frac{\partial(\rho_0 a_0)}{\partial x_\alpha} \tag{16}$$

and

$$\Omega = -\frac{1}{2}\langle\nabla, \mathbf{n}\rangle \tag{17}$$

is the mean curvature of the propagating wavefront Ω_t.

3.1 Transport Equation in the Ray Coordinate System

Using the transformation (7), the mean curvature of the wavefront in the ray coordinate system can be obtained as

$$\Omega = -\frac{1}{2g}\theta_\xi.$$

Also differentiating g in (4) with respect to t and using the ray velocity (13), we arrive at

$$g_t = \left\{a_0 + \left(\frac{\gamma+1}{2}\right)w\right\}\theta_\xi.$$

Substituting the above expression for the mean curvature in (15) and then eliminating θ_ξ from g_t, we get (after a suitable re-arrangement of the terms) the transport equation in the ray coordinate system as

$$\left[ln\left\{ge^{2\frac{(m-a_0)}{a_0}}(m-a_0)^2\right\}\right]_t = \mathscr{A}(\xi, t, m, \theta), \tag{18}$$

where

$$\mathscr{A} = -2\frac{(m-a_0)}{a_0^2}a_{0t} + \frac{2k}{a_0}m \tag{19}$$

and

$$m = a_0 + \left(\frac{\gamma+1}{2}\right)w. \tag{20}$$

The KCL (8) along with the closure relation (18)–(20) forms a closed system of equations, called the *WNLRT system*, which governs the propagation of a weakly nonlinear wavefront in the ray coordinate system (ξ, t). The eigenvalues of the WNLRT system are given by

$$\lambda_1 = -\sqrt{\frac{a_0(m-a_0)}{2g^2}},\quad \lambda_2 = 0,\ \text{and}\ \lambda_3 = \sqrt{\frac{a_0(m-a_0)}{2g^2}},$$

which is hyperbolic provided $m > a_0$.

4 Shock Ray Theory (SRT)

In this section, we derive the closure relations for KCL in the case when the propagating curve is a shock front. As in the previous section, we assume the medium in which the shock propagates is inhomogeneous and is filled with a polytropic gas, where the external velocity is negligible.

As in the case of wavefronts, we denote a shock front in the ray coordinate system by

$$\Omega_t : X = X(\xi, t),\quad Y = Y(\xi, t),\ \xi \in \mathbb{R},$$

and use the notation $\mathbf{X} = (X, Y) \in \mathbb{R}^2$. In order to distinguish the normal to the wavefront and the normal to the shock front, we denote the later by $\mathbf{N} = (\cos\Theta, \sin\Theta)$, where Θ is the angle between the normal to the shock front and the x-axis. Also, we use the notation $w = \varepsilon\tilde{w}$ for the wave amplitude so that $\tilde{w} = O(1)$ and take

$$\mu = \tilde{w}\Big|_{\text{shock front}} = \frac{w}{\varepsilon}\Big|_{\text{shock front}}.$$

A shock front is followed by a 1-parameter family of nonlinear wavefronts in the same mode. A wavefront instantaneously behind the shock front interacts with the shock and disappears. Due to the shortwave assumption, the wavefront instantaneously behind the shock front coincides with the shock front, and therefore, the ray equations (13)–(14) of the WNLRT may be used for the flow behind the shock front. Whereas, for the flow ahead of the shock front, linear rays can be used. Since the shock ray velocity and the shock front rate of rotation is the mean of the ray velocity just behind and ahead of the shock front, the shock ray equations in our case can be

written as (see [6, 12])

$$\frac{d\mathbf{X}}{d\tau} = \frac{1}{2}\left\{a_0\mathbf{N} + \mathbf{N}\left(a_0 + \varepsilon\frac{\gamma+1}{2}\mu\right)\right\}$$
$$= \mathbf{N}\left(a_0 + \varepsilon\frac{\gamma+1}{4}\mu\right) \tag{21}$$

and

$$\frac{d\mathbf{N}}{d\tau} = \frac{1}{2}\left\{-\mathbf{L}_s a_0 + \left(-\mathbf{L}_s a_0 - \varepsilon\frac{(\gamma+1)}{2}\mathbf{L}_s\mu\right)\right\}$$
$$= -\left(\mathbf{L}_s a_0 + \varepsilon\frac{(\gamma+1)}{4}\mathbf{L}_s\mu\right), \tag{22}$$

where τ is the time measured while moving along a shock ray and $\mathbf{L}_s$ is the tangential operator $\nabla - \mathbf{N}\langle\mathbf{N}, \nabla\rangle$ on the shock front Ω_t.

We write the transport equation on a shock front Ω_t in terms of the mean ray velocity as

$$\frac{d\mu}{d\tau} \equiv \left\{\frac{\partial}{\partial t} + \left(a_0 + \varepsilon\frac{\gamma+1}{4}\mu\right)\langle\mathbf{N}, \nabla\rangle\right\}\mu$$
$$= (k + a_0\Omega)\mu - \varepsilon\frac{\gamma+1}{4}\mu\langle\mathbf{N}, \nabla\rangle\tilde{w}, \tag{23}$$

where

$$\Omega = -\frac{1}{2}\langle\nabla, \mathbf{N}\rangle \tag{24}$$

is the mean curvature of the shock front. The derivative $\langle\mathbf{N}, \nabla\rangle\tilde{w}$ does not make sense on the shock front and so we introduce the new variables defined on the shock front as

$$\mu_1 = \varepsilon\left\{\langle\mathbf{N}, \nabla\rangle\tilde{w}\right\}\Big|_{\text{shock front}} \quad \text{and} \quad \mu_2 = \varepsilon^2\left\{\langle\mathbf{N}, \nabla\rangle^2\tilde{w}\right\}\Big|_{\text{shock front}}, \tag{25}$$

where the power of ε appears to make both μ_1 and μ_2 of $O(1)$ since, in shortwave approximation, variation of $\tilde{w}$ with respect to the fast variable ϕ/ε is of $O(1)$. For obtaining second compatibility condition, we differentiate (23) (with μ replaced by w, for more details see [8]) in the ray direction $\langle\mathbf{N}, \nabla\rangle$ to get

$$\frac{d\mu_1}{dt} = \left(k + a_0\Omega\right)\mu_1 - \left(\frac{\gamma+1}{2}\right)\mu_1^2, \tag{26}$$

where we have omitted those terms which contains μ_2 using the proposal made for the new theory of shock dynamics (see [12] for more details). In terms of the primitive

variables (M, Θ, G) in the ray coordinate system (ξ, t), the Eqs. (23) and (26) can be written as

$$\left[ln\left\{G(M-a_0)^2 \exp\left(2\frac{(M-a_0)}{a_0}\right)\right\}\right]_t = \left(\frac{2M}{a_0}\right)(k-V) - 2\left(\frac{M-a_0}{a_0^2}\right)a_{0t} \tag{27}$$

and

$$\left[ln\left\{GV^2 \exp\left(2\frac{(M-a_0)}{a_0}\right)\right\}\right]_t = 2\left(\frac{M-a_0}{a_0}\right)\left(k-V-\frac{a_{0t}}{a_0}\right) + 2(k-2V), \tag{28}$$

where

$$M = \left(a_0 + \varepsilon\frac{\gamma+1}{4}\mu\right), \quad V = \left(\frac{\gamma+1}{4}\right)\mu_1.$$

Eqs. (27) and (28) are first and second compatibility conditions defined on a shock front Ω_t in the ray coordinate system. These two compatibility conditions with KCL (8) (by replacing) form a closed system of equations in the ray coordinate system, called the *SRT system*, which governs the propagation of a shock front Ω_t. The eigenvalues of the SRT system are given by

$$\lambda_1 = -\sqrt{\frac{a_0(M-a_0)}{2G^2}}, \ \lambda_2 = \lambda_3 = 0, \ \ \lambda_4 = \sqrt{\frac{a_0(M-a_0)}{2G^2}}.$$

Clearly, this system is hyperbolic provided $M > a_0$.

5 Numerical Results

We have derived the closure relations for the KCL in the case of WNLRT for wavefront propagation and SRT for the shock front propagation. Using these theories, we can study the geometry and the amplitude of the propagating fronts (wavefront or shock front). To this end, we first solve the KCL system with appropriate closure relations in the ray coordinate system and obtain the primitive variables (m, θ, g) in the case of wavefront propagation and (M, Θ, G, V) for shock front propagation. Then, we transform these variables into the physical coordinate system to obtain the geometry of the propagating front by solving the system of ODEs (5) or equivalently, the system of ODEs (6), with appropriate initial conditions. Note that the WNLRT and the SRT systems may not be solvable exactly to get a closed form solution. So, we use the MUSCL scheme with local Lax–Friedrichs flux (see [17]) in order to get the primitive variables in the ray coordinate system and use the trapezoidal rule to

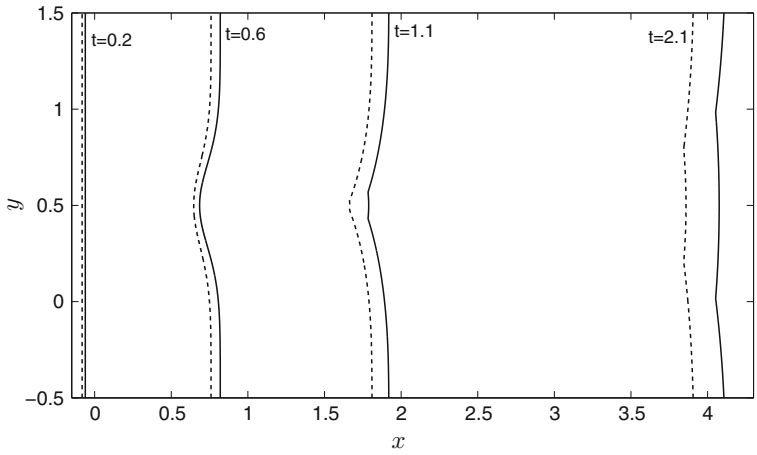


Fig. 1 Propagating wavefront and shock front at different times are obtained as solutions of the WNLRT and SRT systems, which are represented by solid and dashed curves, respectively, in the physical coordinate system (x, y)

integrate the ODEs (6) and get the propagating front as the locus of the tip of the rays emerging from different points on the initial front up to a given time t.

Since the KCL along with the closer relation of both WNLRT and SRT are hyperbolic (under a restricted condition, which we always assume), the weak (entropy) solution of the governing system involves elementary waves in the ray coordinate system, which when transformed into the physical coordinate system, gives rise to the elementary shapes. These elementary shapes are the base for the geometry of propagating fronts in the physical coordinate system.

To illustrate the geometry of a propagating front, let us consider an inhomogeneous medium with the sound speed

$$a_0(x, y) = 2 - 0.8e^{-10(x-0.7)^2} e^{-10(y-0.5)^2}, \tag{29}$$

where the external velocity is neglected.

Note that the sound speed has a circular region where the speed decreases as we approach the center radially. Let us take an initially planar front, for instance, we take the initial conditions for WLNRT and SRT, respectively, as

$$\Big(m, \theta, g\Big)(\xi, 0) = \Big(m_0(\xi), \theta_0(\xi), g_0(\xi)\Big) = (a_0 + 0.2, 0, 1)$$

and

$$\Big(M, V, \Theta, G\Big)(\xi, 0) = \Big(M_0(\xi), V_0(\xi), \Theta_0(\xi), G_0(\xi)\Big) = (a_0 + 0.1, 0, 0, 1), \tag{30}$$

which when transformed to the physical coordinate system, gives a planar front parallel to the y-axis. As this front passes through the circular region, it bends and becomes concave as seen in Fig. 1. As a result, rays emerging from different points tend to converge and make the front to focus. In the case of the linear ray theory, two converging rays intersect each other and forms a caustic region in which the propagating front folds and becomes multi-valued. Consequently, the amplitude blows-up, whereas the nonlinear diffraction effect plays a crucial role in preventing the rays to intersect and hence the amplitude remains finite. However, the amplitude undergoes a sudden change due to the presence of shocks in the primitive variables in the ray coordinate system as depicted in Figs. 2, 3, and 4. These shocks when transformed into the physical coordinate system gives rise to a pair of kinks (also called shock–shock by Whitham [20]) as seen on the fronts at $t = 1.1$ and $t = 2.1$ in Fig. 1.

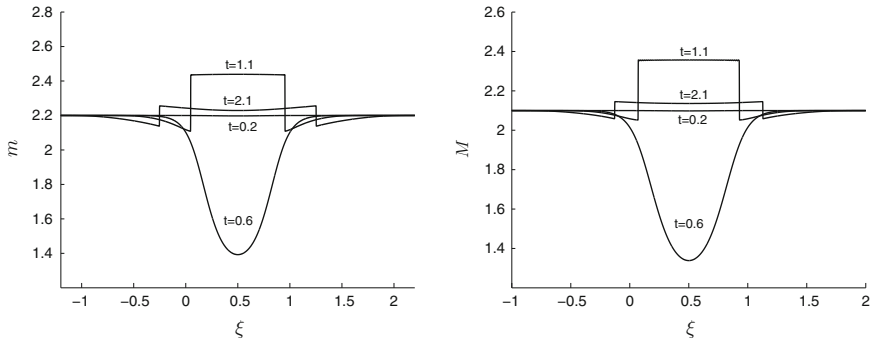


Fig. 2 Wavefront speed m (left figure) and the shock front speed M (right figure) in the ray coordinate system (ξ, t) at different times

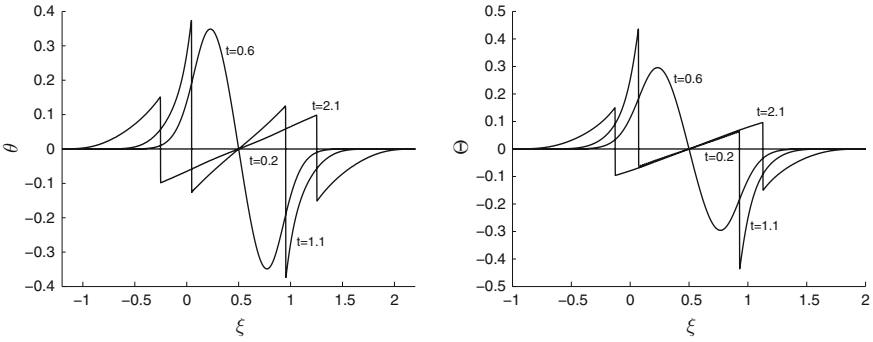


Fig. 3 Angles θ (left figure) and Θ (right figure) in the ray coordinate system (ξ, t) at different times

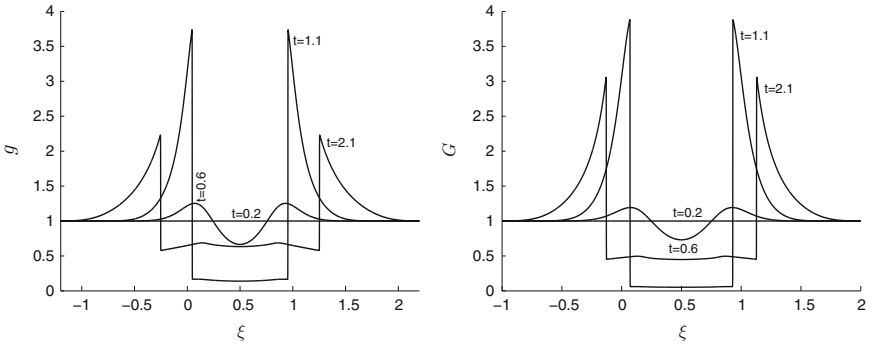


Fig. 4 Metric functions g (left figure) and G (right figure) in the ray coordinate system (ξ, t) at different times

Acknowledgements The first author thank the Council of Scientific and Industrial Research (CSIR), Government of India for giving stipend during his Ph.D. The second author's research is supported by the Department of Science and Technology, Government of India, under the project SR/FTP/MS-013/2010.

References

1. K.R. Arun, P. Prasad, 3-D Kinematical Conservation Laws (KCL): Equations of evolution of a surface. Wave Motion **46**, 293–311 (2009)
2. S. Baskar, P. Prasad, Propagation of curved shock fronts using shock ray theory and comparison with other theories. J. Fluid Mech. **523**, 171–198 (2005)
3. S. Baskar, P. Prasad, Formulation of the problem of sonic boom by a maneuvering aerofoil as a one-parameter family of Cauchy problems. Proc. Indian Acad. Sci. Math. Sci. **116**, 97–119 (2006)
4. J.E. Cates, B. Sturtevant, Shock wave focusing using geometrical shock dynamics. Phys. Fluids **9**, 3058–3068 (1997)
5. M. Giles, P. Prasad, R. Ravindran, *Conservation form of equations of three dimensional front propagation* (Department of Mathematics, Indian Institute of Science, Bangalore, 1996) Technical Report
6. E. Godlewski, P.A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws* (Springer, New York, 1996)
7. N.K.-R. Kevlahan, The propagation of weak shocks in non-uniform flows. J. Fluid Mech. **327**, 161–197 (1996)
8. A. Monica, P. Prasad, Propagation of a curved weak shock. J. Fluid Mech. **434**, 119–151 (2001)
9. K.W. Morton, P. Prasad, R. Ravindran, Conservation form of nonlinear ray equations. Technical Report 2 (Department of Mathematics, Indian Institute of Science, Bangalore, 1992)
10. K.J. Plotkin, State of the art of sonic boom modeling. J. Acoust. Soc. Amer. **111**, 530–536 (2002)
11. K.J. Plotkin, C. Hobbs, V. Sparrow, J. Salamone, K. Elmer, H. R. Welge, J. Ladd, D. Maglieri, A. Piacsek, Superboom caustic analysis and measurement program (scamp) final report. Final Report NASA/CR-2015-218871, WR 12-21, DRFC-E-DAA-TN24049, (NASA Langley Research Center (Wyle), VA 22161, 2015)

12. P. Prasad, *Nonlinear hyperbolic waves in multi-dimensions*, vol. 121 (Chapman & Hall/CRC, Boca Raton, FL, 2001)
13. P. Prasad, Kinematical conservation laws in a space of arbitrary dimensions. Indian J. Pure and Appl. Math. **47**, 641–653 (2016)
14. P. Prasad, K. Sangeeta, Numerical simulation of converging nonlinear wavefronts. J. Fluid Mech. **385**, 1–20 (1999)
15. B. Sturtevant, The physics of shock focusing in the context of extracorporeal shock wave lithotripsy. Tohoku University, Tohoku, 1989, in *Proceedings of the International Workshop on Shock Wave Focusing* (Shock Wave Research Center, Institute of Fluid Science)
16. B. Sturtevant, V.A. Kulkarny, The focusing of weak shock waves. J. Fluid Mech. **73**, 651–671 (1976)
17. E.F. Toro, *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*, 3rd edn. (Springer, 2009)
18. P.A. Varadarajan, P.L. Roe, Geometrical shock dynamics and engine unstart, in *41st AIAA Fluid Dynamics Conference and Exhibit, AIAA 2011-3909* (2011), pp. 1–16
19. G.B. Whitham, On the propagation of weak shock waves. J. Fluid Mech. **1**, 290–318 (1956)
20. G.B. Whitham, *Linear and nonlinear waves*. (Wiley-Interscience, 1974)

Artificial Viscosity for Correction Procedure via Reconstruction Using Summation-by-Parts Operators

Jan Glaubitz, Philipp Öffner, Hendrik Ranocha and Thomas Sonar

Abstract We focus on spectral viscosity in the framework of correction procedure via reconstruction (CPR, also known as flux reconstruction) using summation-by-parts (SBP) operators. In Ranocha et al. (J Comput Phys 342:13–28, 2017), [10], Ranocha et al. (J Comput Phys 311:299–328, 2016), [9], the authors used SBP operators in the CPR framework and were able to recover and extend some results of Gassner (SIAM J Sci Comput 35(3):A1233–A1253, 2013), [1] and Vincent et al. (Comput Methods Appl Mech Eng 296:248–272, 2015), [12]. In this contribution, we introduce a viscosity term for a scalar conservation law and analyse this new setting in the context of CPR methods using SBP operators. We derive conditions on the viscosity term and the basis, which allow us to prove conservation and stability in the semidiscrete setting. Next, we extend semidiscrete stability results to fully discrete stability by an explicit Euler method. Numerical tests are presented, which verify our results (Ranocha, Enhancing stability of correction procedure via reconstruction using summation-by-parts operators I: artificial dissipation, 2016, [8]). This is an extension of the contribution *Correction Procedure via Reconstruction Using Summation-by-parts Operators* by Hendrik Ranocha (J Comput Phys 342:13–28, 2017), [10], Ranocha et al. (J Comput Phys 311:299–328, 2016), [9].

Keywords Artificial dissipation · Summation-by-parts · Correction procedure via reconstruction · Spectral viscosity · Flux reconstruction

J. Glaubitz · P. Öffner (✉) · H. Ranocha · T. Sonar
Institute Computational Mathematics, TU Braunschweig, Pockelsstraße 14,
38106 Braunschweig, Germany
e-mail: p.oeffner@tu-bs.de

J. Glaubitz
e-mail: j.glaubitz@tu-bs.de

H. Ranocha
e-mail: h.ranocha@tu-bs.de

T. Sonar
e-mail: t.sonar@tu-bs.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_28

Mathematics Subject Classification (2010) 65M12 · 65M60 · 65M70

1 Introduction

We continue the consideration of the last contribution by Ranocha, and we examine the correction procedure via reconstruction (CPR), but we focus on the application of *artificial dissipation / spectral viscosity* in the context of CPR methods.

The CPR method is a high-order numerical scheme for conservation laws introduced by Huynh [3], unifying some discontinuous Galerkin, spectral difference and spectral volume in a common framework. In [9], *summation-by-parts* (SBP) operators and *simultaneous approximation terms* (SATs) have been used to create provably stable semidiscretisations. However, by using a simple explicit Euler method as time discretisation, we get stability problems in the fully discrete scheme, as described inter alia in Sect. 3.9 of [9].

Artificial dissipation/spectral viscosity has already been used in the early works of von Neumann and Richtmyer [13] to improve the stability properties of numerical schemes for conservation laws. In [6], the authors used artificial dissipation in the finite different (FD) framework of SBP operators and SATs. There are plenty of further developments, investigations and results about artificial dissipation/spectral viscosity, see inter alia [4, 5, 7, 11] and references therein.

In this contribution, we investigate the application of artificial dissipation/spectral viscosity in the CPR framework using SBP operators. We present a new approach to obtain *stable fully discrete schemes*.

Therefore, we repeat shortly the discretisation of CPR methods. Here, we only consider the linear advection and the Burgers' equation in this framework. For a more general introduction to the CPR methods using SBP operators, we strongly recommend the prior contribution of Ranocha or the papers [9, 10]. In the next section, we focus on the viscosity operator in the continuous setting and we show, that adding a viscosity term to the equation improves stability. Afterwards, we investigate the dissipation term in the semidiscrete setting and by using an explicit Euler method we get fully discrete schemes. We suggest a new algorithm to adapt the strength of the viscosity term, yielding stable fully discrete schemes, if the time step is small enough. The numerical experiments in Sect. 4 augment our theoretical results. Finally, a conclusion is presented in Sect. 5, together with additional topics of further research. Here, we strongly recommend the connection to modal filters, described in [2].

2 CPR Methods Using SBP Operators

As it was already explained in the introduction Sect. 1, we only present the semidiscretisations of the linear advection and Burgers' equation in the CPR framework in details. We apply the same notation as in the prior contribution by Ranocha and in

the papers [9, 10]. The CPR method is a semidiscretisation applying a polynomial approximation on elements. The domain $\Omega \subset \mathbb{R}$ is split into disjoint open intervals $\Omega_i \subset \Omega$ such that $\bigcup_i \overline{\Omega_i} = \Omega$. Each element Ω_i is transferred onto a standard element, which is in our case simply $[-1, 1]$. All calculations are conducted within this standard element. Let $\mathbb{P}^p$ be the space of polynomials of degree $\leq p$ and the solution u is approximated by a polynomial $U \in \mathbb{P}^p$ and in the basic formulation a nodal Lagrange basis is employed. The coefficients of $\underline{u}$ are given by the nodal values $\underline{u}_i = u(\zeta_i), i \in \{0, \ldots, p\}$, where $-1 \leq \zeta_i \leq 1$ are interpolation points in $[-1, 1]$. It can be written in $U(\xi) = \sum_{i=0}^{p} \underline{u}_i l_i(\xi)$, where $l_i(\xi)$ is the ith Lagrange interpolation polynomial that satisfies $l_j(\xi_j) = \delta_{ij}$. The flux $f(u)$ is also approximated by a polynomial, where the coefficients are given by $\underline{f}_i = f\left(\underline{u}_i\right) = f\left(u(\zeta_i)\right)$. The divergence of $\underline{f}$ is $\underline{\underline{D}}\,\underline{f}$, where we apply a discrete derivative matrix $\underline{\underline{D}}$. Since the solutions will probably have discontinuities across elements, we will have this in the discrete flux, too. To avoid this problem, we introduce a numerical flux $\underline{f}^{\text{num}}$ and also a correction term $\underline{\underline{C}}$ at the boundary nodes [9]. The restriction matrix $\underline{\underline{R}}$ performs interpolation to the boundary. With respect to a chosen basis the scalar product approximating the L^2 scalar product is represented by a matrix $\underline{\underline{M}}$ and integration with respect to the outer normal by $\underline{\underline{B}}$. Finally, all operators are introduced and they have to fulfil the SBP property

$$\underline{\underline{M}}\,\underline{\underline{D}} + \underline{\underline{D}}^T\,\underline{\underline{M}} = \underline{\underline{R}}^T\,\underline{\underline{B}}\,\underline{\underline{R}}, \tag{1}$$

in order to mimic integration by parts on a discrete level

$$\underline{u}^T\,\underline{\underline{M}}\,\underline{\underline{D}}\,\underline{v} + \underline{u}^T\,\underline{\underline{D}}^T\,\underline{\underline{M}}\,\underline{v} \approx \int_\Omega u\,(\partial_x v) + \int_\Omega (\partial_x u)\,v = u\,v\big|_{\partial\Omega} \approx \underline{u}^T\,\underline{\underline{R}}^T\,\underline{\underline{B}}\,\underline{\underline{R}}\,\underline{v}. \tag{2}$$

Linear Advection

The linear advection equation with constant velocity is a scalar conservation law with linear flux $f(u) = u$, i.e.

$$\partial_t u + \partial_x u = 0. \tag{3}$$

The semidiscretisation of this Eq. (3) is given by

$$\partial_t \underline{u} = -\underline{\underline{D}}\,\underline{u} - \underline{\underline{C}}\left(\underline{f}^{\text{num}} - \underline{\underline{R}}\,\underline{u}\right). \tag{4}$$

The canonical choice of the correction matrix $\underline{\underline{C}} = \underline{\underline{M}}^{-1}\,\underline{\underline{R}}^T\,\underline{\underline{B}}$ yields the semidiscretisation

$$\partial_t \underline{u} = -\underline{\underline{D}}\,\underline{u} - \underline{\underline{M}}^{-1}\,\underline{\underline{R}}^T\,\underline{\underline{B}}\left(\underline{f}^{\text{num}} - \underline{\underline{R}}\,\underline{u}\right) \tag{5}$$

in the standard element $[-1, 1]$, which is conservative across elements and stable with respect to the discrete norm $\|\cdot\|_M$ induced by $\underline{\underline{M}}$, if an adequate numerical flux is chosen, see inter alia Theorem 5 of [9].

Burgers' Equation

Burgers' equation

$$\partial_t u + \partial_x \frac{u^2}{2} = 0 \tag{6}$$

is nonlinear and since the product of two polynomials of degree $\leq p$ is in general a polynomial of degree $\leq 2p$, it has to be projected onto the lower dimensional space of polynomials of degree $\leq p$. For a nodal Gauß–Legendre or Lobatto–Legendre basis, the collocation approach is used, whereas for a modal Legendre basis, an exact multiplication of polynomials followed by an exact L_2 projection is applied (see for details [10]).

Using the $\underline{\underline{M}}$-adjoint $\underline{\underline{u}}^* = \underline{\underline{M}}^{-1}\underline{\underline{u}}^T\underline{\underline{M}}$, Ranocha et al. [10] presented the semidiscretisation

$$\partial_t \underline{u} = -\frac{1}{3}\underline{\underline{D}}\,\underline{\underline{u}}\,\underline{u} - \frac{1}{3}\underline{\underline{u}}^*\,\underline{\underline{D}}\,\underline{u} + \underline{\underline{M}}^{-1}\underline{\underline{R}}^T\,\underline{\underline{B}}\left(\underline{f}^{\text{num}} - \frac{1}{3}\underline{\underline{R}}\,\underline{\underline{u}}\,\underline{u} - \frac{1}{6}\left(\underline{\underline{R}}\,\underline{u}\right)^2\right), \tag{7}$$

which is conservative across elements and stable in the discrete norm induced by $\underline{\underline{M}}$, if an appropriate numerical flux is chosen, see Theorem 2 of [10] or Theorem 3 of the prior contribution.

3 Artificial Dissipation/Spectral Viscosity

We consider a scalar conservation law in one space dimension

$$\partial_t u(t,x) + \partial_x f\left(u(t,x)\right) = 0, \tag{8}$$

equipped with appropriate initial and boundary conditions. Adding a viscosity term on the right-hand side yields

$$\partial_t u(t,x) + \partial_x f\left(u(t,x)\right) = (-1)^{s+1}\varepsilon\left(\partial_x a(x)\partial_x\right)^s u(t,x), \tag{9}$$

where $s \in \mathbb{N}$ is the order, $\varepsilon \geq 0$ is the strength and $a : \mathbb{R} \to \mathbb{R}$ is a suitable function.

3.1 Continuous Setting

In the continuous setting, we analyse conservation and stability after the introduction of a viscosity term on the right-hand side.

In order to study conservation, we integrate equation (9) over some interval Ω and use integration by parts. We get

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u = \int_\Omega \partial_t u = -\int_\Omega \partial_x f(u) + (-1)^{s+1}\varepsilon \int_\Omega (\partial_x a \partial_x)^s\, u$$
$$\iff \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u = -f(u)\big|_{\partial\Omega} + (-1)^{s+1}\varepsilon\, a\, \partial_x\, (\partial_x a \partial_x)^{s-1}\, u\big|_{\partial\Omega}.$$

Thus, if a vanishes at the boundary $\partial\Omega$, this guarantees conservation.

Investigating L_2 stability, (9) is multiplied with u and integrated over Ω. With the entropy flux $F(u) = uf'(u)$ and integration by parts, we obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2_{L_2(\Omega)} =& \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u^2 = \int_\Omega u\, \partial_t u = -\int_\Omega u\, \partial_x f(u) + (-1)^{s+1}\varepsilon \int_\Omega u\, (\partial_x a \partial_x)^s\, u \\ =& -F(u)\big|_{\partial\Omega} + (-1)^{s+1}\varepsilon\, u\, a\, \partial_x\, (\partial_x a \partial_x)^{s-1}\, u\big|_{\partial\Omega} \\ &+ (-1)^s \varepsilon \int_\Omega (a\partial_x u)\, \partial_x\, (\partial_x a \partial_x)^{s-1}\, u.\end{aligned}$$

Assuming again that a vanishes at the boundary $\partial\Omega$ and using induction, this becomes

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2_{L_2(\Omega)} &= -F(u)\big|_{\partial\Omega} + (-1)^s\varepsilon \int_\Omega (a\partial_x u)\ \partial_x\, (\partial_x a \partial_x)^{s-1}\, u \\ &= -F(u)\big|_{\partial\Omega} + (-1)^{s+1}\varepsilon \int_\Omega \left[(\partial_x a \partial_x)\, u\right]\left[(\partial_x a \partial_x)^{s-1}\, u\right] \\ &= -F(u)\big|_{\partial\Omega} + \begin{cases} (-1)^{s+1}\varepsilon \int_\Omega \left[(\partial_x a \partial_x)^{s/2}\, u\right]^2, & s \text{ even}, \\ (-1)^s \varepsilon \int_\Omega a \left[\partial_x\, (\partial_x a \partial_x)^{\frac{s-1}{2}}\, u\right]^2, & s \text{ odd}. \end{cases}\end{aligned}$$

So, if a vanishes at the boundary $\partial\Omega$, the rate of change of the integral of the L_2 entropy $u \mapsto U(u) = \frac{1}{2}u^2$ is given by the entropy flux $F(u)$ through the surface of $\partial\Omega$ and an additional term, which is non-positive if $a \geq 0$ in Ω. Thus, the right-hand side in Eq. (9) has a stabilising effect. Under these conditions on a, we can ensure conservation and L_2 stability in the continuous setting.

If we discretise now this term, we have to be careful. We need some kind of projection, because the product of au is not in general a polynomial of degree $\leq p$ and the approximation of (au) might not be zero on $\partial\Omega$, even if a vanishes there, see for details [8].

3.2 Semidscrete setting

In the following, we assume that we have a conservative and stable scheme for the conservation law (8) and we augment this with an additional term, a discrete equivalent of the dissipative term (9). We have only to study the discretisation of the viscosity term concerning conservation and stability. In the context of CPR methods

using SBP operators, a direct discretisation of the dissipative term can be written as

$$(-1)^{s+1}\varepsilon \left(\underline{\underline{D}}\,\underline{\underline{a}}\,\underline{\underline{D}}\right)^{s} \underline{u}, \tag{10}$$

where $\underline{\underline{a}}$ represents the multiplication with a.

Analysing (10), conservation across elements and stability for $s = 1$ yields further conditions. Focussing on conservation results in

$$\underline{1}^T \underline{\underline{M}}\,\underline{\underline{D}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u} = \underline{1}^T \underline{\underline{R}}^T\,\underline{\underline{B}}\,\underline{\underline{R}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u} - \underline{1}^T \underline{\underline{D}}^T\,\underline{\underline{M}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u} = \underline{1}^T \underline{\underline{R}}^T\,\underline{\underline{B}}\,\underline{\underline{R}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u},$$

where the SBP property (1) has been used. So, the resulting scheme is conservative if and only if the projection used preserves boundary values. This is the case for a nodal Lobatto–Legendre basis including the boundary points, but not for a nodal Gauß–Legendre or a modal Legendre basis.

Considering now stability for $s = 1$, we multiply the term (10) with $\underline{u}^T \underline{\underline{M}}$ and divide by ε, yielding with the SBP property (1)

$$\underline{u}^T \underline{\underline{M}}\,\underline{\underline{D}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u} = \underline{u}^T \underline{\underline{R}}^T\,\underline{\underline{B}}\,\underline{\underline{R}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u} - \underline{u}^T \underline{\underline{D}}^T\,\underline{\underline{M}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u}.$$

As before, the boundary term does not vanish in general and also the matrix $\underline{\underline{a}}$ has to be self-adjoint and positive semidefinite with respect to $\underline{\underline{M}}$ to ensure that the last term is not positive.

We can avoid these problems by using the SPB property (1) directly in the dissipative term (10) for $s = 1$. This yields

$$\varepsilon\,\underline{\underline{D}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u} = \varepsilon\,\underline{\underline{M}}^{-1}\,\underline{\underline{M}}\,\underline{\underline{D}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u} = \varepsilon\,\underline{\underline{M}}^{-1}\left(\underline{\underline{R}}^T\,\underline{\underline{B}}\,\underline{\underline{R}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u} - \underline{\underline{D}}^T\,\underline{\underline{M}}\,\underline{\underline{a}}\,\underline{\underline{D}}\,\underline{u}\right). \tag{11}$$

Enforcing the boundary term to vanish yields for arbitrary s the discrete form

$$(-1)^{s+1}\varepsilon \left(-\underline{\underline{M}}^{-1}\,\underline{\underline{D}}^T\,\underline{\underline{M}}\,\underline{\underline{a}}\,\underline{\underline{D}}\right)^{s} \underline{u} = -\varepsilon \left(\underline{\underline{M}}^{-1}\,\underline{\underline{D}}^T\,\underline{\underline{M}}\,\underline{\underline{a}}\,\underline{\underline{D}}\right)^{s} \underline{u} \tag{12}$$

of the viscosity term. Now we study conservation and stability for this dissipative term (12). Multiplying (12) with $\underline{1}^T \underline{\underline{M}}$ results in

$$-\varepsilon\,\underline{1}^T\,\underline{\underline{D}}^T\,\underline{\underline{M}}\,\underline{\underline{a}}\,\underline{\underline{D}}\left(\underline{\underline{M}}^{-1}\,\underline{\underline{D}}^T\,\underline{\underline{M}}\,\underline{\underline{a}}\,\underline{\underline{D}}\right)^{s-1} \underline{u} = 0, \tag{13}$$

since the derivative is exact for constants and so the resulting scheme is conservative across elements.

To analyse the L_2 stability, we get by simple calculation

$$
\begin{aligned}
&- \varepsilon \, \underline{u}^T \, \underline{\underline{D}}^T \, \underline{\underline{M}} \, \underline{\underline{a}} \, \underline{\underline{D}} \Big(\underbrace{\underline{\underline{M}}^{-1} \, \underline{\underline{D}}^T \, \underline{\underline{M}} \, \underline{\underline{a}} \, \underline{\underline{D}}}_{:= \underline{\underline{A}}} \Big)^{s-1} \underline{u} \\
&= \begin{cases}
-\varepsilon \left[\left(\underline{\underline{A}} \right)^{s/2} \underline{u} \right]^T \underline{\underline{M}} \left[\left(\underline{\underline{A}} \right)^{s/2} \underline{u} \right], & s \text{ even}, \\
-\varepsilon \left[\left(\underline{\underline{A}} \right)^{\frac{s-1}{2}} \underline{u} \right]^T \underline{\underline{D}}^T \, \underline{\underline{M}} \, \underline{\underline{a}} \, \underline{\underline{D}} \left[\left(\underline{\underline{A}} \right)^{\frac{s-1}{2}} \underline{u} \right], & s \text{ odd}.
\end{cases}
\end{aligned}
$$

with $\underline{\underline{a}}$ self-adjoint. To ensure that these terms are always negative, we have to focus on the different bases and the projections, see for details [8]. Finally, this results in the following lemma.

Lemma 1 (Lemma 1 in [8]). *Augmenting a conservative and stable SBP CPR method for the scalar conservation law* (8)

$$\partial_t u + \partial_x f(u) = 0 \tag{14}$$

with the right-hand side (12)

$$- \varepsilon \left(\underline{\underline{M}}^{-1} \, \underline{\underline{D}}^T \, \underline{\underline{M}} \, \underline{\underline{a}} \, \underline{\underline{D}} \right)^s \underline{u}, \tag{15}$$

where $a|_\Omega \geq 0$ *is a polynomial fulfilling* $a|_{\partial\Omega} = 0$ *results in a conservative and stable semidiscrete scheme if*

- *a nodal basis with diagonal norm matrix* $\underline{\underline{M}}$
- *or a modal basis with exact* L_2 *norm and multiplication using exact* L_2 *projection*

is used. Bases fulfilling these conditions are nodal bases using Gauß–Legendre or Lobatto–Legendre nodes (with lumped mass matrix) and a modal Legendre basis.

3.3 Discrete setting

In order to get a working numerical scheme, a time discretisation has to be used to solve the ordinary differential equation. We apply in this work an explicit Euler method. The development in the standard element during one time step Δt is given by

$$\underline{u} \mapsto \underline{u}_+ := \underline{u} + \Delta t \, \partial_t \underline{u}. \tag{16}$$

Using now an SBP CPR semidiscretisation to compute the time derivative $\partial_t \underline{u}$ without artificial viscosity term, the norm after one Euler step is given by

$$\begin{aligned}\left\|\underline{u}_+\right\|_M^2 = \underline{u}_+^T \underline{\underline{M}}\, \underline{u}_+ &= \underline{u}^T \underline{\underline{M}}\, \underline{u} + 2\Delta t\, \underline{u}^T \underline{\underline{M}} \partial_t \underline{u} + (\Delta t)^2 \partial_t \underline{u}^T \underline{\underline{M}}\, \partial_t \underline{u} \\ &= \left\|\underline{u}\right\|_M^2 + 2\Delta t \left\langle \underline{u}, \partial_t \underline{u} \right\rangle_M + (\Delta t)^2 \left\|\partial_t \underline{u}\right\|_M^2 .\end{aligned} \tag{17}$$

The term $2\Delta t \left\langle \underline{u}, \partial_t \underline{u} \right\rangle_M$ can be estimated in terms of boundary values and can be controlled by the numerical flux. However, the last term $(\Delta t)^2 \left\|\partial_t \underline{u}\right\|_M^2$ causes problems. It is always non-negative and increases the norm. This may trigger instabilities. The main idea is now to introduce artificial dissipation and choose the parameters adequately in order to damp the energy growth. Adding the artificial viscosity term (12) with strength ε yields

$$\partial_t \underline{u}^\varepsilon = \partial_t \underline{u} - \varepsilon \left(\underline{\underline{M}}^{-1} \underline{\underline{D}}^T \underline{\underline{M}}\, \underline{\underline{a}}\, \underline{\underline{D}} \right)^s \underline{u} = \partial_t \underline{u} - \varepsilon \underline{\underline{A}}^s \underline{u}.$$

The norm after one explicit Euler step with artificial dissipation is

$$\begin{aligned}\left\|\underline{u}_+^\varepsilon\right\|_M^2 &= \left\|\underline{u}\right\|_M^2 + 2\Delta t \left\langle \underline{u}, \partial_t \underline{u}^\varepsilon \right\rangle_M + (\Delta t)^2 \left\|\partial_t \underline{u}^\varepsilon\right\|_M^2 \\ &= \left\|\underline{u}\right\|_M^2 + 2\Delta t \left\langle \underline{u}, \partial_t \underline{u} \right\rangle_M - 2\varepsilon \Delta t \left\langle \underline{u}, \underline{\underline{A}}^s \underline{u} \right\rangle_M + (\Delta t)^2 \left\|\partial_t \underline{u}^\varepsilon\right\|_M^2 .\end{aligned} \tag{18}$$

Here again, the term $\left\langle \underline{u}, \partial_t \underline{u} \right\rangle_M$ does not cause any problems. We have to focus now on the last two terms and these terms shall cancel out. Then, we get a similar estimate to the semidiscrete case, i.e.

$$\left\|\underline{u}_+^\varepsilon\right\|_M^2 = \left\|\underline{u}\right\|_M^2 + 2\Delta t \left\langle \underline{u}, \partial_t \underline{u} \right\rangle_M . \tag{19}$$

Thus, we get a conservative and stable fully discrete scheme. To cancel these two terms $-2\varepsilon \Delta t \left\langle \underline{u}, \underline{\underline{A}}^s \underline{u} \right\rangle_M + (\Delta t)^2 \left\|\partial_t \underline{u}^\varepsilon\right\|_M^2$, we have to solve this quadratic equation

$$\begin{aligned}0 &= -2\varepsilon \left\langle \underline{u}, \underline{\underline{A}}^s \underline{u} \right\rangle_M + \Delta t \left\|\partial_t \underline{u}^\varepsilon\right\|_M^2 \\ &= -2\varepsilon \left\langle \underline{u}, \underline{\underline{A}}^s \underline{u} \right\rangle_M + \Delta t \left(\left\|\partial_t \underline{u}\right\|_M^2 - 2\varepsilon \left\langle \partial_t \underline{u}, \right\rangle \underline{\underline{A}}^s \underline{u}_M + \varepsilon^2 \left\|\underline{\underline{A}}^s \underline{u}\right\|_M^2 \right),\end{aligned} \tag{20}$$

which is equivalent to

$$\varepsilon^2 \underbrace{\left(\Delta t \left\|\underline{\underline{A}}^s \underline{u}\right\|_M^2 \right)}_{=:A} + \varepsilon \underbrace{\left(-2\left\langle \underline{u}, \underline{\underline{A}}^s \underline{u} \right\rangle_M - 2\Delta t \left\langle \partial_t \underline{u}, \underline{\underline{A}}^s \underline{u} \right\rangle_M \right)}_{=:B} + \underbrace{\left(\Delta t \left\|\partial_t \underline{u}\right\|_M^2 \right)}_{=:C} = 0. \tag{21}$$

The roots of this equation for $A \neq 0$ are given by

$$\varepsilon_{1/2} = \frac{1}{2A} \left(-B \pm \sqrt{B^2 - 4AC} \right).$$

and for sufficiently small time step Δt the discriminant $B^2 - 4AC$ is non-negative as well as $-B$ and AC. We may estimate the strength in the following way

$$\varepsilon_1 \geq \varepsilon_2 = \frac{1}{2A}\left(-B - \sqrt{B^2 - 4AC}\right) \geq \frac{1}{2A}\left(-B + \sqrt{B^2}\right) = 0.$$

Finally, this results in the following lemma.

Lemma 2 (Lemma 3 of [8]). *If a conservative and stable SBP CPR method for a scalar conservation law (8)*

$$\partial_t u + \partial_x f(u) = 0 \tag{22}$$

is augmented with the artificial dissipation (12)

$$-\varepsilon \left(\underline{\underline{M}}^{-1}\, \underline{\underline{D}}^T\, \underline{\underline{M}}\, \underline{\underline{a}}\, \underline{\underline{D}}\right)^s \underline{u} \tag{23}$$

on the right-hand side, the fully discrete scheme using an explicit Euler method as time discretisation is both conservative and stable if

- *a nodal Gauß–Legendre/Lobatto–Legendre or a modal Legendre basis is used,*
- $\left\langle \underline{u}\, \underline{\underline{A}}^s\, \underline{u}, > \right\rangle 0$, *which will be fulfilled for the choice of a described below if the solution* $\underline{u}$ *is not constant,*
- *the time step* Δt *is small enough such that*

$$B^2 - 4AC > 0, \quad -B > 0, \qquad \text{if } \Delta t \text{ is small enough and } \underline{\underline{A}}^s \underline{u} \neq 0. \tag{24}$$

  is fulfilled,
- *and the strength* $\varepsilon > 0$ *is big enough.*

  If the other conditions are fulfilled, ε *has to obey*

$$\varepsilon \geq \varepsilon_2 = \frac{1}{2A}\left(-B - \sqrt{B^2 - 4AC}\right), \tag{25}$$

where A, B, *and* C *from Eq. (21) are used.*

This result gives us the approach to calculate the strength of the viscosity operator in an adaptive way to ensure stability for the fully discrete schemes.

4 Numerical results

In this section, we verify our results by two numerical test cases. We consider the linear advection (3) and Burgers' equation (6) with smooth initial conditions. Further experiments can be found in [8].

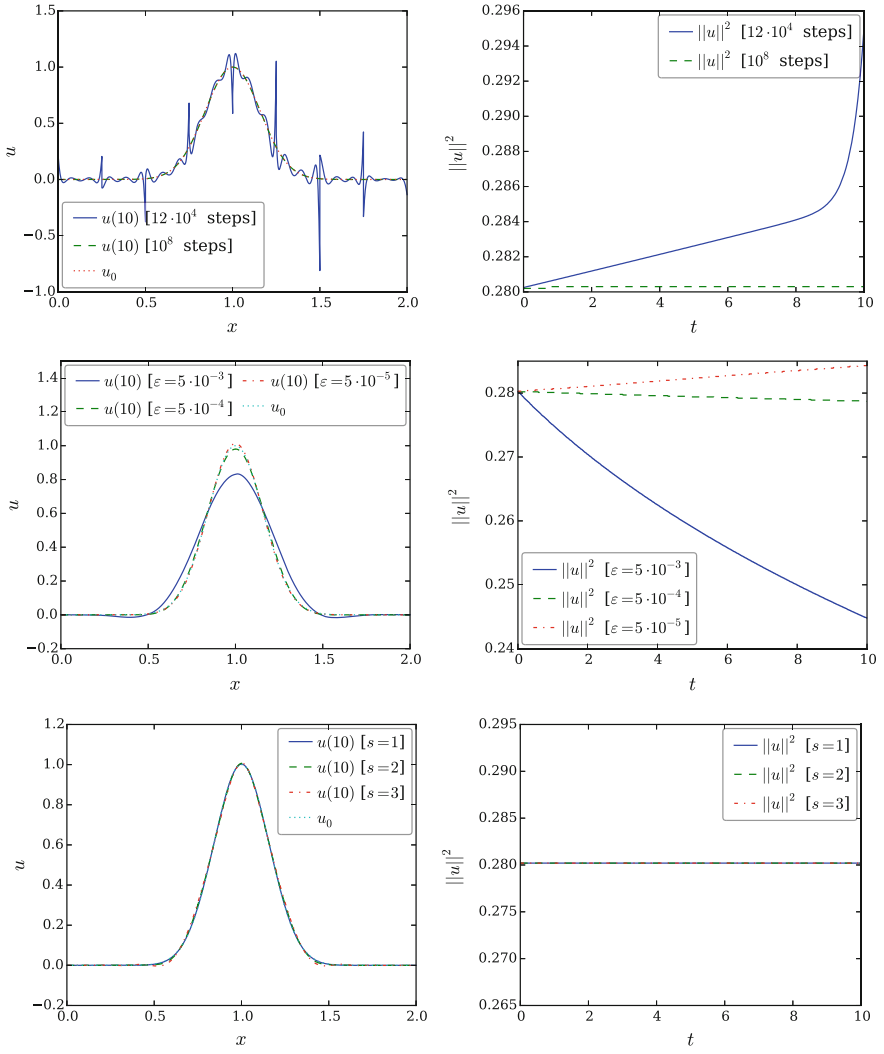


Fig. 1 The numerical solutions (left) and the energies (right) are plotted

We start with the linear advection with constant velocity

$$\partial_t u + \partial_x u = 0, \qquad u(0,x) = u_0(x) = \exp\left(-20(x-1)^2\right).$$

We calculate the numerical solution in the domain $x \in [0, 2]$, equipped with periodic boundary conditions. We choose $N = 8$ elements using a Gauß–Legendre nodal basis of degree $\leq p = 7$, a central numerical flux $f^{\text{num}}(u_-, u_+) = (u_- + u_+)/2$ and for the dissipative term $a(x) = 1 - x^2$. For the time integration, we use $12 \cdot 10^4$ steps in

the time interval [0, 10]. In Fig. 1 on the left side the numerical solution and on the right side the energy of these solutions are given.

In the first row, the computation is without additional artificial dissipation. In the second row, the calculation is done with various strengths ε and in the last row, new adaptive strategy for the strength is applied. As can be seen in the first row, the energy of the solutions using $12 \cdot 10^4$ time steps is increasing, as expected, whereas a simple artificial dissipation of fixed strength has a stabilising effect, see the plots in the second row. However, using the new adaptive technique (Lemma 2) to estimate the strength ε, the energy remains constant and the solutions look as expected. This confirms our results.

In the second test case, we consider the Burgers' equation (6) with smooth initial conditions

$$\partial_t u + \partial_x \frac{u^2}{2} = 0, \quad u(0, x) = u_0(x) = \sin \pi x + 0.01 \tag{26}$$

in the periodic domain [0, 2]. This Eq. (6) is used as a prototypical example of a nonlinear conservation law yielding a discontinuous solution in finite time $t \in [0, 3]$. The stable semidiscretisation (7) with $N = 16$ elements representing polynomials of degree $\leq p = 16$ in nodal Gauß–Legendre bases is used with the local Lax-Friedrichs

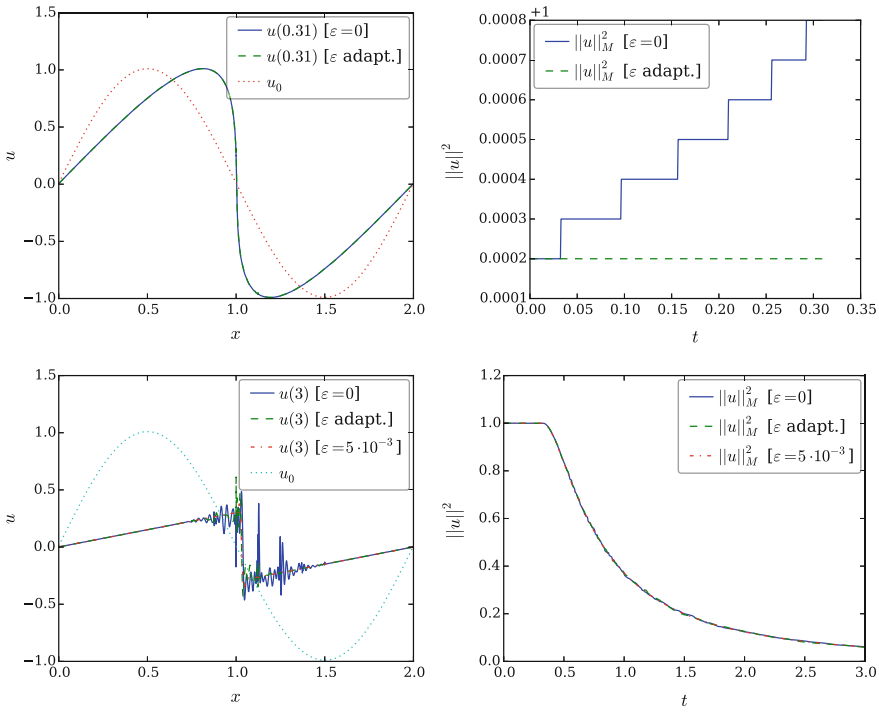


Fig. 2 First row: $t = 0.31$. Second row: $t = 3$

flux $f^{\text{num}}(u_-, u_+) = \frac{u_-^2+u_+^2}{4} - \frac{\max\{|u_-|,|u_+|\}}{2}(u_+ - u_-)$. The explicit Euler method as time integrator uses $15 \cdot 10^3$ steps for the interval $[0, 3]$.

At time $= 0.31$, the solutions computed with only 500 time steps are still smooth, but the energy increases if no artificial dissipation is employed. The application of adaptive spectral viscosity results in a constant energy, see first row in Fig. 2.

At time $t = 3$, the solutions have developed discontinuities and we see oscillations around $x \approx 1$ for the different spectral viscosity, which we use. Here, the scheme with the spectral viscosity of fixed strength $\varepsilon = 5 \cdot 10^{-3}$ demonstrates the best result. It adds enough dissipation to remove these oscillations nearly completely. Nevertheless, all three choices of spectral viscosity yield nearly visually indistinguishable results for the energy.

5 Conclusion and outlook

In this contribution, we considered artificial dissipation/spectral viscosity in the general framework of CPR methods using SBP operators. Stability and conservation results are presented and also a new adaptive strategy to get a conservative and stable fully discrete scheme by an explicit Euler method. Numerical test cases show the advantage of the chosen approach as well some limitations. The authors of [2] present an approach to overcome some of these limitations.

Further research topics are the extensions to different time integration methods and other hyperbolic conservation laws.

References

1. G.J. Gassner, A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods. SIAM J. Sci. Comput. **35**(3), A1233–A1253 (2013)
2. J. Glaubitz, H. Ranocha, P. Öffner, T. Sonar, Enhancing stability of correction procedure via reconstruction using summation-by-parts operators II: modal filtering (2016). Submitted
3. H.T. Huynh, A flux reconstruction approach to high-order schemes including discontinuous Galerkin methods. AIAA paper 2007-4079 (2007)
4. H. Ma, Chebyshev-Legendre spectral viscosity method for nonlinear conservation laws. SIAM J. Numer. Anal. **35**(3), 869–892 (1998)
5. H. Ma, Chebyshev-Legendre super spectral viscosity method for nonlinear conservation laws. SIAM J. Numer. Anal. **35**(3), 893–908 (1998)
6. K. Mattsson, M. Svärd, J. Nordström, Stable and accurate artificial dissipation. J. Sci. Comput. **21**(1), 57–79 (2004)
7. J. Nordström, Conservative finite difference formulations, variable coefficients, energy estimates and artificial dissipation. J. Sci. Comput. **29**(3), 375–404 (2006)
8. H. Ranocha, J. Glaubitz, P. Öffner, T. Sonar, Enhancing stability of correction procedure via reconstruction using summation-by-parts operators I: artificial dissipation (2016). Submitted
9. H. Ranocha, P. Öffner, T. Sonar, Summation-by-parts operators for correction procedure via reconstruction. J. Comput. Phys. **311**, 299–328 (2016)

10. H. Ranocha, P. Öffner, T. Sonar, Extended skew-symmetric form for summation-by-parts operators and varying Jacobians. J. Comput. Phys. **342**, 13–28 (2017). (Accepted)
11. E. Tadmor, Convergence of spectral methods for nonlinear conservation laws. SIAM J. Numer. Anal. **26**(1), 30–44 (1989)
12. P.E. Vincent, A.M. Farrington, F.D. Witherden, A. Jameson, An extended range of stable-symmetric-conservative flux reconstruction correction functions. Comput. Methods Appl. Mech. Eng. **296**, 248–272 (2015)
13. J. von Neumann, R.D. Richtmyer, A method for the numerical calculation of hydrodynamic shocks. J. Appl. Phys. **21**(3), 232–237 (1950)

On a Relation Between Shock Profiles and Stabilization Mechanisms in a Radiating Gas Model

Masashi Ohnawa

Abstract In the present article, the author deals with the asymptotic stability issue of traveling wave solutions or shock waves to a simplified model of radiating gas called the Hamer model. If the shock strength, defined by the difference of the two asymptotic values, exceeds a certain threshold, the shock profiles have discontinuities of the first kind. We prove that all subcritical shock waves are stable to small perturbations due to smoothing effect of radiation, while in the critical case, arbitrary small perturbations could cause blowup in a finite time. In the supercritical cases, however, convection contributes to the recovery of stability under the presence of discontinuity in the asymptotic state. This article basically reviews the author's two papers (Ohnawa, SIAM J Math Anal 46:2136–2159, 2014, [15]) and (Ohnawa, SIAM J Math Anal 48:3820–3839, 2016, [16]).

Keywords Asymptotic stability · Discontinuous shock wave · Radiation gas system

1 Introduction

The present paper deals with a hyperbolic–elliptic coupled system which was derived in Hamer [4] from a system of equations describing gas dynamics with radiation. The system reads

$$u_t + uu_x + q_x = 0, \tag{1}$$
$$-m^2 q_{xx} + q + u_x = 0, \tag{2}$$

where $u(t,x)$ and $q(t,x)$ are real-valued functions for $t \geq 0$ and $x \in \mathbb{R}$ which satisfies

M. Ohnawa (✉)
Tokyo University of Marine Science and Technology, Tokyo 108-8477, Japan
e-mail: ohnawa@m.kaiyodai.ac.jp

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_29

$$u(0,x) = u_0(x) \ \text{ with } \lim_{x\to\pm\infty} u_0(x) = u_\pm, \tag{3}$$

$$\lim_{x\to\pm\infty} q(t,x) = 0 \ \text{ for an arbitrary } \ t \geq 0, \tag{4}$$

and m is a positive constant whose reciprocal m^{-1} corresponding to the absorption coefficient.

For this system with $m = 1$, Kawashima and Nishibata [7] showed (1)–(2) admits traveling wave solutions, i.e., solutions in the form of $(u,q)(t,x) = (U,Q)(x - st)$ for a certain constant s, supplemented by $\lim_{x\to\pm\infty} U(x) = u_\pm$. (By suitably changing variables if necessary, we may assume without loss of generality that $u_- + u_+ = 0$ and $s = 0$.) Moreover, defining the shock strength δ_S by

$$\delta_S := |u_- - u_+|, \tag{5}$$

they revealed that U is smooth only when $\delta_S \leq \sqrt{2}$, and discontinuities appear at one point in the profiles of U if $\delta_S > \sqrt{2}$. Here, discontinuous solutions are defined in the sense of Kružkov [9]. Their results are easily extended to the case of general $m > 0$. To see the effects of m and δ_S, let us put

$$\tilde{\delta}_S := m\delta_S. \tag{6}$$

Proposition 1. *For $\tilde{\delta}_S > 0$, there exists a traveling wave solution to (1)–(2) uniquely up to a shift. The function U is monotonically decreasing and in a suitable coordinate U is an odd function. Moreover, it holds that*

$$|U(x) - u_S(x)| \leq \frac{1}{2}\delta_S e^{-c|x|}, \quad \textit{with } \ u_S(x) := u_\pm \ \textit{for } \ \pm x > 0, \tag{7}$$

where c is a positive constant depending only on δ_S and m.
(i) *In the cases $\tilde{\delta}_S \in (0,\sqrt{2})$ (subcritical) or $\tilde{\delta}_S = \sqrt{2}$ (critical), the solution satisfies $(U,Q) \in C^1(\mathbb{R}) \times C^2(\mathbb{R})$, and*

$$0 > U'(x) \geq U'(0) = \left(-1 + \sqrt{1 - \tilde{\delta}_S^2/2}\right)/2m^2 \ \ \textit{for } \ x \in \mathbb{R}. \tag{8}$$

(ii) *In the case $\tilde{\delta}_S > \sqrt{2}$ (supercritical), the solution is discontinuous at only $x = 0$ and*

$$0 > U'(x) > U'(\pm 0) = -1/m^2 \ \ \textit{for } \ x \in \mathbb{R}_0 := \mathbb{R} \setminus \{0\}. \tag{9}$$

These features are not intrinsic to simplified models. For the original system from which Hamer's model was derived, the appearance of discontinuities when δ_S or m are sufficiently large was known since the early works by physicists [5, 19] and is still being actively studied [11, 12]. Some of these properties of shock waves are recently mathematically validated in [2, 13].

Along with these studies, the stability of shock waves has been extensively studied. The L^1-stability was proved in a comprehensive manner by Serre [17, 18]; arbitrary strong shock waves are shown to be L^1-stable to arbitrary large L^1 perturbations. However, physically important feature of the appearance of discontinuities in the supercritical shock profiles seems to be more closely related to L^∞-stability issue. The earliest result in this direction was obtained in [7] for shock waves with $\delta_S < \sqrt{6}/2 (< \sqrt{2})$ in the case of $m = 1$. Improvements have been made in various aspects since then [3, 10, 14], but all of the results were applicable only to a portion of subcritical cases.

The objective of this paper is to review facts on L^∞-stability of traveling waves for subcritical, critical, and supercritical cases obtained by the author in [15, 16], and to give an insight into the appearance of discontinuity in the supercritical shock profiles from the view point of stability, which was revealed in [7] by phase plane analysis of the underlying system of ordinary differential equations.

More precisely, the author clarified that all subcritical shock waves are L^∞-stable to small perturbations, while the critical shock wave, if added certain types of perturbations, blows up in a finite time whatever small they may be. In the supercritical cases, however, convection contributes to the recovery of stability under the presence of discontinuity in the asymptotic state.

Notations: For a constant $p \in [1, \infty]$, $|f|_p$ denotes the canonical L^p norm of a function f. The kth-order Sobolev space in the L^2 sense is denoted by H^k and is equipped with the norm $\|\cdot\|_{H^k}$. We often simplify it as $\|\cdot\|_k$. For nonnegative integer n, we denote by $C_b^n(\mathbb{R})$ a subspace of $C^n(\mathbb{R})$ with derivatives being bounded up to nth order, which is equipped with the norm of $\|f\|_{C^n(\mathbb{R})} := \sum_{k=0}^n \sup_{x\in\mathbb{R}} |f^{(k)}(x)|$. Finally, c and C denote generic positive constants.

Here we formulate the stability problems depending on the magnitude of $\tilde{\delta}_S$ relative to the threshold value of $\sqrt{2}$.

• **Subcritical case** ($\tilde{\delta}_S < \sqrt{2}$)

We consider an initial value which satisfies

$$u_0 - u_S \in L^1, \quad \text{where } u_S(x) := u_\pm \ (\pm x > 0). \tag{10}$$

Then we may assume

$$\int_{-\infty}^{\infty} (u_0(x) - U(x))\, dx = 0, \tag{11}$$

and $U(0) = 0$ by suitably changing variables. The initial perturbation and its anti-derivative are defined by

$$\phi_0(x) := u_0(x) - U(x), \quad \Phi_0(x) := \int_{-\infty}^{x} \phi_0(y)\, dy \ \text{ for an arbitrary } x \in \mathbb{R},$$

respectively. Conditions (7) and (10) assure the well-definedness of $\Phi_0(x)$.

Theorem 1. *In the case* $\tilde{\delta}_S \in (0, \sqrt{2})$*, assume (10)–(11),*

$$\Phi_0,\ \phi_0 \in L^2(\mathbb{R}), \tag{12}$$

and

$$u_0 \in C_b^1(\mathbb{R}),\quad \inf_{x\in\mathbb{R}} u_0'(x) > \left(-1 - \sqrt{1 - \tilde{\delta}_S^2/2}\right)/2m^2. \tag{13}$$

If $\|\Phi_0\|_1$ *is sufficiently small, the initial value problem (1)–(4) has a unique global solution* $(u, q) \in C_b^1 \times C_b^2$*. The solution converges uniformly to the shock wave:*

$$\sup_{x\in\mathbb{R}} |u(t,x) - U(x),\ q(t,x) - Q(x)| \to 0 \ \ as\ \ t \to \infty. \tag{14}$$

• **Critical case** $(\tilde{\delta}_S = \sqrt{2})$
In contrast to the theorems above, the critical shock wave, which is still continuous, blows up the first-order derivative in a finite time if added certain types of perturbations whatever small they may be.

Theorem 2. *Let* (U, Q) *be a traveling wave solution in the case* $\tilde{\delta}_S = \sqrt{2}$*. For an arbitrary* $\phi_0(\not\equiv 0) \in C_b^1(\mathbb{R})$ *which satisfies*

$$\phi_0(-x) = -\phi_0(x) \ \ and\ \ \phi_0(x) \le 0 \ \ for\ \ x \ge 0, \tag{15}$$

the solution to (1)–(4) with $u_0(x) = U(x) + \phi_0(x)$ *blows up in a finite time, i.e.,*

$$\inf_{x\in\mathbb{R}} u_x(t,x) \to -\infty \ \ as\ \ t \to T_* - 0 \tag{16}$$

for a certain finite value $T_*(> 0)$*.*

Next theorem presents another kind of blowup set. It shows passing $\tilde{\delta}_S$ formally to $\sqrt{2}$ in Theorem 1 is not valid. See [15] for the proof.

Theorem 3. *Consider the case* $\tilde{\delta}_S = \sqrt{2}$*. For an arbitrary positive constant* ε*, there exists an initial data* $u_0 \in C_b^1(\mathbb{R})$ *satisfying (10)–(12),* $\|\Phi_0\|_1 < \varepsilon$ *as well as* $\inf_{x\in\mathbb{R}} u_0'(x) > -1/2m^2$ *such that the solution to (1)–(4) blows up the first order derivative in a finite time.*

• **Supercritical case** $(\tilde{\delta}_S > \sqrt{2})$
We consider a piecewise smooth initial data $u_0(x)$ which has exactly one discontinuity of the first kind. Assuming (10), there uniquely exists a traveling wave U such that $U(\pm\infty) = u_\pm$ and (11) hold. We redefine the origin to be the location of the jump in U and fix the shift of U so that (11) holds. In view of Proposition 1, the origin may be regarded as the center of mass of the initial data. Denoting the location of the discontinuity of u_0 by $x = d_0$, we define the initial perturbation and its anti-derivative by

$$\phi_0(x) := u_0(d_0 + x) - U(x), \quad \Phi_0(x) := \int_{\pm\infty}^{x} \phi_0(y)\, dy \ \text{ for } \ \pm x > 0.$$

Theorem 4. *In the case $\tilde{\delta}_S > \sqrt{2}$, assume conditions stated above and*

$$\Phi_0 \in H^3(\mathbb{R}_0), \quad \text{where } \mathbb{R}_0 := \mathbb{R} \setminus \{0\}. \tag{17}$$

If $\tilde{\delta}_S$ is sufficiently larger than $\sqrt{2}$, and $\|\Phi_0\|_3$ is sufficiently small, the initial value problem (1)–(4) has a unique global solution which satisfies

$$\phi(t,x) \in \bigcap_{k=0}^{2} C^k([0,\infty); H^{2-k}(\mathbb{R}_0)), \ \text{and} \ \ \psi(t,x) \in \bigcap_{k=0}^{2} C^k([0,\infty); H^{3-k}(\mathbb{R}_0)), \tag{18}$$

where

$$(\phi, \psi)(t,x) := (u, q)(t, d(t) + x) - (U, Q)(x). \tag{19}$$

Here $d(t)$ is a C^1-function representing the location of the sole discontinuity in $u(t,\cdot)$, which converges to the center of mass of the initial data:

$$d(t) \to 0 \ \text{ as } t \to \infty. \tag{20}$$

Furthermore, the solution converges uniformly to the shock wave:

$$\sup_{x \in \mathbb{R}_0} (\phi, \psi)(t,x) \to 0 \ \text{ as } t \to \infty. \tag{21}$$

The condition $\tilde{\delta}_S \gg \sqrt{2}$ in Theorem 4 is imposed in order to handle terms arising from the mobility of the discontinuity. In the case $u_0(x)$ is odd, so is $u(t,\cdot)$ for an arbitrary $t > 0$ and $d(t)$ is identically zero. Thus the 'largeness' of $\tilde{\delta}_S$ can be removed.

Theorem 5. *If $u_0(x)$ is an odd function, all conclusions of Theorem 4 hold for an arbitrary $\tilde{\delta}_S > \sqrt{2}$.*

In what follows, we give sketches for the proofs of Theorem 1 in Sect. 2, of Theorem 2 in Sect. 3, and of Theorem 5 in Sect. 4. Since the essence of the stability mechanism of supercritical shock waves is manifested in the proof of Theorem 5, Theorem 4 is not proved here. Interested readers are referred to [16].

2 Subcritical case (Proof of Theorem 1)

By standard arguments (e.g., [1] Sect. 3), a local solution $(u, q) \in C^1(\mathbb{R}) \times C^2(\mathbb{R})$ to (1)–(4) is obtained over a life span $[0, T]$, where T is determined by the norm $\|u_0\|_{C^1}$. Following [8], we have a maximum principle-type estimate

$$\inf_{x\in\mathbb{R}} u_0(x) \le u(t,x) \le \sup_{x\in\mathbb{R}} u_0(x), \quad u_x(t,x) \le \sup_{x\in\mathbb{R}} u_0'(x). \tag{22}$$

Variables for the perturbation defined by

$$(\phi,\psi)(t,x) := (u,q)(t,x) - (U,Q)(x)$$

satisfy

$$\phi_t + (U+\phi)\phi_x + U'\phi + \psi_x = 0, \tag{23}$$
$$-m^2\psi_{xx} + \psi + \phi_x = 0, \tag{24}$$

with the initial value given by

$$\phi(0,x) = \phi_0(x) = u_0(x) - U(x). \tag{25}$$

Lemma 1. *Assuming (10), the perturbation $\phi(t,\cdot)$ belongs to $L^1(\mathbb{R})$ and satisfies*

$$|\phi(t)|_1 \le |\phi_0|_1, \quad t\in[0,T]. \tag{26}$$

The anti-derivative of ϕ defined by $\Phi(t,x) := \displaystyle\int_{-\infty}^{x} \phi(t,y)dy$ satisfies

$$\Phi_t + U\Phi_x + \frac{1}{2}\Phi_x^2 + \psi = 0. \tag{27}$$

Furthermore, assuming (10)–(12), $\Phi(t,\cdot), \phi(t,\cdot), \psi(t,\cdot), \psi_x(t,\cdot) \in L^2(\mathbb{R})$ holds at each time $t\in[0,T]$.

Proof. L^1-boundedness (26) is obtained in [7]. L^2-boundedness is proved using Reynolds' transport theorem.

Lemma 2. *Assume (10)–(12). If $\|\Phi_0\|_1$ is sufficiently small, it holds*

$$\|\Phi(t)\|_1^2 + \int_0^t \left(|\phi(s,\cdot)|_2^2 + \|\psi(s,\cdot)\|_1^2\right) ds \le C\|\Phi_0\|_1^2 \tag{28}$$

for an arbitrary $t\in[0,T]$, where C is a positive constant independent of T.

Proof. Multiply (23) by 2ϕ and (27) by 2Φ respectively to get

$$\partial_t\Phi^2 + \partial_x\left\{U\Phi^2 + 2(\Phi + m^2\psi)(m^2\psi_x - \phi)\right\}$$
$$-U'\Phi^2 + (2+\Phi)\phi^2 - 2m^2(\psi^2 + m^2\psi_x^2) = 0, \tag{29}$$
$$\partial_t\phi^2 + \partial_x\left\{U\phi^2 + \frac{2}{3}\phi^3 - 2\psi(m^2\psi_x - \phi)\right\} + U'\phi^2 + 2(\psi^2 + m^2\psi_x^2) = 0. \tag{30}$$

From these, we first see that $\|\Phi(t,\cdot)\|_1$ grows at most exponentially fast so that $\sup_{0\le t\le T}|\Phi(t,\cdot)|_\infty$ can be arbitrary small by letting $\|\Phi_0\|_1$ be small accordingly. Under this condition, multiplication of (30) by $3m^2/2$ and addition to (29) yield the desired estimate recalling (8).

Lemma 3. *Assume the same conditions as in Theorem 1. If $\|\Phi_0\|_1$ is sufficiently small, then*

$$\inf_{x\in\mathbb{R}} u_x(t,x) \ge \min\left\{\inf_{x\in\mathbb{R}} u_0'(x), \frac{-1/2m^2 + U'(0)}{2}\right\}, \tag{31}$$

where $U'(0)$ is given in (8).

Proof. For arbitrary $t\in[0,T]$ and $x\in\mathbb{R}$, there exists a unique characteristic curve $\{(s,X(s))|s\in[0,t]\}$ which reaches (t,x). Along this trajectory, it holds

$$\frac{d}{dt}u_x(t,X(t)) = -u_x^2 - m^{-2}u_x + m^{-2}K_m * U' + m^{-2}K_m' * \phi, \tag{32}$$

where K_m is defined by $K_m(x) = \exp(-|x|/m)/2m$. Since

$$|K_m' * \phi|_\infty(t) \le |K_m'|_2|\phi|_2(t) \le C|K_m'|_2|(\Phi_0,\phi_0)|_2 \le C|(\Phi_0,\phi_0)|_2 \tag{33}$$

follows from Young's inequality and Lemma 2 provided $|(\Phi_0,\phi_0)|_2$ is sufficiently small, and

$$K_m * U'(x) = -Q(x) = \frac{U(x)^2}{2} - \frac{(\delta_S/2)^2}{2} \ge -\frac{\delta_S^2}{8}, \tag{34}$$

substitution of (33) and (34) into (32) results in

$$\begin{aligned}\frac{d}{dt}u_x(t,X(t)) &\ge -u_x^2 - m^{-2}u_x - \frac{1}{8}\delta_S^2 - |K' * \phi|_\infty(t)\\ &= -(u_x - a_-)(u_x - a_+) - |K' * \phi|_\infty(t),\end{aligned} \tag{35}$$

where $a_\pm := \left(-1 \pm \sqrt{1-\tilde{\delta}_S^2/2}\right)/2m^2$. Note $a_+ = U'(0)$ holds. If in addition (13), that is $\inf_{x\in\mathbb{R}} u_0'(x) > a_-$ holds and $\|\Phi_0\|_1$ is further small if necessary, we have (31).

Proof of Theorem 1 The local existence theorem and the uniform C^1 bounds (22) and (31) assure the global solution $(u,q)\in C_b^1\times C_b^2$. These C^1 bounds together with $\lim_{t\to\infty}|\phi|_2(t) = 0$ deduced from (28) conclude (21).

3 Critical Case (Proof of Theorem 2)

Suppose the conclusion is false, and we have a global solution of C^1 class. It is apparent that $u(t,\cdot)$ is an odd function for an arbitrary t for odd u_0. Then a characteristic curve initially within $x\ge 0$ remains always $x\ge 0$. By (32) and (34), we have

$$\frac{d}{dt}u_x(t,0) = -(u_x(t,0) + 1/2m^2)^2 + m^{-2}K_m' * \phi(t,0). \tag{36}$$

Lemma 4.

$$\phi(t,x) \le 0 \ \textit{for arbitrary} \ t \ge 0 \ \textit{and} \ x \ge 0. \tag{37}$$

Proof. Consider an arbitrary characteristic curve $\{(t, X(t)) \mid t \ge 0\}$ departing from $(0, X_0)$ with $X_0 \ge 0$. Substitute $\psi_x = m^{-2}(\phi - K_m * \phi)$ into (23) and integrate the result along that characteristic curve to have

$$\begin{aligned}\phi(t,X(t)) &= \phi_0(X_0)\exp\left(-\int_0^t (m^{-2} + U'(X(s)))ds\right) \\ &\quad + m^{-2}\int_0^t (K_m * \phi)(\tau, X(\tau))\exp\left(-\int_\tau^t (m^{-2} + U'(X(s)))ds\right)d\tau. \end{aligned}\tag{38}$$

Since $\phi(t,\cdot)$ is odd while $K_m(\cdot)$ is even and $K_m(a) \ge K_m(b)$ if $|a| \le |b|$, it holds for $x \ge 0$ that

$$\begin{aligned} K_m * \phi(\tau,x) &= \int_0^\infty \phi(\tau,y)(K_m(x-y) - K_m(x+y))dy \\ &\le \sup_{y\ge 0}\phi(\tau,y)\int_0^\infty (K_m(x-y) - K_m(x+y))dy = \sup_{y\ge 0}\phi(\tau,y)\,(1 - e^{-x/m}). \end{aligned}$$

Noting $\phi_0(X_0) \le 0$, $U'(\cdot) \in [-1/2m^2, 0)$ and $\sup_{y\ge 0}\phi(\tau,y) \ge \phi(\tau,0) = 0$, we have from (38) that $\phi(t,X(t)) \le C\int_0^t \sup_{y\ge 0}\phi(\tau,y)d\tau$ for a certain positive constant C. Taking the supremum among all characteristic curves departing from the right half line, we readily obtain (37) by the Gronwall inequality.

Proof of Theorem 2 Since both K_m' and $\phi(t,\cdot)$ are odd functions,

$$K_m' * \phi\,(t,0) = \int_{-\infty}^\infty K_m'(-x)\phi(t,x)dx = m^{-1}\int_0^\infty e^{-x/m}\phi(t,x)dx. \tag{39}$$

The conditions $\phi_0 \in C_b^1(\mathbb{R})$ and $\phi_0 \not\equiv 0$ imply $\phi(t,\cdot) \in C_b^1(\mathbb{R})$ and $\phi(t,\cdot) \not\equiv 0$. Thus, (37) and (39) imply

$$K_m' * \phi\,(t,0) < 0 \ \text{ for an arbitrary } \ t \ge 0. \tag{40}$$

Substituting (40) into (36) and noting $u_x(0,0) \le U'(0) = -1/2m^2$, we conclude

$$u_x(t,0) \to -\infty \ \text{ as } \ t \to T_* - 0$$

for a certain finite value T_*, which contradicts the initial assumption.

4 Supercritical Case (Proof of Theorem 5)

As already mentioned, if $u_0(x)$ is odd, then discontinuity of $u(t,\cdot)$ always remain at $x=0$. In that case, (ϕ,ψ) is governed by (23) and (24). The local solution to them is constructed in the class of (18) by Kato's method [6] provided the initial perturbation is sufficiently small so that the entropy condition $u_0(0-0) > u_0(0+0)$ holds. Defining $\Phi(t,x) := \int_{\pm\infty}^{x} \phi(t,y)\,dy$ for $\pm x > 0$, analogous results to Lemma 1 follow. In this section, we give a priori estimates for the local solution. Set

$$N(T) := \sup_{t\in[0,T]} \|\Phi(t,\cdot)\|_3, \quad L_0 := |(U')^{-1}(-1/3m^2)|, \quad \Omega_0 := \{x\in\mathbb{R}_0 \,|\, |x| > L_0\}.$$

4.1 Energy Estimates Away from the Discontinuity

Lemma 5. *If $N(T)$ is sufficiently small, it holds for an arbitrary $t\in[0,T]$ that*

$$|(\Phi,\phi)|_2^2(t) + \int_{\Omega_0} \left(\phi_x^2+\phi_{xx}^2\right)(t,x)dx + \int_0^t \left(|(\phi,\psi,\psi_x)|_2^2(s) + \int_{\Omega_0} (\phi_x^2+\phi_{xx}^2)(s,x)dx \right) ds \le C\|\Phi_0\|_3^2, \quad (41)$$

where C is a positive constant independent of T.

Proof. First we note (28) is still valid if $N(T)$ is sufficiently small. Differentiate (23) in x and multiply the result by ϕ_x and use (24) to obtain

$$\partial_t\left(\frac{1}{2}\phi_x^2\right) + \partial_x\left(\frac{1}{2}(\phi+U)\phi_x^2\right) + \left(m^{-2}+\frac{3}{2}U' + \frac{1}{2}\phi_x\right)\phi_x^2 + m^{-2}\phi_x\psi + U''\phi\phi_x = 0. \quad (42)$$

Now we integrate (42) over Ω_0. Since $U''\in L^\infty$, the integrals of the last two terms are estimated as

$$\left| \int_{\Omega_0} \left(m^{-2}\phi_x\psi + U''\phi\phi_x\right) dx \right| \le \varepsilon m^{-2}\int_{\Omega_0}\phi_x^2 dx + C\varepsilon^{-1}|(\phi,\psi)|_2^2, \quad (43)$$

where ε is an arbitrary positive constant. Noting $m^{-2}+3U'(x)/2 > 1/2m^2$ for $x\in\Omega_0$, if $N(T)\ll 1$ so that $|\phi|_\infty \ll |U(\pm 0)|$ $(\le |U(\pm L_0)|)$ and $|\phi_x|_\infty \ll 1$ hold, letting ε in (43) suitably small and integrating in time yield

$$\int_{\Omega_0}\phi_x(t,x)^2dx + \int_0^t\int_{\Omega_0}\phi_x(s,x)^2dxds \le C\int_{\Omega_0}\phi_0'(x)^2dx + C\int_0^t |(\phi,\psi)|_2^2(s)ds. \quad (44)$$

In the similar way, the second-order derivative is estimated as

$$\int_{\Omega_0} \phi_{xx}(t,x)^2 dx + \int_0^t \int_{\Omega_0} \phi_{xx}(s,x)^2 dx ds$$
$$\leq C \int_{\Omega_0} \phi_0''(x)^2 dx + C \int_0^t \left(|(\phi, \psi_x)|_2^2(s) + \int_{\Omega_0} \phi_x(s,x)^2 dx \right) ds. \quad (45)$$

Combination of (28), (44), and (45) yields the desired estimate.

4.2 Energy Estimates over the Entire Domain

In this subsection, we assume that $\|\Phi_0\|_3$ is sufficiently small so that the local solution has a life span T long enough for characteristic curves starting from $x = \pm L_0$ at $t = 0$ to reach $x = \pm 0$ by $t = T/2$.

For an arbitrary $s \geq 0$, we solve an ordinary differential equation

$$dX(t)/dt = U(X(t)) + \phi(t, X(t)) \text{ for } t > s \text{ with } X(s) = -L_0, \quad (46)$$

and define $T_1(y; s)$ for an arbitrary $y \in [-L_0, 0)$ so that $X(T_1(y; s)) = y$ holds. The solvability of (46) is assured by the boundedness of U' and ϕ_x. The smallness of $\|\Phi_0\|_3$ implies the existence of a limit $\lim_{y \to -0} T_1(y; 0)$, which we denote by T_0.

For an arbitrary $t > 0$, consider a characteristic curve subject to (46) arriving at $x = 0$ from left at time t and denote its location at an arbitrary time $\tau \in [0, t)$ by $a_-(\tau; t)$. For an arbitrary $t (\geq T_0)$, we define $s_0(t)$ (≥ 0) such that $a_-(s_0(t); t) = -L_0$ holds, and we set $s_0(t) = 0$ for $t \in [0, T_0)$. Letting T be the existence time of the solution, for an arbitrary $s \leq s_0(T)$ define $t_1(s)$ by $\lim_{y \to -0} T_1(y; s)$.

Lemma 6. *If $N(T)$ is sufficiently small, there exists a positive constant T_c independent of T such that $t - s_0(t) < T_c$ for an arbitrary $t \in (0, T]$, and $t_1(s) - s < T_c$ for an arbitrary $s \in [0, s_0(T)]$ hold. Moreover, the function $t_1(s)$ is differentiable almost everywhere in $s \in [0, s_0(T)/2]$ and its derivative is bounded by a constant which is independent of T.*

Proof. First two statements are obvious provided $N(T)$ is small enough. By definition, it holds for $y \in [-L_0, 0)$ that $T_1(y; s) = s + \int_{-L_0}^{y} (U(z) + \phi(T_1(z; s), z))^{-1} dz$. Taking the difference quotient of this equality with respect to s, standard arguments lead us to the conclusions.

Lemma 7. *If $N(T)$ is sufficiently small, it holds that*

$$\|\Phi\|_3(t) \leq C \|\Phi_0\|_3 \text{ for an arbitrary } t \in [0, T],$$

where C is a positive constant independent of T.

Proof. For arbitrary $t \in (0, T]$ and $\tau \in [0, t)$, define a time-dependent domain $\Omega(\tau; t)$ by $\Omega(\tau; t) := \{x < a_-(\tau; t) | x \in \mathbb{R}\}$. Applying Reynolds' transport theorem to (42), we have

$$\frac{d}{d\tau}\int_{\Omega(\tau;t)} \frac{1}{2}\phi_x^2 dx = \frac{1}{2}\phi_x^2(\tau, a_-(\tau;t))\partial_\tau a_-(\tau;t) - \int_{\Omega(\tau;t)} \partial_x\left(\frac{1}{2}(\phi+U)\phi_x^2\right)dx$$
$$-\int_{\Omega(\tau;t)} \left(m^{-2}\phi_x\psi + U''\phi\phi_x\right)dx - \int_{\Omega(\tau;t)} \left(m^{-2} + \frac{3}{2}U' + \frac{1}{2}\phi_x\right)\phi_x^2 dx. \quad (47)$$

Since $\partial_\tau a_-(\tau; t) = U(a_-(\tau; t)) + \phi(\tau, a_-(\tau; t))$, the first two terms in the right-hand side cancel. Integrating (47) in τ over $[s_0(t), t]$ and using Lemma 6, we have

$$\int_{-\infty}^0 \phi_x^2(t, x)dx \le e^{C(t-s_0(t))}\int_{\Omega(s_0(t),t)} \phi_x(s_0(t), x)^2 dx + C\int_{s_0(t)}^t e^{C(t-\tau)}|(\phi, \psi)|_2(\tau)^2 d\tau$$
$$\le C\int_{\Omega(s_0(t),t)} \phi_x(s_0(t), x)^2 dx + C\int_{s_0(t)}^t |(\phi, \psi)|_2(\tau)^2 d\tau \quad (48)$$

for an arbitrary $t \in [0, T]$, where C is independent of T. In the same way, we have

$$\int_{-\infty}^0 (\phi_x^2 + \phi_{xx}^2)(t, x)dx \le C\int_{\Omega(s_0(t),t)} \left(\phi_x^2 + \phi_{xx}^2\right)(s_0(t), x)dx + C\int_{s_0(t)}^t |(\phi, \psi, \psi_x)|_2(s)^2 ds. \quad (49)$$

By Lemma 5, the last term in the right-hand side of (49) is bounded above by $C\|\Phi_0\|_3^2$. In the case of $t \le T_0$, then $s_0(t) = 0$ and $\Omega_0 \subset \Omega(s_0(t), t) \subset (-\infty, 0)$. Therefore, the first term in the right-hand side of (49) is not greater than $C\|\Phi_0\|_3^2$. In the case of $t > T_0$, this term appears in the left-hand side of (41) because $a_-(s_0(t), t) = -L_0$ for $t > T_0$ and hence $\Omega(s_0(t), t) = \Omega_0$. In any case, $\int_{-\infty}^0 (\phi_x^2(t, x) + \phi_{xx}^2(t, x))dx \le C\|\Phi_0\|_3^2$ holds for an arbitrary $t \in [0, T]$, where C is independent of T. By this estimate and Lemma 5, we have the desired estimate.

Lemma 8. *If $N(T)$ is sufficiently small, it holds for an arbitrary $t \in [0, T]$ that $\|\Phi\|_3^2(t) + \int_0^t \|(\phi, \psi)\|_2^2(s)ds \le C\|\Phi_0\|_3^2$, where C is independent of T.*

Proof. Choose an arbitrary interval $[0, t] \subset [0, T]$ and integrate (49) with t replaced by τ over $[0, t]$ to get

$$\int_0^t \int_{-\infty}^0 (\phi_x^2(\tau, x) + \phi_{xx}^2(\tau, x))dxd\tau$$
$$\le C\int_0^t \int_{\Omega(s_0(\tau),\tau)} \left(\phi_x^2 + \phi_{xx}^2\right)(s_0(\tau), x)dxd\tau + C\int_0^t \int_{s_0(\tau)}^\tau |(\phi, \psi, \psi_x)|_2(s)^2 dsd\tau. \quad (50)$$

The first term in the right-hand side of (50) is estimated as

$$\begin{aligned}
&\int_0^t \int_{\Omega(s_0(\tau),\tau)} \left(\phi_x^2 + \phi_{xx}^2\right)(s_0(\tau), x) dx d\tau \\
=& \int_0^{T_0} \int_{\Omega(s_0(\tau),\tau)} \left(\phi_x^2 + \phi_{xx}^2\right)(s_0(\tau), x) dx d\tau + \int_{T_0}^t \int_{\Omega(s_0(\tau),\tau)} \left(\phi_x^2 + \phi_{xx}^2\right)(s_0(\tau), x) dx d\tau \\
\leq & T_0 \int_{-\infty}^0 \left((\phi_0')^2 + (\phi_0'')^2\right)(x) dx + \int_0^{s_0(t)} \int_{\Omega_0} \left(\phi_x^2 + \phi_{xx}^2\right)(s, x) dx \frac{dt_1(s)}{ds} ds \leq C\|\Phi_0\|_3^2, \qquad (51)
\end{aligned}$$

where we used $s_0(\tau) = 0$ for $\tau \leq T_0$, $\Omega(s_0(\tau), \tau) = \Omega_0$ for $\tau > T_0$, $s_0^{-1}(s) = t_1(s)$ for $s > 0$, and Lemma 5 and 6. The second term in the right-hand side of (50) is estimated by using Lemma 5 and 6 as

$$\begin{aligned}
&\int_0^t \int_{s_0(\tau)}^{\tau} |(\phi, \psi, \psi_x)|_2(s)^2 ds d\tau \\
=& \int_0^{T_0} \int_0^{\tau} |(\phi, \psi, \psi_x)|_2(s)^2 ds d\tau + \int_{T_0}^t \int_{s_0(\tau)}^{\tau} |(\phi, \psi, \psi_x)|_2(s)^2 ds d\tau \\
\leq & T_0 \int_0^{T_0} |(\phi, \psi, \psi_x)|_2(s)^2 ds + \sup_{s \leq s_0(t)} (t_1(s) - s) \int_{T_0}^t |(\phi, \psi, \psi_x)|_2(s)^2 ds \\
\leq & (T_0 + T_C) \int_0^t |(\phi, \psi, \psi_x)|_2(s)^2 ds \leq C\|\Phi_0\|_3^2. \qquad (52)
\end{aligned}$$

Substituting (51) and (52) into (50), we obtain

$$\int_0^t \int_{-\infty}^0 (\phi_x^2(\tau, x) + \phi_{xx}^2(\tau, x)) dx d\tau \leq C\|\Phi_0\|_3^2 \text{ for an arbitrary } t \in [0, T], \quad (53)$$

where C is independent of T. Lemma 5, 7, (53) and (24) complete the proof.

Proof of Theorem 5 Local existence theorem and Lemma 7 conclude the unique existence of the global solution. Standard arguments deduce (21) from Lemma 8.

Acknowledgements The author is supported by the Grants-in-Aid for Scientific Research (KAKENHI, Grant No.26800076, 15KT0014, 17K05313, 17K05376) from the Japan Society for the Promotion of Science.

References

1. A. Bressan, *Hyperbolic Systems of Conservation Laws: The One Dimensional Cauchy Problem* (Oxford University Press, Oxford, 2000)
2. J.-F. Coulombel, T. Goudon, P. Lafitte, C. Lin, Analysis of large amplitude shock profiles for non-equilibrium radiative hydrodynamics: formation of Zeldovich spikes. Shock Waves **22**, 181–197 (2012)
3. R. Duan, K. Fellner, C. Zhu, Energy method for multi-dimensional balance laws with non-local dissipation. J. Math. Pures Appl. **93**, 572–598 (2010)

4. K. Hamer, Nonlinear effects on the propagation of sounds waves in a radiating gas. Quarter J. Mech. Appl. Math. **24**, 155–168 (1971)
5. M.A. Heaslet, B.S. Baldwin, Predictions of the structure of radiation resisted shock waves. Phys. Fluids **6**, 781–791 (1963)
6. T. Kato, Linear evolution equations of "hyperbolic" type. II. J. Math. Soc. Jpn. **25**, 648–666 (1973)
7. S. Kawashima, S. Nishibata, Shock waves for a model system of the radiating gas. SIAM J. Math. Anal. **30**, 95–117 (1998)
8. S. Kawashima, S. Nishibata, Cauchy problem for a model system of the radiating gas: weak solutions with a jump and classical solutions. Math. Models. Meth. Sci. **9**, 69–91 (1999)
9. S.N. Kružkov, First order quasilinear equations in several independent variables. Math. USSR Sbor. **10**, 217–243 (1970)
10. C. Lattanzio, C. Mascia, T. Nguyen, R.G. Plaza, K. Zumbrun, Stability of scalar radiative shock profiles. SIAM J. Math. Anal. **41**, 2165–2206 (2009)
11. R.B. Lowrie, J.D. Edwards, Radiative shock solutions with grey nonequilibrium diffusion. Shock Waves **18**, 129–143 (2008)
12. R.B. Lowrie, R.M. Rauenzahn, Radiative shock solutions in the equilibrium diffusion limit. Shock Waves **16**, 445–453 (2007)
13. C. Mascia, Small, medium and large shock waves for radiative Euler equations. Physica D **245**, 46–56 (2013)
14. T. Nguyen, R.G. Plaza, K. Zumbrun, Stability of radiative shock profiles for hyperbolic-elliptic coupled systems. Physica D **239**, 428–453 (2010)
15. M. Ohnawa, L^∞-stability of continuous shock waves in a radiating gas model. SIAM J. Math. Anal. **46**, 2136–2159 (2014)
16. M. Ohnawa, L^∞-stability of discontinuous traveling waves in a hyperbolic-elliptic coupled system. SIAM J. Math. Anal. **48**, 3820–3839 (2016)
17. D. Serre, L^1-stability of constants in a model for radiating gases. Commun. Math. Sci. **1**, 197–205 (2003)
18. D. Serre, L^1-stability of nonlinear waves in scalar conservation laws, in *Evolutionary Equations Handbook of Differential Equations*, vol. 1, ed. by C.M. Dafermos, E. Feireisl (Elsevier, North Holland, Amsterdam, 2004), pp. 473–553
19. YaB Zel'dovich, Shock waves of large amplitude in air. Soviet Phys. JETP **5**, 919–927 (1957)

On the Longtime Behavior of Almost Periodic Entropy Solutions to Scalar Conservation Laws

Evgeny Yu. Panov

Abstract We found the precise condition for the decay as $t \to \infty$ of Besicovitch almost periodic entropy solutions of multidimensional scalar conservation laws. Moreover, in the case of one space variable we establish asymptotic convergence of the entropy solution to a traveling wave (in the Besicovitch norm). Besides, the flux function turns out to be affine on the minimal segment containing the essential range of the limit profile while the speed of the traveling wave coincides with the slope of the flux function on this segment.

Keywords Almost periodic entropy solutions · Decay property
Scalar conservation laws · Spectrum · Traveling waves

1 Introduction

In the half-space $\Pi = \mathbb{R}_+ \times \mathbb{R}^n$, where $\mathbb{R}_+ = (0, +\infty)$, we consider a conservation law

$$u_t + \operatorname{div}_x \varphi(u) = 0, \quad u = u(t, x), \ (t, x) \in \Pi. \tag{1}$$

The flux vector $\varphi(u) = (\varphi_1(u), \dots, \varphi_n(u))$ is supposed to be merely continuous: $\varphi(u) \in C(\mathbb{R}, \mathbb{R}^n)$. Recall the notion of Kruzhkov entropy solution of the Cauchy problem for Eq. (1) with initial condition

$$u(0, x) = u_0(x) \in L^\infty(\mathbb{R}^n). \tag{2}$$

Definition 1 ([6])**.** A bounded measurable function $u = u(t, x) \in L^\infty(\Pi)$ is called an entropy solution (e.s.) of (1), (2) if for all $k \in \mathbb{R}$

E. Yu. Panov (✉)
Novgorod State University, 41, B. St-Peterburgskaya str., 173003 Veliky Novgorod, Russia
e-mail: Eugeny.Panov@novsu.ru

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_30

$$\frac{\partial}{\partial t}|u-k|+\operatorname{div}_x[\operatorname{sign}(u-k)(\varphi(u)-\varphi(k))]\le 0 \tag{3}$$

in the sense of distributions on Π (in $\mathscr{D}'(\Pi)$), and

$$\operatorname*{ess\,lim}_{t\to 0+} u(t,\cdot)=u_0 \quad \text{in } L^1_{loc}(\mathbb{R}^n). \tag{4}$$

Here $\operatorname{sign} u=\begin{cases}1, & u>0,\\ -1, & u\le 0\end{cases}$ and relation (3) means that for each test function $h=h(t,x)\in C_0^1(\Pi)$, $h\ge 0$,

$$\int_\Pi [|u-k|h_t+\operatorname{sign}(u-k)(\varphi(u)-\varphi(k))\cdot\nabla_x h]dtdx\ge 0,$$

where $\cdot$ denotes the inner product in $\mathbb{R}^n$.

Taking in (3) $k=\pm R$, where $R\ge\|u\|_\infty$, we obtain that $u_t+\operatorname{div}_x\varphi(u)=0$ in $\mathscr{D}'(\Pi)$; that is, an e.s. $u=u(t,x)$ is a weak solution of this equation as well.

The existence of e.s. of (1), (2) follows from the general result of [12, Theorem 3]. In the case under consideration when the flux vector is only continuous the effect of infinite speed of propagation appears, which may even lead to the nonuniqueness of e.s. if $n>1$, and see examples in [7, 8, 12], where exact sufficient conditions of the uniqueness were also found. Nevertheless, if an initial function u_0 is periodic in $\mathbb{R}^n$ (at least in $n-1$ independent directions), then the e.s. of (1), (2) is unique and x-periodic; see [11], as well as the more general result [12, Theorem 11].

We will study problem (1), (2) in the class of Besicovitch almost periodic functions. Let C_R be the cube

$$\{\,x=(x_1,\dots,x_n)\in\mathbb{R}^n \mid |x|_\infty=\max_{i=1,\dots,n}|x_i|\le R/2\,\},\quad R>0.$$

We define the seminorm

$$N_1(u)=\limsup_{R\to+\infty} R^{-n}\int_{C_R}|u(x)|dx,\quad u(x)\in L^1_{loc}(\mathbb{R}^n).$$

Recall (see [1, 9]) that the Besicovitch space $\mathscr{B}^1(\mathbb{R}^n)$ is the closure of trigonometric polynomials, i.e., finite sums $\sum a_\lambda e^{2\pi i\lambda\cdot x}$ with $i^2=-1$, $\lambda\in\mathbb{R}^n$, in the quotient space $B^1(\mathbb{R}^n)/B_0^1(\mathbb{R}^n)$, where

$$B^1(\mathbb{R}^n)=\{u\in L^1_{loc}(\mathbb{R}^n)\mid N_1(u)<+\infty\},\ B_0^1(\mathbb{R}^n)=\{u\in L^1_{loc}(\mathbb{R}^n)\mid N_1(u)=0\}.$$

The space $\mathscr{B}^1(\mathbb{R}^n)$ is equipped with the norm $\|u\|_1=N_1(u)$ (we identify classes in the quotient space $B^1(\mathbb{R}^n)/B_0^1(\mathbb{R}^n)$ and their representatives). The space $\mathscr{B}^1(\mathbb{R}^n)$ is a Banach space, and it is isomorphic to the completeness of the space $AP(\mathbb{R}^n)$ of

Bohr almost periodic functions with respect to the norm N_1. It is known (see, for instance, [1]) that for each function $u \in \mathscr{B}^1(\mathbb{R}^n)$ there exists the mean value

$$\bar{u} = \fint_{\mathbb{R}^n} u(x)dx \doteq \lim_{R\to+\infty} R^{-n} \int_{C_R} u(x)dx$$

and, more generally, the Bohr–Fourier coefficients

$$a_\lambda = \fint_{\mathbb{R}^n} u(x)e^{-2\pi i\lambda\cdot x}dx, \quad \lambda \in \mathbb{R}^n.$$

The set

$$Sp(u) = \{\, \lambda \in \mathbb{R}^n \mid a_\lambda \neq 0 \,\}$$

is called the spectrum of an almost periodic function $u(x)$. It is known [1] that the spectrum $Sp(u)$ is at most countable.

Now we assume that the initial function $u_0(x) \in \mathscr{B}^1(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$. Let $I = \fint_{\mathbb{R}^n} u_0(x)dx$, and M_0 be the smallest additive subgroup of $\mathbb{R}^n$ containing $Sp(u_0)$.

It was shown in [17] that an e.s. $u(t,x)$ of (1), (2) is almost periodic with respect to spatial variables. Moreover, $u(t,x) \in C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$ (after possible correction on a set of null measure) and $Sp(u(t,\cdot)) \subset M_0$, $\fint_{\mathbb{R}^n} u(t,x)dx = I$ for all $t \geq 0$. The uniqueness of e.s. $u(t,x)$ in the space $C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$ is a consequence of the following general result [17, Proposition 1.3], which holds for arbitrary bounded and measurable initial functions.

Theorem 1. *Let $u(t,x), v(t,x) \in L^\infty(\Pi)$ be e.s. of (1), (2) with initial functions $u_0(x), v_0(x) \in L^\infty(\mathbb{R}^n)$, respectively. Then for a.e. $t > 0$*

$$N_1(u(t,\cdot) - v(t,\cdot)) \leq N_1(u_0 - v_0). \tag{5}$$

For completeness, we reproduce the proof.

Proof. Applying Kruzhkov doubling of variables method, we obtain the relation (see [6, 12])

$$|u-v|_t + \operatorname{div}_x[\operatorname{sign}(u-v)(\varphi(u)-\varphi(v))] \leq 0 \ \text{ in } \mathscr{D}'(\Pi). \tag{6}$$

We choose a function $g(y) \in C_0^1(\mathbb{R}^n)$ such that $0 \leq g(y) \leq 1$, and $g(y) \equiv 1$ in the cube C_1, $g(y) \equiv 0$ in the complement of the cube C_k, $k > 1$, and a function $h = h(t) \in C_0^1(\mathbb{R}_+)$, $h \geq 0$. Applying (6) to the test function $f = R^{-n}h(t)g(x/R)$ with $R > 0$, we obtain

$$\int_0^\infty \left(R^{-n} \int_{\mathbb{R}^n} |u(t,x) - v(t,x)| g(x/R)dx \right) h'(t)dt +$$
$$R^{-n-1} \int_\Pi \operatorname{sign}(u-v)(\varphi(u)-\varphi(v)) \cdot \nabla_y g(x/R)h(t)dtdx \geq 0. \tag{7}$$

Making the change $y = x/R$ in the last integral in (7), we derive the estimate

$$R^{-n-1}\left|\int_{\Pi} \operatorname{sign}(u-v)(\varphi(u)-\varphi(v))\cdot\nabla_y g(x/R)h(t)dtdx\right| \le$$
$$R^{-1}\|\varphi(u)-\varphi(v)\|_\infty \int_{\Pi}|\nabla_y g|(y)h(t)dtdy \le \frac{A}{R}\int_0^{+\infty} h(t)dt, \tag{8}$$

where $A = \|\varphi(u)-\varphi(v)\|_\infty \int_{\mathbb{R}^n}|\nabla_y g|(y)dy$. Here and below we use the notation $|z|$ for the Euclidean norm of a finite-dimensional vector z. Let

$$I_R(t) = R^{-n}\int_{\mathbb{R}^n}|u(t,x)-v(t,x)|g(x/R)dx.$$

From (7) and (8), it follows that

$$\int_0^{+\infty}(I_R(t)-At/R)h'(t)dt = \int_0^{+\infty} I_R(t)h'(t)dt + \frac{A}{R}\int_0^{+\infty} h(t)dt \ge 0$$

for all $h(t)\in C_0^1((0,+\infty))$, $h(t)\ge 0$. This means that the generalized derivative $\frac{d}{dt}(I_R(t)-At/R)\le 0$, which readily implies that there exists a set $F\subset(0,+\infty)$ of full Lebesgue measure (which can be defined as the set of common Lebesgue points of functions $I_R(t)$, $R\in\mathbb{Q}$) such that $\forall t_2, t_1\in F, t_2>t_1, \forall R\in\mathbb{Q}$ $I_R(t_2)-At_2/R \le I_R(t_1)-At_1/R$, that is $I_R(t_2)\le I_R(t_1)+A(t_2-t_1)/R$. By the evident continuity of $I_R(t)$ with respect to R, the latter relation remains valid for all $R>0$. In the limit as $F\ni t_1\to 0$ we obtain, taking into account the initial conditions for e.s. u, v, that $\forall t_2 = t\in F$ for all $R>0$

$$I_R(t)\le I_R(0)+At/R, \tag{9}$$

where $I_R(0) = R^{-n}\int_{\mathbb{R}^n}|u_0(x)-v_0(x)|g(x/R)dx$. By the properties of $g(y)$, we find the inequalities

$$R^{-n}\int_{C_R}|u(t,x)-v(t,x)|dx \le I_R(t) \le$$
$$R^{-n}\int_{C_{kR}}|u(t,x)-v(t,x)|dx = k^n(kR)^{-n}\int_{C_{kR}}|u(t,x)-v(t,x)|dx,$$

which imply that

$$N_1(u(t,\cdot)-v(t,\cdot)) \le \limsup_{R\to+\infty} I_R(t) \le k^n N_1(u(t,\cdot)-v(t,\cdot)). \tag{10}$$

In view of (10), we derive from (9) in the limit as $R\to+\infty$ that $N_1(u(t,\cdot)-v(t,\cdot))\le k^n N_1(u_0-v_0)$ for all $t\in F$. To complete the proof, it only remains to notice that $k>1$ is arbitrary.

Remark 1. As was established in [13, Corollary 7.1], after possible correction on a set of null measure any e.s. $u(t,x) \in C(\mathbb{R}_+, L^1_{loc}(\mathbb{R}^n))$. In particular, without loss of generality, we may claim that relation (9) holds for all $t > 0$. This implies in the limit as $R \to +\infty$ that the statement of Theorem 1 holds for all $t > 0$ as well. The continuity property allows also to replace the essential limit in initial condition (4) by the usual one.

Our main results are contained in Theorems 2, 4, indicated below.

Theorem 2. *Assume that the following non-degeneracy condition holds for the flux components in "resonant" directions $\xi \in M_0$:*

$$\forall \xi \in M_0, \xi \neq 0 \text{ the functions } u \to \xi \cdot \varphi(u) \text{ are not affine in any vicinity of } I = \overline{u_0}. \tag{11}$$

Then, an e.s. $u(t,x) \in C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$ satisfies the decay property

$$\lim_{t\to+\infty} u(t,\cdot) = I \ \text{ in } \mathscr{B}^1(\mathbb{R}^n). \tag{12}$$

Condition (11) is precise: if it fails, then there exists an initial data $u_0 \in \mathscr{B}^1(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ with the properties $Sp(u_0) \subset M_0$, $\overline{u_0} = I$, such that the corresponding e.s. $u(t,x)$ of (1), (2) does not satisfy (12).

Remark 2. The decay of almost periodic e.s. was firstly studied by H. Frid [5] in the class of Stepanov almost periodic function. This class is natural for the case of smooth flux vector $\varphi(u)$, when an e.s. $u(t,x)$ of (1), (2) exhibits the property of finite speed of propagation. The decay of such solutions was established in the stronger Stepanov norm but under rather restrictive assumptions on the dependence of the length of inclusion intervals for ε-almost periods of u_0 on the parameter ε.

Notice that in the case of a periodic function u_0 the group M_0 coincides with the dual lattice $\mathscr{L}'$ to the lattice $\mathscr{L}$ of periods of u_0, and in this case, Theorem 2 reduces to the following result [15] (see also the earlier paper [14]):

Theorem 3. *Under the condition*

$$\forall \xi \in \mathscr{L}', \xi \neq 0 \text{ the functions } u \to \xi \cdot \varphi(u) \text{ are not affine in any vicinity of } I = \int_{\mathbb{T}^n} u_0(x)dx \tag{13}$$

an e.s. $u(t,x) \in C([0,+\infty), L^1(\mathbb{T}^n))$ satisfies the decay property

$$\lim_{t\to+\infty} \int_{\mathbb{T}^n} |u(t,x) - I| dx = 0. \tag{14}$$

Here $\mathbb{T}^n = \mathbb{R}^n/\mathscr{L}$ is the n-dimensional torus, and dx is the normalized Lebesgue measure on $\mathbb{T}^n$.

Remark that in the case $\varphi(u) \in C^2(\mathbb{R}, \mathbb{R}^n)$ the assertion of Theorem 3 was established in [3]. Now we consider the case of one space variable $n = 1$ when (1) has the form

$$u_t + \varphi(u)_x = 0, \tag{15}$$

where $\varphi(u) \in C(\mathbb{R})$. As above, we assume that $u_0 \in \mathscr{B}^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and that M_0 is the additive subgroup of $\mathbb{R}$ generated by $Sp(u_0)$. For an almost periodic function $v(x) \in \mathscr{B}^1(\mathbb{R})$, we denote by $S(v)$ the minimal segment $[a, b]$ containing essential values of $v(x)$. This segment can be defined by the relations

$$b = \min\{\, k \in \mathbb{R} \mid (v-k)^+ = \max(v-k, 0) = 0 \text{ in } \mathscr{B}^1(\mathbb{R}) \,\},$$
$$a = \max\{\, k \in \mathbb{R} \mid (k-v)^+ = 0 \text{ in } \mathscr{B}^1(\mathbb{R}) \,\}.$$

As is easy to verify, the above minimal and maximal values exist and $a \le b$.

Our second result is the following unconditional asymptotic property of convergence of an e.s. $u(t, x)$ to a traveling wave:

Theorem 4. *There is a constant $c \in \mathbb{R}$ (speed) and a function $v(y) \in \mathscr{B}^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ (profile) such that*

$$\lim_{t \to +\infty} (u(t, x) - v(x - ct)) = 0 \ \text{ in } \mathscr{B}^1(\mathbb{R}). \tag{16}$$

Moreover, $Sp(v) \subset M_0$, $\bar{v} = I = \overline{u_0}$, and $\varphi(u) - cu = \text{const}$ on the segment $S(v)$.

We remark, in addition to Theorem 4, that the profile $v(y)$ of the traveling wave and, if $v \not\equiv \text{const}$, its speed c are uniquely defined. Indeed, if (16) holds with $v = v_1, v_2, c = c_1, c_2$, respectively, then $v_1(x - c_1 t) - v_2(x - c_2 t) \to 0$ in $\mathscr{B}^1(\mathbb{R})$ as $t \to +\infty$, which implies the relation

$$\lim_{t \to +\infty} (v_1(y) - v_2(y + (c_1 - c_2)t)) = 0 \ \text{ in } \mathscr{B}^1(\mathbb{R}). \tag{17}$$

By the known property of almost periodic functions (see, for example, [1]), there exists a sequence $t_r \to +\infty$ such that $v_2(y + (c_1 - c_2)t_r) \underset{r \to \infty}{\to} v_2(y)$ in $\mathscr{B}^1(\mathbb{R})$ (this is evident if $c_1 = c_2$). On the other hand, in view of (17) $v_2(y + (c_1 - c_2)t_r) \underset{r \to \infty}{\to} v_1(y)$ in $\mathscr{B}^1(\mathbb{R})$ and hence $v_1 = v_2$ in $\mathscr{B}^1(\mathbb{R})$. Further, if $\Delta c = c_1 - c_2 \ne 0$, then it follows from (17) in the limit as $t = t_r + h/\Delta c \to +\infty$ that $v_2(y) = v_2(y + h)$ in $\mathscr{B}^1(\mathbb{R})$ for each $h \in \mathbb{R}$. Therefore,

$$v_2(y) = \fint_{\mathbb{R}} v_2(y + h)dh = \fint_{\mathbb{R}} v_2(h)dh = \overline{v_2} = \text{const}.$$

Thus, for the nonconstant profile $v = v_2$ the speed $c_1 = c_2 = c$ is uniquely determined. We also remark that $\|v\|_\infty \le \|u_0\|_\infty$ because by the maximum principle $|u(t, x)| \le \|u_0\|_\infty$ a.e. in Π.

Theorem 4 defines the nonlinear operator T on $\mathscr{B}^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$, which associates an initial function u_0 with the profile $v(y) = T(u_0)(y)$ of the limit traveling wave for the corresponding e.s. of problem (15), (2). In Theorem 5 below, we establish that T does not increase the distance in $\mathscr{B}^1(\mathbb{R})$.

Remark 3. In the case $n = 1$, the statement of Theorem 2 follows from Theorem 4. Indeed, under the assumptions of Theorem 2, $v(y) = I$ in $\mathscr{B}^1(\mathbb{R})$. Otherwise, $a < I < b$, where $[a, b] = S(v)$ and, by Theorem 4, $\varphi(u) = cu + \text{const}$ in the vicinity (a, b) of I. But the latter contradicts to assumption (11) of Theorem 2.

Note that in the periodic case Theorems 4, 5 were proved in [16].

2 Proof of Theorem 2

We assume firstly that the initial function is a trigonometric polynomial $u_0(x) = \sum_{\lambda\in\Lambda} a_\lambda e^{2\pi i\lambda\cdot x}$. Here $\Lambda = Sp(u_0) \subset \mathbb{R}^n$ is a finite set. The minimal additive subgroup $M_0 \doteq M(u_0)$ of $\mathbb{R}^n$ containing Λ is a finite generated torsion-free abelian group, and therefore, it is a free abelian group of finite rank (see [10]). Therefore, there is a basis $\lambda_j \in M_0$, $j = 1, \dots, m$, so that every element $\lambda \in M_0$ can be uniquely represented as $\lambda = \lambda(\bar{k}) = \sum_{j=1}^m k_j\lambda_j$, $\bar{k} = (k_1, \dots, k_m) \in \mathbb{Z}^m$. In particular, the vectors λ_j, $j = 1, \dots, m$, are linearly independent over the field of rational numbers $\mathbb{Q}$. We introduce the finite set $J = \{\, \bar{k} \in \mathbb{Z}^m \mid \lambda(\bar{k}) \in \Lambda \,\}$ and represent the initial function as

$$u_0(x) = \sum_{\bar{k}\in J} a_{\bar{k}} e^{2\pi i \sum_{j=1}^m k_j\lambda_j\cdot x}, \quad a_{\bar{k}} \doteq a_{\lambda(\bar{k})}.$$

By this representation $u_0(x) = v_0(y(x))$, where

$$v_0(y) = \sum_{\bar{k}\in J} a_{\bar{k}} e^{2\pi i\bar{k}\cdot y}$$

is a periodic function on $\mathbb{R}^m$ with the standard lattice of periods $\mathbb{Z}^m$ while $y(x)$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ defined by the equalities $y_j = \lambda_j \cdot x = \sum_{i=1}^n \lambda_{ji}x_i$, λ_{ji}, $i = 1, \dots, n$, being coordinates of the vectors λ_j, $j = 1, \dots, m$. We consider the conservation law

$$v_t + \operatorname{div}_y\tilde{\varphi}(v) = 0, \quad v = v(t, y),\ t > 0,\ y \in \mathbb{R}^m, \tag{18}$$

$\tilde{\varphi}(v) = (\tilde{\varphi}_1(v), \dots, \tilde{\varphi}_m(v))$, where

$$\tilde{\varphi}_j(v) = \lambda_j \cdot \varphi(u) = \sum_{i=1}^n \lambda_{ji}\varphi_i(v) \in C(\mathbb{R}), \quad j = 1, \dots, m.$$

As was shown in [11, 12], there exists a unique e.s. $v(t,y) \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^m)$ of the Cauchy problem for Eq. (18) with initial function $v_0(y)$ and this e.s. is y-periodic, i.e., $v(t, y+e) = v(t,y)$ a.e. in $\mathbb{R}_+ \times \mathbb{R}^m$ for all $e \in \mathbb{Z}^m$. Besides, in view of [13, Corollary 7.1], we may suppose that $v(t,\cdot) \in C([0,+\infty), L^1(\mathbb{T}^m))$, where $\mathbb{T}^m = \mathbb{R}^m/\mathbb{Z}^m$ is an m-dimensional torus (which may be identified with the fundamental cube $[0,1)^m$). Formally, for $u(t,x) = v(t,y(x))$

$$u_t + \operatorname{div}_x \varphi(u) = v_t + \sum_{i=1}^{n}\sum_{j=1}^{m} (\varphi_i(v))_{y_j} \frac{\partial y_j(x)}{\partial x_i} =$$

$$v_t + \sum_{i=1}^{n}\sum_{j=1}^{m} (\varphi_i(v))_{y_j} \lambda_{ji} = v_t + \sum_{j=1}^{m} (\tilde{\varphi}_j(v))_{y_j} = 0.$$

However, these reasons are correct only for classical solutions. In the general case $v(t,y) \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^m)$, the range of $y(x)$ may be a proper subspace of $\mathbb{R}^m$ (for example, this is always true if $m > n$), and the composition $v(t,y(x))$ is not even defined. The situation is saved by introduction of additional variables $z \in \mathbb{R}^m$. Namely, the linear change $(z,x) \to (z + y(x), x)$ is not degenerated; i.e., it is a linear automorphism of $\mathbb{R}^m \times \mathbb{R}^n$. Since $v(t,y)$ is an e.s. of Eq. (18) considered in the extended half-space $t > 0$, $(y,x) \in \mathbb{R}^{m+n}$, then the function $u(t,z,x) = v(t, z + y(x))$ satisfies the relations

$$|u-k|_t + \operatorname{div}_x[\operatorname{sign}(u-k)(\varphi(u) - \varphi(k))] =$$

$$|v-k|_t + \sum_{i=1}^{n}\sum_{j=1}^{m} [\operatorname{sign}(v-k)(\varphi_i(v) - \varphi_i(k))]_{y_j} \frac{\partial y_j(x)}{\partial x_i} =$$

$$|v-k|_t + \sum_{j=1}^{m}\sum_{i=1}^{n} [\operatorname{sign}(v-k)(\varphi_i(v) - \varphi_i(k))]_{y_j} \lambda_{ji} =$$

$$|v-k|_t + \sum_{j=1}^{m} [\operatorname{sign}(v-k)(\tilde{\varphi}_j(u) - \tilde{\varphi}_j(k))]_{y_j} \le 0 \ \text{ in } \mathscr{D}'(\mathbb{R}_+ \times \mathbb{R}^{m+n}).$$

Evidently, the initial condition

$$\lim_{t \to 0+} u(t,z,x) = u_0(z,x) \doteq v_0(z + y(x)) \ \text{ in } L^1_{loc}(\mathbb{R}^{m+n})$$

is also satisfied; therefore, $u(t,z,x)$ is an e.s. of (1), (2) in the extended domain $\mathbb{R}_+ \times \mathbb{R}^{m+n}$. Since Eq. (1) does not contain the auxiliary variables $z \in \mathbb{R}^m$, then (cf. [17, Theorem 2.1]) for all $z \in E \subset \mathbb{R}^m$, where E is a set of full measure, the function $v(t, z + y(x))$ is an e.s. of (1), (2) with initial data $v_0(z + y(x)) \in \mathscr{B}^1(\mathbb{R}^n)$. Therefore, $v(t, z + y(x)) = u^z(t,x)$ a.e. in Π, where, in accordance with [17, Theorem 1.6], $u^z(t,x) \in C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$ is a unique almost periodic e.s. of (1),

(2). Therefore, we may find a countable dense set $S \subset \mathbb{R}_+$ and a subset $E_1 \subset E$ of full measure such that $u^z(t,x) = v(t, z + y(x))$ in $\mathscr{B}^1(\mathbb{R})$ for all $t \in S$, $z \in E_1$.

Further, as follows from independence of the vectors λ_j, $j = 1, \dots, m$, over $\mathbb{Q}$, the action of the additive group $\mathbb{R}^n$ on the torus $\mathbb{T}^m$ defined by the shift transformations $T_x z = z + y(x)$, $x \in \mathbb{R}^n$ is ergodic; see [17] for details. By the variant of Birkhoff individual ergodic theorem [4, Chap. VIII] for every $w(y) \in L^1(\mathbb{T}^m)$ for a.e. $z \in \mathbb{T}^m$ there exists the mean value

$$\fint_{\mathbb{R}^n} w(z + y(x))dx = \int_{\mathbb{T}^m} w(y)dy. \tag{19}$$

In view of (19), there exists a set $E_2 \subset E_1$ of full measure such that for $z \in E_2$ and all $t \in S$

$$\fint_{\mathbb{R}^n} |u^z(t,x) - I|dx = \fint_{\mathbb{R}^n} |v(t, z + y(x)) - I|dx = \int_{\mathbb{T}^m} |v(t,y) - I|dy.$$

Since $u^z(t,x) \in C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$, $v(t,\cdot) \in C([0,+\infty), L^1(\mathbb{T}^m))$, while the set S is dense in $[0,+\infty)$, we find that property

$$\fint_{\mathbb{R}^n} |u^z(t,x) - I|dx = \int_{\mathbb{T}^m} |v(t,y) - I|dy \tag{20}$$

remains valid for all $t \geq 0$. Observe that $v_0(z + y(x)) \to v_0(y(x)) = u_0(x)$ as $z \to 0$ in $\mathscr{B}^1(\mathbb{R}^n)$ (and even in $AP(\mathbb{R}^n)$). Hence, by Theorem 1 in the limit as $E_2 \ni z \to 0$ $u^z(t,x) \to u(t,x)$ in $C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$, where $u(t,x)$ is the e.s. of original problem (1), (2). Therefore, relation (20) in the limit as $z \to 0$ implies the equality

$$\fint_{\mathbb{R}^n} |u(t,x) - I|dx = \int_{\mathbb{T}^m} |v(t,y) - I|dy. \tag{21}$$

Further, for every $\bar{k} = (k_1, \dots, k_m) \in \mathbb{Z}^m$

$$\bar{k} \cdot \tilde{\varphi}(u) = \sum_{j=1}^{m} \sum_{i=1}^{n} k_j \lambda_{ji} \varphi_i(u) = \lambda(\bar{k}) \cdot \varphi(u),$$

where $\lambda(\bar{k}) = \sum_{j=1}^m k_j \lambda_j \in M_0$. By condition (11), the functions $u \to \bar{k} \cdot \tilde{\varphi}(u)$ are not affine in any vicinity of $I = \overline{u_0} = \int_{\mathbb{T}^m} v_0(y)dy$. We see that non-degeneracy requirement (13) is satisfied, and by [15, Theorem 1.3]

$$\lim_{t \to +\infty} \int_{\mathbb{T}^m} |v(t,y) - I|dy = 0.$$

Now it follows from (21) that

$$\lim_{t\to+\infty} \fint_{\mathbb{R}^n} |u(t,x) - I|dx = 0,$$

i.e., (12) holds.

In the general case $u_0 \in \mathscr{B}^1(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$, we choose a sequence $u_{0m}, m \in \mathbb{N}$, of trigonometric polynomials converging to u_0 in $\mathscr{B}^1(\mathbb{R}^n)$ and such that $Sp(u_{0m}) \subset M_0$, $\overline{u_{0m}} = I$ (for instance, we may choose the Bochner–Fejér trigonometric polynomials; see [1]). Let $u_m(t,x)$ be the corresponding sequence of e.s. of (1), (2) with initial data $u_{0m}(x)$, $m \in \mathbb{N}$. By Theorem 1 and Remark 1, this sequence converges as $m \to \infty$ to the e.s. $u(t,x)$ of the original problem in $C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$. We have already established that under condition (11) e.s. $u_m(t,x)$ satisfy the decay property

$$\lim_{t\to+\infty} u_m(t,\cdot) = I \ \text{ in } \mathscr{B}^1(\mathbb{R}^n).$$

Passing to the limit as $m \to \infty$ in this relation and taking into account the uniform convergence $u_m(t,\cdot) \underset{m\to\infty}{\to} u(t,\cdot)$ in $\mathscr{B}^1(\mathbb{R}^n)$, we obtain (12).

In conclusion, we demonstrate that condition (11) is precise. Indeed, if this condition is violated, then there is a nonzero vector $\xi \in M_0$ such that $\xi \cdot \varphi(u) = \tau u + c$ on some segment $[I-\delta, I+\delta]$, where $\tau, c, \delta \in \mathbb{R}$, and $\delta > 0$. Obviously, the function

$$u(t,x) = I + \delta \sin(2\pi(\xi \cdot x - \tau t))$$

is an e.s. of (1), (2) with the periodic initial function $u_0(x) = I + \delta \sin(2\pi(\xi \cdot x))$. We see that $\overline{u_0} = I$, $Sp(u_0) \subset \{-\xi, 0, \xi\} \subset M_0$ but the e.s. $u(t,x)$ does not converge to a constant in $\mathscr{B}^1(\mathbb{R}^n)$ as $t \to +\infty$.

The proof of Theorem 2 is complete.

3 Proof of Theorem 4

If the flux function $\varphi(u)$ is not affine in any vicinity of I, then by Theorem 2 the function $v(y) \equiv I$, and the segment $S(v) = [I, I] = \{I\}$. Otherwise, suppose that the function $\varphi(u)$ is affine in a certain maximal interval (a,b), where $-\infty \le a < I < b \le +\infty$: $\varphi(u) - cu = \text{const}$ in (a,b).

Assuming that $b < +\infty$, we define $u_+ = u_+(t,x)$ as the e.s. of (15), (2) with initial function $u_0(x) + b - I > u_0$. By the comparison principle [7, 8, 11, 12] $u_+ \ge u$ a.e. in Π. We note that $\fint_{\mathbb{R}}(u_0(x) + b - I)dx = b$ while $\varphi(u)$ is not affine in any vicinity of b (otherwise, $\varphi(u)$ is affine on a larger interval (a,b'), $b' > b$, which contradicts the maximality of (a,b)). By Theorem 2 $u_+(t,\cdot) \to b$ in $\mathscr{B}^1(\mathbb{R})$ as $t \to +\infty$, and it follows from the inequality $u \le u_+$ that $(u(t,\cdot) - b)^+ \to 0$ as $t \to +\infty$ in $\mathscr{B}^1(\mathbb{R})$. Similarly, if $a > -\infty$, then $u \ge u_-$, where $u_- = u_-(t,x)$ is an e.s. of (15), (2) with initial function $u_0(x) + a - I < u_0$. By Theorem 2 again the function $u_-(t,\cdot) \to a$ as $t \to +\infty$ in $\mathscr{B}^1(\mathbb{R})$ because $\fint_{\mathbb{R}}(u_0(x) + a - I)dx = a$ while the

function $\varphi(u)$ is not affine in any vicinity of a. Therefore, $(a - u(t,\cdot))^+ \underset{t\to+\infty}{\to} 0$ in $\mathscr{B}^1(\mathbb{R})$. The obtained limit relations can be represented in the form

$$u(t,\cdot) - s_{a,b}(u(t,\cdot)) \underset{t\to+\infty}{\to} 0 \ \text{ in } \mathscr{B}^1(\mathbb{R}), \tag{22}$$

where $s_{a,b}(u) = \min(b, \max(a,u))$ is the cut-off function at the levels a, b (it is possible that $a = -\infty$ or $b = +\infty$).

We set $w(t,x) = s_{a,b}(u(t,x))$ and choose a strictly increasing sequence $t_k > 0$ such that $t_k \to +\infty$ and $N_1(u(t_k,\cdot) - w(t_k,\cdot)) \le 2^{-k}$. Since $a \le w(t,x) \le b$ while $\varphi(u) = cu + \text{const}$ on (a,b), then the e.s. of (15) with initial data $w(t_k,x)$ at $t = t_k$ has the form $u = w(t_k, x - c(t - t_k))$. By Theorem 1 (with the initial time t_k) for all $t > t_k$

$$\begin{aligned}\int_{\mathbb{R}} |w(t,x) - w(t_k, x - c(t-t_k))|dx &= \int_{\mathbb{R}} |s_{a,b}(u(t,x)) - s_{a,b}(w(t_k, x - c(t-t_k)))|dx \\ &\le \int_{\mathbb{R}} |u(t,x) - w(t_k, x - c(t-t_k))|dx \le \int_{\mathbb{R}} |u(t_k,x) - w(t_k,x)|dx \le 2^{-k}.\end{aligned}$$

Substituting $t = t_l$, where $l > k$, into this inequality, we obtain

$$\int_{\mathbb{R}} |w(t_l, x + ct_l) - w(t_k, x + ct_k)|dx = \int_{\mathbb{R}} |w(t,x) - w(t_k, x - c(t_l - t_k))]dx \le 2^{-k}.$$

Thus, $w(t_k, x + ct_k)$, $k \in \mathbb{N}$, is a Cauchy sequence in $\mathscr{B}^1(\mathbb{R})$. Therefore, this sequence converges as $k \to \infty$ to some function $v(x) \in \mathscr{B}^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ in $\mathscr{B}^1(\mathbb{R})$. It is clear that the segment $S(v) \subset [a,b]$ and therefore $\varphi(u) - cu = \text{const}$ on $S(v)$. Since $Sp(w(t_k, x + ct_k)) = Sp(w(t_k,\cdot)) \subset Sp(u(t_k,\cdot)) \subset M_0$, the same inclusion holds for the limit function: $Sp(v) \subset M_0$. Finally, as follows from Theorem 1, for $t > t_k$

$$\begin{aligned}\int_{\mathbb{R}} |u(t,x) - v(x - ct)|dx \le \int_{\mathbb{R}} |u(t_k,x) - w(t_k,x)|dx + \int_{\mathbb{R}} |w(t_k,x) - v(x - ct_k)|dx = \\ \int_{\mathbb{R}} |u(t_k,x) - w(t_k,x)|dx + \int_{\mathbb{R}} |w(t_k, x + ct_k) - v(x)|dx \le \\ 2^{-k} + N_1(w(t_k,\cdot + ct_k) - v) \to 0\end{aligned}$$

as $t \to +\infty$ (then also $k = \max\{\, l \mid t > t_l \,\} \to +\infty$). We see that relation (16) is satisfied. To complete the proof of Theorem 4, it only remains to notice that

$$\forall t > 0 \quad \overline{u(t,\cdot)} = \int_{\mathbb{R}} u(t,x)dx = I, \ \overline{v} = \int_{\mathbb{R}} v(x - ct)dx$$

and (16) implies that $\overline{v} = I$.

In conclusion, we show that the operator $u_0 \to v = T(u_0)$, defined in the Introduction, does not increase the distance in $\mathscr{B}^1(\mathbb{R})$.

Theorem 5. *Let* $u_{01}(x), u_{02}(x) \in \mathscr{B}^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ *and* $v_1 = T(u_{01})(x)$, $v_2 = T(u_{02})(x)$. *Then*

$$\fint_{\mathbb{R}} |v_1(x) - v_2(x)|dx \le \fint_{\mathbb{R}} |u_{01}(x) - u_{02}(x)|dx. \tag{23}$$

Proof. Let $u_1(t,x), u_2(t,x) \in C([0,+\infty), \mathscr{B}^1(\mathbb{R})) \cap L^\infty(\Pi)$ be e.s. of (15), (2) with initial data u_{01}, u_{02}, respectively. By Theorem 4

$$\delta(t) = \fint_{\mathbb{R}} |u_1(t,x) - v_1(x - c_1 t)|dx + \fint_{\mathbb{R}} |u_2(t,x) - v_2(x - c_2 t)|dx \underset{t\to+\infty}{\to} 0,$$

where c_1, c_2 are constants. We can choose a sequence $t_k > 0$ such that $t_k \to +\infty$ as $k \to \infty$, and $N_1(v_2(x + (c_1 - c_2)t_k) - v_2(x)) \le 1/k$. Then, with property (5) taken into account,

$$\begin{aligned}
\fint_{\mathbb{R}} |v_1(x) - v_2(x)|dx &= \fint_{\mathbb{R}} |v_1(x - c_1 t_k) - v_2(x - c_1 t_k)|dx \le \\
\fint_{\mathbb{R}} |v_1(x - c_1 t_k) - v_2(x - c_2 t_k)|dx &+ \fint_{\mathbb{R}} |v_2(x - c_2 t_k) - v_2(x - c_1 t_k)|dx = \\
\fint_{\mathbb{R}} |v_1(x - c_1 t_k) - v_2(x - c_2 t_k)|dx &+ \fint_{\mathbb{R}} |v_2(x + (c_1 - c_2)t_k) - v_2(x)|dx \le \\
\fint_{\mathbb{R}} |u_1(t_k, x) - u_2(t_k, x)|dx + \delta(t_k) + 1/k &\le \fint_{\mathbb{R}} |u_{01}(x) - u_{02}(x)|dx + \delta(t_k) + 1/k.
\end{aligned}$$

In the limit as $k \to \infty$, this inequality implies (23).

Remark 4. In view of Theorem 1 the map F, which associates an initial data $u_0 \in \mathscr{B}^1(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ with the e.s. $u(t,x) \in C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$ of problem (1), (2), is a uniformly continuous map from $\mathscr{B}^1(\mathbb{R}^n)$ to $C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$. Therefore, it admits the unique continuous extension on the whole space $\mathscr{B}^1(\mathbb{R}^n)$. By analogy with [2], the corresponding function $F(u_0) = u(t,x) \in C([0,+\infty), \mathscr{B}^1(\mathbb{R}^n))$ may be called a renormalized solution of (1), (2) with possibly unbounded almost periodic initial data u_0. By the approximation techniques, all our results can be extended to the case of renormalized almost periodic solutions.

Acknowledgements This work was supported by the Ministry of Education and Science of the Russian Federation (project no. 1.445.2016/1.4) and by the Russian Foundation for Basic Research (grant no. 18-01-00258-a).

References

1. A.S. Besicovitch, *Almost Periodic Functions* (Cambridge University Press, Cambridge, 1932)
2. Ph. Bénilan, J. Carrillo, P. Wittbold, Renormalized entropy solutions of scalar conservation laws. Ann. Scuola Norm. Sup. Pisa Cl. Sci. **29**, 313–327 (2000)
3. C.M. Dafermos, Long time behavior of periodic solutions to scalar conservation laws in several space dimensions. SIAM J. Math. Anal. **45**, 2064–2070 (2013)
4. N. Danford, J.T. Schwartz, *Linear Operators. General Theory (Part I)* (Interscience Publishers, New York-London, 1958)
5. H. Frid, Decay of almost periodic solutions of conservation laws. Arch. Ration. Mech. Anal. **161**, 43–64 (2002)
6. S.N. Kruzhkov, First order quasilinear equations in several independent variables. Math. USSR Sb. **10**, 217–243 (1970)
7. S.N. Kruzhkov, E.Yu. Panov, First-order conservative quasilinear laws with an infinite domain of dependence on the initial data. Soviet Math. Dokl. **42**, 316–321 (1991)
8. S.N. Kruzhkov, E.Yu. Panov, Osgood's type conditions for uniqueness of entropy solutions to Cauchy problem for quasilinear conservation laws of the first order. Ann. Univ. Ferrara Sez. VII (N.S.) **40**, 31–54 (1994)
9. B.M. Levitan, *Almost Periodic Functions* (Gostekhizdat, Moscow, 1953)
10. S. Lang, *Algebra, Revised*, 3rd edn. (Springer, New York, 2002)
11. E.Yu. Panov, A remark on the theory of generalized entropy sub- and supersolutions of the Cauchy problem for a first-order quasilinear equation. Differ. Equ. **37**, 272–280 (2001)
12. E.Yu. Panov, On generalized entropy solutions of the Cauchy problem for a first order quasilinear equation in the class of locally summable functions. Izv. Math. **66**, 1171–1218 (2002)
13. E.Yu. Panov, Existence of strong traces for generalized solutions of multidimensional scalar conservation laws. J. Hyperbolic Differ. Equ. **2**, 885–908 (2005)
14. E.Yu. Panov, On decay of periodic entropy solutions to a scalar conservation law. Ann I. H. Poincare-AN **30**, 997–1007 (2013)
15. E.Yu. Panov, On a condition of strong precompactness and the decay of periodic entropy solutions to scalar conservation laws. Netw. Heterog. Media **11**, 349–367 (2016)
16. E.Yu. Panov, Long time asymptotics of periodic generalized entropy solutions of scalar conservation laws. Math. Notes **100**, 112–121 (2016)
17. E.Yu. Panov, On the Cauchy problem for scalar conservation laws in the class of Besicovitch almost periodic functions: global well-posedness and decay property. J. Hyperbolic Differ. Equ. **13**, 633–659 (2016)

Structure Preserving Schemes for Mean-Field Equations of Collective Behavior

Lorenzo Pareschi and Mattia Zanella

Abstract In this paper, we consider the development of numerical schemes for mean-field equations describing the collective behavior of a large group of interacting agents. The schemes are based on a generalization of the classical Chang–Cooper approach and are capable to preserve the main structural properties of the systems, namely nonnegativity of the solution, physical conservation laws, entropy dissipation, and stationary solutions. In particular, the methods here derived are second order accurate in transient regimes, whereas they can reach arbitrary accuracy asymptotically for large times. Several examples are reported to show the generality of the approach.

Keywords Collective behavior · Fokker-Planck equations
Mean-field equations · Structure preserving methods

1 Introduction

The description of social dynamics characterized by emerging collective behaviors has gained increasing popularity in the recent years [1, 5, 9, 13, 14, 20]. Typical examples are groups of animals/humans with a tendency to flock or herd but also interacting agents in a financial market, potential voters during political elections, and connected members of a social network.

In the mathematical description, classical particles are replaced by more complex structures (agents, active particles, etc.) which take into account additional

L. Pareschi (✉)
Department of Mathematics and Computer Science, University of Ferrara,
Via Machiavelli 35, 44121 Ferrara, Italy
e-mail: lorenzo.pareschi@unife.it

M. Zanella
Department of Mathematical Computer Sciences, Politecnico di Torino,
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: mattia.zanella@polito.it

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_31

aspects related to the various specific fields of application, like behavioral characteristics, visual perception, experience/knowledge. Various microscopic models have been introduced in different communities with the aim to reproduce qualitatively the dynamics and to capture some essential stylized facts (clusters, power laws, consensus, flocking, etc.).

In spite of many differences between classical particle dynamics and systems of interacting agents (equation are not a consequence of fundamental physical laws derived from first principles), one can apply similar methodological approaches. In particular, to analyze the formation of stylized facts and reduce the computational complexity of the agents' dynamics, it is of utmost importance to derive the corresponding mesoscopic/kinetic description [1, 2, 6, 8, 9, 15, 20, 21].

These kinetic equations are derived in the limit of a large number of interacting agents and describe the evolution of a nonnegative distribution function $f(x, w, t)$, $t \geq 0$, $x \in \mathbb{R}^{d_x}$, $w \in \mathbb{R}^{d_w}$, $d_x, d_w \geq 1$, which satisfies a mean-field-type equation of the general form

$$\partial_t f + \mathscr{L}[f] = \nabla_w \cdot \Big[\mathscr{B}[f] f + \nabla_w (Df) \Big], \tag{1}$$

where $\mathscr{L}[\cdot](x, w, t)$ is an operator describing the agents' dynamics with respect to the x-variable, $\mathscr{B}[\cdot](x, w, t)$ is an alignment operator in the w-variable, and $D = D(x, w) \geq 0$ is a diffusion function.

The most celebrated example is given by the mean-field *Cucker–Smale* model [8, 9, 13, 20] which, in the absence of diffusion, corresponds to the choices

$$\mathscr{L}[f] = w \cdot \nabla_x f, \qquad \mathscr{B}[f] = \int_{\mathbb{R}^{d_v} \times \mathbb{R}^{d_x}} H(x, y)(w - v) f(y, v, t)\, dy\, dv, \tag{2}$$

where

$$H(x, y) = \frac{1}{(1 + (x - y)^2)^\gamma}, \qquad \gamma \geq 0. \tag{3}$$

The model describes the alignment process in a multidimensional group of agents (birds, insects, etc.), when all agents are aligned with equal speed a flocking state is reached. For the above choice of H, it has been proved that if $\gamma \leq 1/2$, independently on their initial state, all agents tend to move exponentially fast with the same velocity, while their relative distances tend to remain constant. The addition of a diffusion term weighted by $D \in \mathbb{R}^+$ has been studied in [3, 4] among others.

Another example is the nonhomogeneous mean-field *Cordier–Pareschi–Toscani* model [12, 21] which describes the evolution of the distribution $f(x, w, t)$ of wealth $w \in \mathbb{R}^+$ in a set of agents with a given propensity to invest $x \in [0, 1]$. In our notations, it corresponds to

$$\mathscr{L}[f] = \phi(x, w) \partial_x f, \qquad \mathscr{B}[f] = \int_{\mathbb{R}^+} (w - v) f(y, v, t)\, dv, \qquad D = \frac{\sigma^2}{2} w^2. \tag{4}$$

The equilibrium solutions in the homogeneous case, $f = f(w,t)$ independent of x, present the formation of power laws and read

$$f_\infty(w) = \frac{(\mu-1)^\mu}{\Gamma(\mu) w^{1+\mu}} \exp\left(-\frac{\mu-1}{w}\right), \tag{5}$$

with $\mu = 1 + 2/\sigma^2 > 1$ the *Pareto exponent* and $\int_{\mathbb{R}^+} f_\infty(w) w\, dw = 1$.

Finally, a third example is represented by the mean-field *Albi–Pareschi–Zanella* model [1, 2] describing the opinion dynamics of a group of interacting agents over a social network. The evolution of the distribution $f(x,w,t)$ of agents with a given opinion $w \in [-1,1]$ and a certain amount of discrete connections $x \in \{0,1,\dots,c_{max}\}$ is characterized by

$$\begin{aligned}
\mathscr{L}[f] &= -\frac{2V_r(f;w)}{\gamma+\beta}\left[(x+1+\beta) f(x+1,w,t) - (x+\beta) f(x,w,t)\right] \\
&\quad - \frac{2V_a(f;w)}{\gamma+\alpha}\left[(x-1+\alpha) f(x-1,w,t) - (x+\alpha) f(x,w,t)\right], \\
\mathscr{B}[f] &= \sum_{y=0}^{c_{max}} \int_{[-1,1]} P(w,v;x,y)(w-v) f(v,y,t)\, dv,
\end{aligned} \tag{6}$$

where $P(\cdot,\cdot;\cdot,\cdot) \in [0,1]$ is a compromise function, $\gamma = \gamma(t)$ is the mean density of connectivity $\gamma(t) = \sum_{x=0}^{c_{max}} x \int_{[-1,1]} f(x,w,t)\, dw$, $\alpha, \beta > 0$ are attraction coefficients, and $V_r(f;w) \geq 0$, $V_a(f;w) \geq 0$ are characteristic rates of the connections removal and adding processes, respectively.

Different equilibrium solutions in the case $f = f(w,t)$ independent of x are possible depending on the choices of P and D. For example, if $P \equiv 1$ and $D = \sigma^2(1-w^2)^2/2$, the steady state reads

$$f_\infty(w) = C_0 (1+w)^{-2+\bar{m}/\sigma^2} (1-w)^{-2-\bar{m}/\sigma^2} \exp\left\{-\frac{(1-\bar{m}w)}{\sigma^2(1-w^2)}\right\}, \tag{7}$$

where $\bar{m} = \int_{[-1,1]} w f_\infty(w)\, dw$ and C_0 is such that $\int_{[-1,1]} f_\infty(w)\, dw = 1$.

The development of numerical methods for the above class of equations is challenging due to the intrinsic structural properties of the solution [6, 7, 10, 11, 16, 19, 22]. Nonnegativity of the distribution function, conservation of invariant quantities (like moments in w of the distribution function), entropy dissipation, and homogeneous steady states are essential in order to compute qualitatively correct solutions of the mean-field equation.

In this paper, we focus on the construction of numerical methods which preserves such structural properties and, in particular, which is able to capture the correct steady state of the mean-field problem with arbitrary order of accuracy. The schemes are based on a suitable generalization of the Chang–Cooper approach to nonlinear

problems of Fokker–Planck type and are derived in the next section. Their properties are then discussed in Sect. 3. Finally, numerical results are presented in Sect. 4.

2 Derivation of the Schemes

Since most of the structural properties are related to the right-hand side in (1) in the following, we will focus on the homogeneous case $f = f(w, t)$. Connections with the full problem are then recovered using splitting methods or other partitioned time discretization schemes, like additive Runge–Kutta methods [18].

Under this assumption, we can rewrite the mean-field Eq. (1) as

$$\partial_t f(w,t) = \nabla_w \cdot [(\mathscr{B}[f](w,t) + \nabla_w D(w)) f(w,t) + D(w)\nabla_w f(w,t)]. \tag{8}$$

We define the $d-$dimensional flux function

$$\mathscr{F}[f](w,t) = (\mathscr{B}[f](w,t) + \nabla_w D(w)) f(w,t) + D(w)\nabla_w f(w,t), \tag{9}$$

so that the equation may be written in conservative form as

$$\partial_t f(w,t) = \nabla_w \cdot \mathscr{F}(w,t). \tag{10}$$

2.1 One-dimensional Case

Let us consider for notation simplicity the one-dimensional case

$$\partial_t f(w,t) = \partial_w \mathscr{F}[f](w,t), \tag{11}$$

where

$$\mathscr{F}[f](w,t) = (\mathscr{B}[f](w,t) + D'(w)) f(w,t) + D(w)\partial_w f(w,t) \tag{12}$$

and we used the notation $D'(w) = \partial_w D(w)$ and assume $D(w)$ strictly positive in the internal points of the computational domain. We introduce a uniform spatial grid w_i, $i = 0, \ldots, N$ such that $w_{i+1} - w_i = \Delta w$. We denote as usual $w_{i\pm1/2} = w_i \pm \Delta/2$ and consider the conservative discretization of Eq. (11)

$$\frac{d}{dt} f_i(t) = \frac{\mathscr{F}_{i+1/2}[f](t) - \mathscr{F}_{i-1/2}[f](t)}{\Delta w}, \tag{13}$$

where for each $t \geq 0$, $f_i(t)$ is an approximation of $f(w_i, t)$ and $\mathscr{F}_{i\pm1/2}[f](t)$ is the flux function characterizing the discretization.

Let us set $\mathscr{C}[f](w,t) = \mathscr{B}[f](w,t) + D'(w)$ and adopt the notations $\mathscr{B}_{i+1/2} = \mathscr{B}[f](w_{i+1/2},t)$, $D_{i+1/2} = D(w_{i+1/2})$, $D'_{i+1/2} = D'(w_{i+1/2})$. We will consider a general flux function which is combination of the grid points $i+1$ and i as in [11, 22]

$$\mathscr{F}_{i+1/2}[f] = \tilde{\mathscr{C}}_{i+1/2}\tilde{f}_{i+1/2} + D_{i+1/2}\frac{f_{i+1}-f_i}{\Delta w}, \tag{14}$$

where

$$\tilde{f}_{i+1/2} = (1-\delta_{i+1/2})f_{i+1} + \delta_{i+1/2}f_i. \tag{15}$$

Here, we aim at deriving suitable expressions for $\delta_{i+1/2}$ and $\tilde{\mathscr{C}}_{i+1/2}$ in such a way that the method yields nonnegative solutions, without restrictions on Δw, and preserves the steady state of the system with arbitrary accuracy.

For example, the standard approach based on central difference is obtained taking $\delta_{i+1/2} = 1/2$ and $\tilde{\mathscr{C}}_{i+1/2} = \mathscr{B}_{i+1/2}$, $\forall\, i$. It is well known, however, that such a discretization method is subject to restrictive conditions over the mesh size Δw in order to keep nonnegativity of the solution.

First, observe that at the steady state the numerical flux equal should vanish. From (14), we get

$$\frac{f_{i+1}}{f_i} = \frac{-\delta_{i+1/2}\tilde{\mathscr{C}}_{i+1/2} + \dfrac{D_{i+1/2}}{\Delta w}}{(1-\delta_{i+1/2})\tilde{\mathscr{C}}_{i+1/2} + \dfrac{D_{i+1/2}}{\Delta w}}. \tag{16}$$

Similarly, if we consider the analytical flux at the steady state, we have

$$D(w)\partial_w f(w,t) = -(\mathscr{B}[f] + D'(w))f(w,t), \tag{17}$$

which is in general not solvable, except in some special cases due to the nonlinearity on the right-hand side. We may overcome this difficulty in the quasi-steady-state approximation integrating equation (17) on the cell $[w_i, w_{i+1}]$

$$\int_{w_i}^{w_{i+1}} \frac{1}{f(w,t)}\partial_w f(w,t)dw = -\int_{w_i}^{w_{i+1}} \frac{1}{D(w)}(\mathscr{B}[f](w,t) + D'(w))dw, \tag{18}$$

which gives

$$\frac{f_{i+1}}{f_i} = \exp\Big\{-\int_{w_i}^{w_{i+1}} \frac{1}{D(w)}(\mathscr{B}[f](w,t) + D'(w))dw\Big\}, \tag{19}$$

for all $i = 1, \dots, N-1$.

Now, by equating the ratio f_{i+1}/f_i of the numerical and the exact flux and setting

$$\tilde{\mathscr{C}}_{i+1/2} = \frac{D_{i+1/2}}{\Delta w}\int_{w_i}^{w_{i+1}} \frac{\mathscr{B}[f](w,t) + D'(w)}{D(w)}dw \tag{20}$$

we recover

$$\delta_{i+1/2} = \frac{1}{\lambda_{i+1/2}} + \frac{1}{1 - \exp(\lambda_{i+1/2})}, \tag{21}$$

where

$$\lambda_{i+1/2} = \int_{w_i}^{w_{i+1}} \frac{\mathscr{B}[f](w,t) + D'(w)}{D(w)} dw. \tag{22}$$

Remark 1. A second-order method is obtained by discretizing (22) through the midpoint rule

$$\int_{w_i}^{w_{i+1}} \frac{\mathscr{B}[f](w,t) + D'(w)}{D(w)} dw \approx \frac{\Delta w(\mathscr{B}_{i+1/2} + D'_{i+1/2})}{D_{i+1/2}}, \tag{23}$$

therefore

$$\lambda_{i+1/2}^{\text{mid}} = \frac{\Delta w(\mathscr{B}_{i+1/2} + D'_{i+1/2})}{D_{i+1/2}} \tag{24}$$

and

$$\delta_{i+1/2}^{\text{mid}} = \frac{D_{i+1/2}}{\Delta w(\mathscr{B}_{i+1/2} + D'_{i+1/2})} + + \frac{1}{1 - \exp(\lambda_{i+1/2}^{\text{mid}})}. \tag{25}$$

Higher order accuracy of the steady-state solution may be obtained by higher order approximations of the integral (20).

2.2 The Multidimensional Case

In order to extend the previous approach to multidimensional situations, we consider here the case of two-dimensional problems. We introduce a mesh consisting of the cells $C_{ij} = [w_{i-1/2}, w_{i+1/2}] \times [v_{j-1/2}, v_{j+1/2}]$ assumed to be of uniform size $\Delta w \Delta v$, where as usual $\Delta w := w_{i+1/2} - w_{i-1/2}$ and $\Delta v := v_{j+1/2} - v_{j-1/2}$ for all $i = 0, \ldots, N_1$ and $j = 0, \ldots, N_2$. Integration of the general mean-field equation in dimension $d \geq 1$ introduced in (10) yields

$$\frac{d}{dt} f_{i,j} = \frac{\mathscr{F}_{i+1/2,j}[f] - \mathscr{F}_{i-1/2,j}[f]}{\Delta w} + \frac{\mathscr{F}_{i,j+1/2}[f] - \mathscr{F}_{i,j-1/2}[f]}{\Delta v}, \tag{26}$$

being $\mathscr{F}_{i\pm1/2,j}[f]$, $\mathscr{F}_{i,j\pm1/2}[f]$ flux functions characterizing the numerical discretization. The quasi-stationary approximations over the cell $[w_i, w_{i+1}] \times [v_i, v_{i+1}]$ of the two-dimensional problem read

$$\begin{aligned}\int_{w_i}^{w_{i+1}} \frac{1}{f(w,v_j,t)}\partial_w f(w,v_j,t)dw &= -\int_{w_i}^{w_{i+1}} \frac{\mathscr{B}[f](w,v_j,t)+\partial_w D(w,v_j)}{D(w,v_j)}dw,\\ \int_{v_j}^{v_{j+1}} \frac{1}{f(w_i,v,t)}\partial_v f(w_i,v,t)dv &= -\int_{v_j}^{v_{j+1}} \frac{\mathscr{B}[f](w_i,v,t)+\partial_v D(w_i,v)}{D(w_i,v)}dv.\end{aligned} \tag{27}$$

Therefore, setting

$$\begin{aligned}\tilde{\mathscr{C}}_{i+1/2,j} &= \frac{D_{i+1/2,j}}{\Delta w}\int_{w_i}^{w_{i+1}} \frac{\mathscr{B}[f](w,v_j,t)+\partial_w D(w,v_j)}{D(w,v_j)}dw\\ \tilde{\mathscr{C}}_{i,j+1/2} &= \frac{D_{i,j+1/2}}{\Delta v}\int_{v_j}^{v_{j+1}} \frac{\mathscr{B}[f](w_i,v,t)+\partial_v D(w_i,v)}{D(w_i,v)}dv\end{aligned} \tag{28}$$

and by considering an analogous flux components by components as in the one-dimensional case

$$\begin{aligned}\mathscr{F}_{i+1/2,j}[f] &= \tilde{\mathscr{C}}_{i+1/2,j}\tilde{f}_{i+1/2,j} + D_{i+1/2,j}\frac{f_{i+1,j}-f_{i,j}}{\Delta w}\\ \tilde{f}_{i+1/2,j} &= (1-\delta_{i+1/2,j})f_{i+1,j} + \delta_{i+1/2,j}f_{i,j}\\ \mathscr{F}_{i,j+1/2}[f] &= \tilde{\mathscr{C}}_{i,j+1/2}\tilde{f}_{i,j+1/2} + D_{i,j+1/2}\frac{f_{i,j+1}-f_{i,j}}{\Delta v}\\ \tilde{f}_{i,j+1/2} &= (1-\delta_{i,j+1/2})f_{i,j+1} + \delta_{i,j+1/2}f_{i,j},\end{aligned} \tag{29}$$

we define $\delta_{i+1/2,j}$ and $\delta_{i,j+1/2}$ in such a way that we preserve the steady-state solution for each dimension, i.e.,

$$\begin{aligned}\delta_{i+1/2,j} &= \frac{1}{\lambda_{i+1/2,j}} + \frac{1}{1-\exp(\lambda_{i+1/2,j})},\\ \delta_{i,j+1/2} &= \frac{1}{\lambda_{i,j+1/2}} + \frac{1}{1-\exp(\lambda_{i,j+1/2})}\\ \lambda_{i+1/2,j} &= \frac{\Delta w\tilde{\mathscr{C}}_{i+1/2,j}}{D_{i+1/2,j}}, \qquad \lambda_{i,j+1/2} = \frac{\Delta v\tilde{\mathscr{C}}_{i,j+1/2}}{D_{i,j+1/2}}.\end{aligned} \tag{30}$$

The cases of higher dimension $d \geq 3$ may be derived in a similar way.

3 Main Properties

In order to study the structural properties of the numerical scheme, like nonnegativity and entropy property, we restrict to the one-dimensional case.

3.1 Nonnegativity

We introduce a time discretization $t^n = n\Delta t$ with $\Delta t > 0$ and $n = 0, \dots, T$ and consider the simple forward Euler method

$$f_i^{n+1} = f_i^n + \Delta t \frac{\mathscr{F}_{i+1/2}^n - \mathscr{F}_{i-1/2}^n}{\Delta w}, \tag{31}$$

with no-flux boundary conditions $F_{N+1/2}^n = \mathscr{F}_{-1/2}^n = 0$.

Lemma 1. *Let us consider the scheme (31) with no-flux boundary conditions. We have for all* $n \in \mathbb{N}$

$$\sum_{i=0}^{N} f_i^{n+1} = \sum_{i=0}^{N} f_i^n. \tag{32}$$

Proof. From Eq. (31), we have

$$\sum_{i=0}^{N} f_i^{n+1} = \sum_{i=0}^{N} f_i^n + \frac{\Delta t}{\Delta w} \sum_{i=0}^{N} (\mathscr{F}_{i+1/2}^n - \mathscr{F}_{i-1/2}^n). \tag{33}$$

Now since

$$\sum_{i=0}^{N} (\mathscr{F}_{i+1/2}^n - \mathscr{F}_{i-1/2}^n) = \mathscr{F}_{N+1/2}^n - \mathscr{F}_{-1/2}^n, \tag{34}$$

by imposing no-flux boundary conditions, we conclude.

Note that mass conservation holds true also in the backward Euler case by imposing $\mathscr{F}_{N+1/2}^{n+1} = \mathscr{F}_{-1/2}^{n+1} = 0$.

Concerning nonnegativity, we can prove [22]

Proposition 1. *Under the time-step restriction*

$$\Delta t \le \frac{\Delta w^2}{2(M\Delta w + D)}, \quad M = \max_{0 \le i \le N} |\tilde{\mathscr{C}}_{i+1/2}^n|, \quad D = \max_{0 \le i \le N} D_{i+1/2}, \tag{35}$$

the explicit scheme (31) preserves nonnegativity, i.e $f_i^{n+1} \ge 0$ *if* $f_i^n \ge 0$, $i = 0, \dots, N$.

Proof. The scheme reads

$$f_i^{n+1} = f_i^n + \frac{\Delta t}{\Delta w}\Bigg[\left((1-\delta_{i+1/2}^n)\tilde{\mathscr{C}}_{i+1/2}^n + \frac{D_{i+1/2}}{\Delta w}\right) f_{i+1}^n$$
$$+\left(\tilde{\mathscr{C}}_{i+1/2}^n\delta_{i+1/2}^n - \tilde{\mathscr{C}}_{i-1/2}^n(1-\delta_{i-1/2}^n) - \frac{1}{\Delta w}(D_{i+1/2}+D_{i-1/2})\right) f_i^n \tag{36}$$
$$-\left(\tilde{\mathscr{C}}_{i-1/2}^n\delta_{i-1/2}^n - \frac{D_{i-1/2}}{\Delta w}\right) f_{i-1}^n\Bigg].$$

From (36), the coefficients of f_{i+1}^n and f_{i-1}^n should satisfy

$$(1-\delta_{i+1/2})\tilde{\mathscr{C}}_{i+1/2}^n + \frac{D_{i+1/2}}{\Delta w} \geq 0, \qquad -\delta_{i-1/2}\tilde{\mathscr{C}}_{i-1/2}^n + \frac{D_{i-1/2}}{\Delta w} \geq 0, \tag{37}$$

that is equivalent to show that

$$\lambda_{i+1/2}\left(1 - \frac{1}{1-\exp\lambda_{i+1/2}}\right) \geq 0, \qquad \frac{\lambda_{i-1/2}}{\exp\lambda_{i-1/2}-1} \geq 0, \tag{38}$$

which holds true thanks to the properties of the exponential function. In order to ensure the nonnegativity of the scheme the, time step should satisfy the restriction $\Delta t \leq \Delta w/\nu$, with

$$\nu = \max_{0\leq i\leq N}\left\{\tilde{\mathscr{C}}_{i+1/2}^n\delta_{i+1/2}^n - \tilde{\mathscr{C}}_{i-1/2}^n(1-\delta_{i-1/2}^n) - \frac{D_{i+1/2}+D_{i-1/2}}{\Delta w}\right\}. \tag{39}$$

Being M defined in (35), and $0 \leq \delta_{i\pm1/2} \leq 1$, we obtain the prescribed bound.

Remark 2. Higher order SSP methods [17] are obtained by considering a convex combination of forward Euler methods. Therefore, the non negativity result can be extended to general SSP methods.

In practical applications, it is desirable to avoid the parabolic restriction $\Delta t = O((\Delta w)^2)$ of explicit schemes. Unfortunately, fully implicit methods originate a nonlinear system of equations. However, we can prove that nonnegativity of the solution holds true also for the semi-implicit case

$$f_i^{n+1} = f_i^n + \Delta t\frac{\hat{\mathscr{F}}_{i+1/2}^{n+1} - \hat{\mathscr{F}}_{i-1/2}^{n+1}}{\Delta w}, \tag{40}$$

where

$$\hat{\mathscr{F}}_{i+1/2}^{n+1} = \tilde{\mathscr{C}}_{i+1/2}^n\left[(1-\delta_{i+1/2}^n)f_{i+1}^{n+1} + \delta_{i+1/2}f_i^{n+1}\right] + D_{i+1/2}\frac{f_{i+1}^{n+1} - f_i^{n+1}}{\Delta w}. \tag{41}$$

We have [22]

Proposition 2. *Under the time-step restriction*

$$\Delta t < \frac{\Delta w}{2M}, \qquad M = \max_{0\le i\le N} |\tilde{\mathscr{B}}^n_{i+1/2}| \tag{42}$$

the semi-implicit scheme (40) preserves nonnegativity, i.e $f_i^{n+1} \ge 0$ *if* $f_i^n \ge 0$, $i = 0, \dots, N$.

Proof. Setting $\alpha^n_{i+1/2} = \dfrac{\lambda^n_{i+1/2}}{\exp(\lambda^n_{i+1/2}) - 1}$ and

$$\begin{aligned} R_i^n &= 1 + \frac{\Delta t}{\Delta w^2}\left[D_{i+1/2}\alpha^n_{i+1/2} + D_{i-1/2}\alpha^n_{i-1/2}\exp(\lambda^n_{i-1/2})\right] \\ Q_i^n &= \frac{\Delta t}{\Delta w^2} D_{i+1/2}\alpha^n_{i+1/2}\exp(\lambda^n_{i+1/2}) \\ P_i^n &= \frac{\Delta t}{\Delta w^2} D_{i-1/2}\alpha^n_{i-1/2}, \end{aligned} \tag{43}$$

Equation (40) corresponds to

$$R_i^n f_i^{n+1} - Q_i^n f_{i+1}^{n+1} - P_i^n f_{i-1}^{n+1} = f_i^n. \tag{44}$$

If we introduce the matrix

$$(\mathscr{A}[f^n])_{ij} = \begin{cases} R_i^n, & j = i \\ -Q_i^n, & j = i+1,\ 1 \le i \le N \\ -P_i^n, & j = i-1,\ 0 \le i \le N-1, \end{cases} \tag{45}$$

with $R_i^n > 0$, $Q_i^n > 0$, $P_i^n > 0$ defined in (43), the semi-implicit scheme may be expressed in matrix form as follows

$$\mathscr{A}[\mathbf{f}^n]\mathbf{f}^{n+1} = \mathbf{f}^n, \tag{46}$$

with $\mathbf{f}^n = \left(f_0^n, \dots, f_N^n\right)$. Now, the matrix $\mathscr{A}$ is strictly diagonally dominant if and only if

$$|R_i^n| > |Q_i^n| + |P_i^n|, \qquad i = 0, 1 \dots, N, \tag{47}$$

condition which holds true if

$$\begin{aligned} 1 &> \frac{\Delta t}{\Delta w^2}\left[D_{i+1/2}\alpha^n_{i+1/2}\left(\exp(\lambda^n_{i+1/2}) - 1\right) - D_{i-1/2}\alpha^n_{i-1/2}\left(\exp(\lambda^n_{i-1/2}) - 1\right)\right] \\ &= \frac{\Delta t}{\Delta w^2}\left[D_{i+1/2}\lambda^n_{i+1/2} - D_{i-1/2}\lambda^n_{i-1/2}\right] = \frac{\Delta t}{\Delta w}\left[\tilde{\mathscr{B}}^n_{i+1/2} - \tilde{\mathscr{B}}^n_{i-1/2}\right]. \end{aligned} \tag{48}$$

3.2 Entropy Property

In order to discuss the entropy property, we consider the prototype equation [15, 22]

$$\partial_t f(w,t) = \partial_w \left[(w-u) f(w,t) + \partial_w (D(w) f(w,t)) \right], \qquad w \in I = [-1,1], \tag{49}$$

with $-1 < u < 1$ a given constant and boundary conditions

$$\partial_w (D(w) f(w,t)) + (w-u) f(w,t) = 0, \qquad w = \pm 1. \tag{50}$$

If the stationary state f^∞ exists, Eq. (49) may be written in the form

$$\partial_t f(w,t) = \partial_w \left[D(w) f^\infty(w) \partial_w \left(\frac{f(w,t)}{f^\infty(w)} \right) \right]. \tag{51}$$

We define the relative entropy for all positive functions $f(w,t)$, $g(w,t)$ as follows

$$\mathscr{H}(f,g) = \int_I f(w,t) \log \left(\frac{f(w,t)}{g(w,t)} \right), \tag{52}$$

and we have [15]

$$\frac{d}{dt} \mathscr{H}(f, f^\infty) = -\mathscr{I}_D(f, f^\infty), \tag{53}$$

where the dissipation functional $\mathscr{I}_D(\cdot,\cdot)$ is defined as

$$\begin{aligned} \mathscr{I}_D(f, f^\infty) &= \int_{\mathscr{I}} D(w) f(w,t) \left(\partial_w \log \left(\frac{f(w,t)}{f^\infty(w)} \right) \right)^2 dw, \\ &= \int_{\mathscr{I}} D(w) f^\infty(w,t) \partial_w \log \left(\frac{f(w,t)}{f^\infty(w)} \right) \partial_w \left(\frac{f}{f^\infty} \right) dw. \end{aligned} \tag{54}$$

Lemma 2. *In the case $\mathscr{B}[f](w,t) = \mathscr{B}(w)$, the numerical flux function (14)–(15) with $\tilde{\mathscr{B}}_{i+1/2}$ and $\delta_{i+1/2}$ given by (20)–(21) can be written in the form (51) and reads*

$$\mathscr{F}_{i+1/2} = \frac{D_{i+1/2}}{\Delta w} \hat{f}^\infty_{i+1/2} \left(\frac{f_{i+1}}{f^\infty_{i+1}} - \frac{f_i}{f^\infty_i} \right), \tag{55}$$

with

$$\hat{f}^\infty_{i+1/2} = \frac{f^\infty_{i+1} f^\infty_i}{f^\infty_{i+1} - f^\infty_i} \log \left(\frac{f^\infty_{i+1}}{f^\infty_i} \right). \tag{56}$$

Proof. In the hypothesis $\mathscr{B}[f](w,t) = \mathscr{B}(w)$, the definition of $\lambda_{i+1/2}$ does not depend on time, i.e., $\lambda_{i+1/2} = \lambda^\infty_{i+1/2}$, and if a steady state exists, we may write

$$\log f_i^\infty - \log f_{i+1}^\infty = \lambda_{i+1/2}. \tag{57}$$

Furthermore, the flux function $\mathscr{F}_{i+1/2}$ assumes the following form

$$\begin{aligned}\mathscr{F}_{i+1/2} &= \frac{D_{i+1/2}}{\Delta w}\left[\lambda_{i+1/2}\tilde{f}_{i+1/2} + (f_{i+1} - f_i)\right] \\ &= \frac{D_{i+1/2}}{\Delta w}\left[\lambda_{i+1/2}(f_{i+1} + \delta_{i+1/2}(f_i - f_{i+1})) + (f_{i+1} - f_i)\right],\end{aligned} \tag{58}$$

where

$$\delta_{i+1/2} = \frac{1}{\log f_i^\infty - \log f_{i+1}^\infty} + \frac{f_{i+1}^\infty}{f_{i+1}^\infty - f_i^\infty}. \tag{59}$$

Hence, we have

$$\begin{aligned}\mathscr{F}_{i+1/2}^n &= \frac{D_{i+1/2}}{\Delta w}\log\left(\frac{f_i^\infty}{f_{i+1}^\infty}\right)\left[f_{i+1} + \left(\frac{f_i - f_{i+1}}{\log f_i^\infty - \log f_{i+1}^\infty} + \frac{f_{i+1}^\infty(f_i - f_{i+1})}{f_{i+1}^\infty - f_i^\infty}\right)\right. \\ &\qquad \left. + \frac{f_{i+1} - f_i}{\log f_i^\infty - \log f_{i+1}^\infty}\right], \\ &= \frac{D_{i+1/2}}{\Delta w}\log\left(\frac{f_i^\infty}{f_{i+1}^\infty}\right)\left(\frac{f_{i+1}^\infty f_i - f_i^\infty f_{i+1}}{f_{i+1}^\infty - f_i^\infty}\right)\end{aligned} \tag{60}$$

which gives (55).

Theorem 1. *Let us consider $\mathscr{B}[f](w,t) = w - u$ as in Eq. (49). The numerical flux (14)–(15) with $\tilde{\mathscr{B}}_{i+1/2}$ and $\delta_{i+1/2}$ given by (20)–(21) satisfies the discrete entropy dissipation*

$$\frac{d}{dt}\mathscr{H}_\Delta(f, f^\infty) = -\mathscr{I}_\Delta(f, f^\infty), \tag{61}$$

where

$$\mathscr{H}_{\Delta w}(f, f^\infty) = \Delta w \sum_{i=0}^{N} f_i \log\left(\frac{f_i}{f_i^\infty}\right) \tag{62}$$

and I_Δ is the positive discrete dissipation function

$$\mathscr{I}_\Delta(f, f^\infty) = \sum_{i=0}^{N}\left[\log\left(\frac{f_{i+1}}{f_{i+1}^\infty}\right) - \log\left(\frac{f_i}{f_i^\infty}\right)\right]\cdot\left(\frac{f_{i+1}}{f_{i+1}^\infty} - \frac{f_i}{f_i^\infty}\right)\hat{f}_{i+1/2}^\infty D_{i+1/2} \geq 0. \tag{63}$$

Proof. From the definition of relative entropy, we have

$$\begin{aligned}\frac{d}{dt}\mathscr{H}(f, f^\infty) &= \Delta w \sum_{i=0}^{N} \frac{df_i}{dt}\left(\log\left(\frac{f_i}{f_i^\infty}\right)+1\right)\\ &= \Delta w \sum_{i=0}^{N}\left(\log\left(\frac{f_i}{f_i^\infty}\right)+1\right)(\mathscr{F}_{i+1/2}-\mathscr{F}_{i-1/2}),\end{aligned} \tag{64}$$

and after summation by parts we get

$$\frac{d}{dt}\mathscr{H}(f, f^\infty) = -\Delta w \sum_{i=0}^{N}\left[\log\left(\frac{f_{i+1}}{f_{i+1}^\infty}\right)-\log\left(\frac{f_i}{f_i^\infty}\right)\right]\mathscr{F}_{i+1/2}. \tag{65}$$

Thanks to the identity of Lemma 2, we may conclude since the function $(x - y)\log(x/y)$ is nonnegative for all $x, y \geq 0$.

4 Numerics

In this section, we present several numerical tests for the proposed structure-preserving schemes. In particular, we show that the schemes accurately describe the steady-state solution of mean-field equations.

Test 1: Accuracy and Steady States
Let us consider the evolution of a distribution described by Eq. (49) with

$$u = \int_I v f(v,t)dv, \qquad D(w) = \frac{\sigma^2}{2}(1-w^2)^2. \tag{66}$$

We consider as initial distribution

$$f(w,0) = \beta\left[\exp\{-c(w+1/2)\}+\exp\{-c(w-1/2)\}\right], \qquad c = 30, \tag{67}$$

and $\beta > 0$ a normalization constant. The stationary solution in this case can be explicitly computed and is given by (7).

We compute the relative L^1 error of the solution with respect to the stationary state using $N = 41$ points. In Fig. 1, we show the evolution of the mean-field equation and the relative L^1 error in approximating the steady state solution. We used open Newton–Cotes formulas of various orders and Gaussian quadrature to evaluate (22). It is possible to observe how the different integration methods capture the steady state with different accuracies. In particular using Gaussian quadrature, we essentially reached machine precision.

In Table 1, we estimate the overall order of convergence of the scheme for various integration methods. Here, we used $N = 41, 81, 161$ grid points. The time integration has been performed with an explicit RK4 method, and the time step is chosen in such

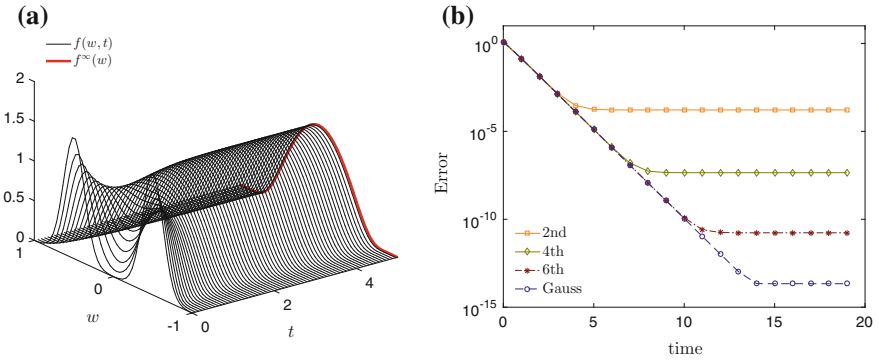


Fig. 1 Test 1. **a** Time evolution of the density $f(w,t)$ for problem (49) with initial datum (67) over the time interval $[0,5]$ for $\sigma^2/2 = 0.1$, $\Delta w = 0.05$. **b** Evolution of the relative L^1 error with respect to the stationary solution (7) for various quadrature methods

Table 1 Test 1. Estimation of the order of convergence toward the reference stationary state for each integration method at different times

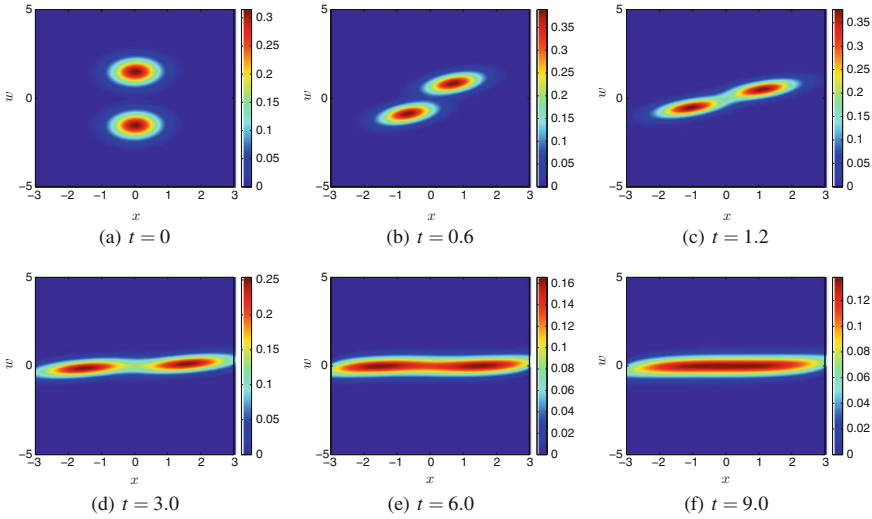
	Second	Fourth	Sixth	Gauss
$T = 1$	1.8676	1.9972	1.9958	1.9958
	1.9840	1.9991	1.9987	1.9987
$T = 5$	1.9348	3.2518	2.3578	2.3344
	2.0043	2.6218	2.0948	2.0930
$T = 10$	1.9289	3.9178	6.4645	7.3482
	2.0034	3.9185	6.3630	7.9217
$T = 15$	1.9289	3.9178	6.4701	7.3512
	2.0034	3.9786	6.6021	7.9954

a way that the CFL condition for the positivity of the scheme is satisfied; therefore, $\Delta t = O((\Delta w)^2)$. As expected, the methods are second order accurate in transient regimes and, as they approach the steady state, they reach the order of the quadrature method. Clearly, the order of Gaussian quadrature is bounded by the maximum observable order which is eight due to the choice of the time discretization method.

Test 2: Flocking Dynamics

We consider a mean-field Cucker–Smale flocking model as introduced in (2). The space variable is discretized using a third-order WENO scheme, and the transport and interaction process are combined using a second-order Strang splitting scheme. For the mean-field term, we considered a semi-implicit scheme with Gaussian quadrature of the weights. This choice guarantees spectral accuracy for the description of the steady-state solution of the equation.

In Fig. 2, we report the evolution of the solution $f(x,w,t)$ in the phase space $(x,w) \in [-3,3] \times [-5,5]$ with $\Delta x = 6 \cdot 10^{-2}$ and $\Delta w = 5 \cdot 10^{-2}$. The time step has been chosen in order to satisfy the CFL condition $\Delta t/\Delta x = 0.25/\max(w)$.



(a) $t = 0$ (b) $t = 0.6$ (c) $t = 1.2$

(d) $t = 3.0$ (e) $t = 6.0$ (f) $t = 9.0$

Fig. 2 Test 2. Mean-field Cucker–Smale model for $(x, w) \in [-3, 3] \times [-5, 5]$ with $\Delta x = 6 \cdot 10^{-2}$ and $\Delta w = 5 \cdot 10^{-2}$, $\Delta t / \Delta x = 0.25 / \max(w)$. We considered $\gamma = 0.1$ in (3) and a constant diffusion function $D = 0.1$

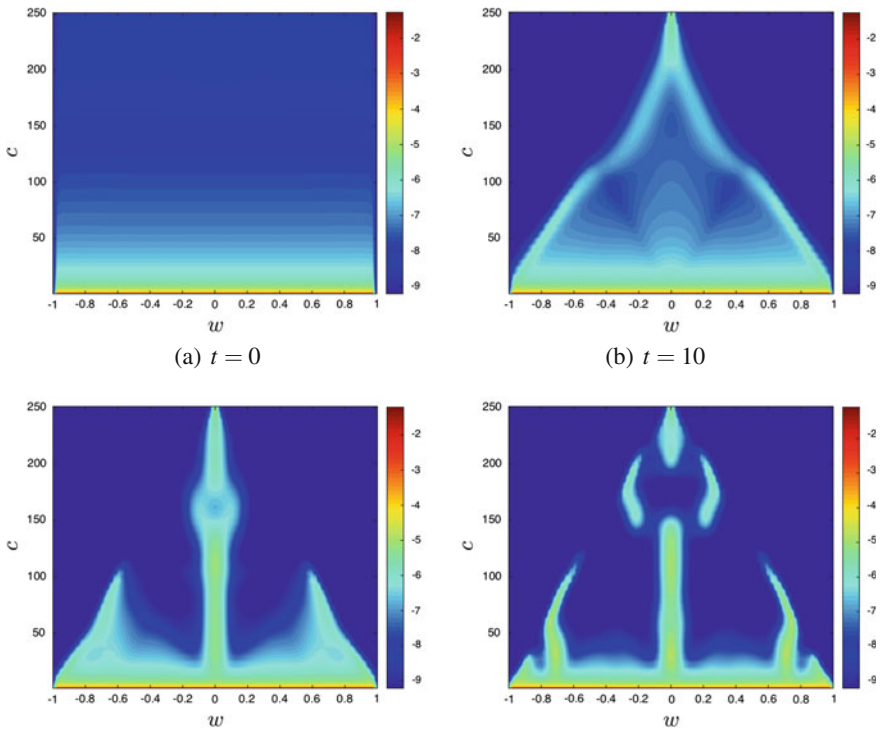
We considered $\gamma = 0.1 < 1/2$ in the Cucker–Smale interaction function (3) and a constant diffusion $D(x, w) = 0.1$. The initial datum is here given by a multivariate population which shares the same average space location $x = 0$ and is strongly clustered around opposite velocities $v = \pm 1.5$. As expected, the whole system converges to the same velocity; i.e., the distribution tends to concentrate in the velocity space and to be distributed uniformly along the spatial dimension.

Test 3: Opinion on Networks

Finally, we consider the model of opinion on networks (6). We focus on the case of a connection-dependent bounded confidence model, where the agents interact only within a certain range of confidence. Hence, we define the compromise function [2]

$$P(w, v; x, y) = \chi_{\{|w-v| \le \Delta(x)\}}(v), \tag{68}$$

where $\Delta(x) = d_0 \dfrac{x}{c_{\max}}$ and $D(w, x) = (1 - w^2)^2$. This choice reflects a behavior where agents with higher number of connections are prone to larger level of confidence. We report in Fig. 3 the evolution of the solution (where in order to better show its evolution we plotted $\log(f(w, x, t) + \varepsilon)$, with $\varepsilon = 0.001$). We can observe how the introduction of the function $\Delta(c)$ creates a heterogeneous emergence of clusters with respect to the connectivity level: For higher level of connectivity, consensus is reached, since the bounded confidence level is larger; instead for lower levels of connectivity, multiple clusters appear. In the limiting case $c = 0$, the opinions are not influenced by the consensus dynamics.



(a) $t = 0$ (b) $t = 10$

(c) $t = 50$ (d) t=100

Fig. 3 Test 3. Evolution of the solution of the mean-field model (6) with uniform initial opinion and power-law-type connection distribution. The interaction is described by (68) with $d_0 = 1.01$, in the time interval [0, 100]. The other parameters are $\sigma^2 = 10^{-3}$, $c_{\max} = 250$, $V_r = V_a = 1$, $\gamma(0) = 30$, $\alpha = 10^{-1}$, $\beta = 0$

References

1. G. Albi, L. Pareschi, G. Toscani, M. Zanella, Recent advances in opinion modeling: control and social influence, in *Active Particles, Volume 1. Modeling and Simulation in Science, Engineering and Technology*, ed. by N. Bellomo, P. Degond, E. Tadmor (Birkhäuser, cham, 2017)
2. G. Albi, L. Pareschi, M. Zanella, Opinion dynamics over complex networks: kinetic modeling and numerical methods. Kinet. Relat. Models **10**(1), 1–32 (2017)
3. A.B.T. Barbaro, P. Degond, Phase transition and diffusion among socially interacting self-propelled agents. Discret. Contin. Dyn. Syst. - Ser. B **19**, 1249–1278 (2014)
4. F. Bolley, J.A. Carrillo, Stochastic mean-field limit: non-Lipschitz forces and swarming. Math. Models Methods Appl. Sci. **21**(11), 2179 (2011)
5. N. Bellomo, G. Ajmone Marsan, A. Tosin, *Complex Systems and Society. Modeling and Simulation*, Springer Briefs in Mathematics (Springer, Berlin, 2013)
6. C. Buet, S. Dellacherie, On the Chang and Cooper numerical scheme applied to a linear Fokker-Planck equation. Commun. Math. Sci. **8**(4), 1079–1090 (2010)
7. C. Buet, S. Cordier, V. Dos Santos, A conservative and entropy scheme for a simplified model of granular media. Transp. Theory Stat. Phys. **33**(2), 125–155 (2004)

8. J.A. Carrillo, M. Fornasier, J. Rosado, G. Toscani, Asymptotic flocking dynamics for the kinetic Cucker-Smale model. SIAM J. Math. Anal. **42**(1), 218–236 (2010)
9. J.A. Carrillo, M. Fornasier, G. Toscani, F. Vecil, in *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*. Particle, Kinetic and Hydrodynamic Models of Swarming (Birkhuser, Boston, 2010), pp. 297–336
10. J.A. Carrillo, A. Chertock, Y. Huang, A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. Commun. Comput. Phys. **17**, 233–258 (2015)
11. J.S. Chang, G. Cooper, A practical difference scheme for Fokker-Planck equations. J. Comput. Phys. **6**(1), 1–16 (1970)
12. S. Cordier, L. Pareschi, G. Toscani, On a kinetic model for a simple market economy. J. Stat. Phys. **120**(1–2), 253–277 (2005)
13. F. Cucker, S. Smale, Emergent behavior in flocks. IEEE Trans. Autom. Control **52**(5), 852–862 (2007)
14. M.R. D'Orsogna, Y.L. Chuang, A.L. Bertozzi, L. Chayes, Self-propelled particles with soft-core interactions: patterns, stability and collapse. Phys. Rev. Lett. **96**, 104302 (2006)
15. G. Furioli, A. Pulvirenti, E. Terraneo, G. Toscani, Fokker-Planck equations in the modelling of socio-economic phenomena. Math. Models Methods Appl. Sci. **27**(1), 115–158 (2017)
16. L. Gosse, *Computing qualitatively correct approximations of Balance Laws. Exponential-Fit, Well-Balanced and Asymptotic-Preserving*, SEMA SIMAI Springer Series (Springer, Berlin, 2013)
17. S. Gottlieb, C.W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods. SIAM Rev. **43**(1), 89–112 (2001)
18. E. Hairer, S.P. Norsett, G. Wanner, *Solving Ordinary Differential Equation I: Nonstiff Problems*, vol. 8, Springer Series in Comput. Mathematics (Springer, Berlin, 1987). Second revised edition 1993
19. E.W. Larsen, C.D. Levermore, G.C. Pomraning, J.G. Sanderson, Discretization methods for one-dimensional Fokker-Planck operators. J. Comput. Phys. **61**(3), 359–390 (1985)
20. L. Pareschi, G. Toscani, *Interacting Multiagent Systems: Kinetic Equations and Monte Carlo Methods* (Oxford University Press, Oxford, 2013)
21. L. Pareschi, G. Toscani, Wealth distribution and collective knowledge: a Boltzmann approach. Philos. Trans. R. Soc. Lond. Ser. A. Math. Phys. Eng. Sci. **372**(2028), 20130396 (2014)
22. L. Pareschi, M. Zanella, Structure preserving schemes for nonlinear Fokker-Planck equations and applications. J. Sci. Comput. **74**(3), 1575–1600 (2018)

A Numerical Model for Three-Phase Liquid–Vapor–Gas Flows with Relaxation Processes

Tore Flåtten, Marica Pelanti and Keh-Ming Shyue

Abstract We are interested in three-phase flows involving the liquid and vapor phases of one species and a third inert gaseous phase. We describe these flows by a single-velocity multiphase flow model composed of the phasic mass and total energy equations, the volume fraction equations, and the mixture momentum equation. The model includes stiff mechanical and thermal relaxation source terms for all the phases and chemical relaxation terms to describe mass transfer between the liquid and vapor phases of the species that may undergo transition. The homogeneous hyperbolic portion of the equations is solved numerically via a finite volume wave propagation scheme. Relaxation terms are treated by routines that exploit algebraic equilibrium conditions for the relaxed states. We present numerical results for a three-phase cavitation tube test, showing that the predicted wave speed for different levels of activation of instantaneous relaxation processes agrees with the theoretical findings on the sub-characteristic interlacing of the wave speeds of the corresponding hierarchy of relaxed models. A two-dimensional simulation of an underwater explosion is also presented.

Keywords Multiphase compressible flows · Relaxation processes
Phase transition · Finite volume schemes · Riemann solvers

Mathematics Subject Classification 65M08 · 76T10

T. Flåtten
Norwegian University of Science and Technology, Kolbjørn Hejes vei 2, 7491 Trondheim, Norway
e-mail: toref@math.uio.no

M. Pelanti (✉)
IMSIA, UMR 9219 ENSTA ParisTech - EDF - CNRS - CEA, Université Paris-Saclay, 828 Boulevard des Maréchaux,91762 Palaiseau Cedex, France
e-mail: marica.pelanti@ensta-paristech.fr

K.-M. Shyue
Department of Mathematics, National Taiwan University,Taipei 106, Taiwan
e-mail: shyue@ntu.edu.tw

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_32

1 Introduction

We are interested in the simulation of three-phase flows involving the liquid and vapor phases of one species and a third non-condensable gaseous phase. Applications are, for instance, the simulation of flows around high speed cavitating underwater devices [13] and the modeling of underwater explosions [2, 15]. We describe these multiphase flows by a hyperbolic single-velocity compressible flow model with stiff pressure relaxation, which extends the two-phase formulation that we have considered in previous work [11]. The model includes thermal relaxation terms to account for heat transfer processes between all the phases and chemical relaxation terms to describe mass transfer between the liquid and vapor phases of the species that may undergo transition. Similar multiphase models have been, for instance, presented in [7, 13]. The formulation that we adopt here with phasic total energy equations is particularly convenient to develop a *mixture-energy-consistent* numerical model, in the sense defined in [11] for the two-phase case (see also Sect. 3). The homogeneous hyperbolic portion of the equations is solved numerically via a finite volume wave propagation scheme that uses a simple HLLC-type Riemann solver. Stiff relaxation source terms are handled by efficient numerical procedures that exploit algebraic equilibrium conditions for the relaxed states. One special focus of this work is the study of the effects of heat and mass transfer on the speed of wave propagation. We first derive analytical expressions of the speed of sound of the relaxed multiphase models associated with the different levels of activation of infinitely fast relaxation processes, and we demonstrate that sub-characteristic conditions hold. We then show through a one-dimensional three-phase cavitation tube experiment that the behavior of the wave speed predicted numerically is consistent with our theoretical findings. This paper is organized as follows. In Sect. 2, we present the multiphase flow model under study. Here we also analyze the characteristic speeds of the relaxed models associated with the parent relaxation model. In Sect. 3, we illustrate the numerical method that we have developed to solve the three-phase flow equations. Some numerical experiments are finally presented in Sect. 4, including a two-dimensional simulation of an underwater explosion.

2 Single-Velocity Multiphase Compressible Flow Model

We consider an inviscid compressible flow composed of N phases that we assume in kinematic equilibrium with velocity $\mathbf{u}$. In this work, we are specifically interested in three-phase flows, $N = 3$; nonetheless, we shall present here a general multiphase flow formulation. The volume fraction, density, internal energy per unit volume, and pressure of each phase will be denoted by α_k, ρ_k, $\mathcal{E}_k$, p_k, $k = 1, \ldots N$, respectively. We will denote the total energy for the kth phase with $E_k = \mathcal{E}_k + \rho_k \frac{|\mathbf{u}|^2}{2}$. The saturation condition is $\sum_{k=1}^{N} \alpha_k = 1$. The mixture density is $\rho = \sum_{k=1}^{N} \alpha_k \rho_k$, the mixture internal energy is $\mathcal{E} = \sum_{k=1}^{N} \alpha_k \mathcal{E}_k$, and the mixture total energy is

$E = \sum_{k=1}^{N} \alpha_k E_k = \mathcal{E} + \rho \frac{|\mathbf{u}|^2}{2}$. Mechanical and thermal transfer processes are considered in general for all the phases. We assume that one species in the mixture can undergo phase transition, so that it can exist as a vapor or a liquid phase, and mass transfer terms are accounted for this species only. We will use the subscripts 1 and 2 to denote the liquid and vapor phases of this species. We describe the N-phase flow under consideration by a compressible flow model that extends the six-equation two-phase flow system that we studied in [11]. The model system is composed of the volume fraction equations for $N-1$ phases, the mass and total energy equations for all the N phases, and d mixture momentum equations, where d denotes the spatial dimension:

$$\partial_t \alpha_k + \mathbf{u} \cdot \nabla \alpha_k = \sum_{j=1}^{N} \mathcal{P}_{kj}\,, \quad k = 1, 3, \ldots, N, \tag{1a}$$

$$\partial_t (\alpha_1 \rho_1) + \nabla \cdot (\alpha_1 \rho_1 \mathbf{u}) = \mathcal{M}\,, \tag{1b}$$

$$\partial_t (\alpha_2 \rho_2) + \nabla \cdot (\alpha_2 \rho_2 \mathbf{u}) = -\mathcal{M} \tag{1c}$$

$$\partial_t (\alpha_k \rho_k) + \nabla \cdot (\alpha_k \rho_k \mathbf{u}) = 0\,, \quad k = 3, \ldots, N\,, \tag{1d}$$

$$\partial_t (\rho \mathbf{u}) + \nabla \cdot \left(\rho \mathbf{u} \otimes \mathbf{u} + \left(\sum_{k=1}^{N} \alpha_k p_k \right) \mathbb{I} \right) = 0\,, \tag{1e}$$

$$\partial_t (\alpha_1 E_1) + \nabla \cdot (\alpha_1 (E_1 + p_1) \mathbf{u}) + \Upsilon_1 = -\sum_{j=1}^{N} p_{\mathrm{I}1j} \mathcal{P}_{1j} + \sum_{j=1}^{N} \mathcal{Q}_{1j} + \left(g_{\mathrm{I}} + \frac{|\mathbf{u}|^2}{2} \right) \mathcal{M}\,, \tag{1f}$$

$$\partial_t (\alpha_2 E_2) + \nabla \cdot (\alpha_2 (E_2 + p_2) \mathbf{u}) + \Upsilon_2 = -\sum_{j=1}^{N} p_{\mathrm{I}2j} \mathcal{P}_{2j} + \sum_{j=1}^{N} \mathcal{Q}_{2j} - \left(g_{\mathrm{I}} + \frac{|\mathbf{u}|^2}{2} \right) \mathcal{M}\,, \tag{1g}$$

$$\partial_t (\alpha_k E_k) + \nabla \cdot (\alpha_k (E_k + p_k) \mathbf{u}) + \Upsilon_k = -\sum_{j=1}^{N} p_{\mathrm{I}kj} \mathcal{P}_{kj} + \sum_{j=1}^{N} \mathcal{Q}_{kj}\,, \quad k = 3, \ldots, N\,. \tag{1h}$$

The non-conservative terms Υ_k appearing in the phasic total energy Eqs. (1f)–(1h) are given by

$$\Upsilon_k = \mathbf{u} \cdot \left(Y_k \nabla \left(\sum_{j=1}^{N} \alpha_j p_j \right) - \nabla (\alpha_k p_k) \right)\,, \quad k = 1, \ldots, N\,, \tag{1i}$$

where $Y_k = \frac{\alpha_k \rho_k}{\rho}$ denotes the mass fraction of phase k. In the system above, $\mathcal{P}_{kj}$ and $\mathcal{Q}_{kj}$ represent the volume transfer and the heat transfer, respectively, between the phases k and j, $k, j = 1, \ldots N$. The term $\mathcal{M}$ indicates the mass transfer between the liquid and vapor phases indexed with 1 and 2. The transfer terms are defined as relaxation terms:

$$\mathcal{P}_{kj} = \mu_{kj} (p_k - p_j)\,, \quad \mathcal{Q}_{kj} = \vartheta_{kj} (T_j - T_k)\,, \quad \mathcal{M} = \nu (g_2 - g_1)\,, \tag{2}$$

where T_k denotes the phasic temperature, g_k the phasic chemical potential, and where we have introduced the mechanical, thermal, and chemical relaxation parameters $\mu_{kj} = \mu_{jk} \geq 0$, $\vartheta_{kj} = \vartheta_{jk} \geq 0$, and $\nu = \nu_{12} = \nu_{21} \geq 0$, respectively. Note that $\mathcal{P}_{kj} = -\mathcal{P}_{jk}$ and $\mathcal{Q}_{kj} = -\mathcal{Q}_{jk}$. The quantities $p_{\mathrm{I}kj} = p_{\mathrm{I}jk}$ are interface pressures and g_{I} is an interface chemical potential. We shall assume that mechanical equilibrium is reached instantaneously for all the phases, $\mu_{kj} = \mu_{jk} \equiv \mu \to +\infty$; that is, mechanical relaxation processes are infinitely fast. Following [14], we then consider that

thermal and chemical relaxation processes are either inactive, $\vartheta_{kj} = 0$, $\nu = 0$, or they act infinitely fast, $\vartheta_{kj} \to +\infty$, $\nu \to +\infty$. Heat and mass transfer may be activated at selected locations, for instance, at interfaces for a phase pair (k, j), identified by $\min(\alpha_k, \alpha_j) > \epsilon$, where ϵ is a tolerance.

The closure of the system (1) is obtained through the specification of an equation of state (EOS) for each phase $p_k = p_k(\mathcal{E}_k, \rho_k)$, $T_k = T_k(p_k, \rho_k)$. Here in particular we will adopt the widely used stiffened gas (SG) equation of state:

$$p_k(\mathcal{E}_k, \rho_k) = (\gamma_k - 1)\mathcal{E}_k - \gamma_k \varpi_k - (\gamma_k - 1)\eta_k \rho_k \quad \text{and} \quad T_k(p_k, \rho_k) = \frac{p_k + \varpi_k}{\kappa_{vk}\rho_k(\gamma_k - 1)}, \tag{3}$$

where γ_k, ϖ_k, η_k, and κ_{vk} are constant material-dependent parameters. The corresponding expression for the phasic entropy is $s_k = \kappa_{vk} \log(T_k^{\gamma_k}(p_k + \varpi_k)^{-(\gamma_k - 1)}) + \eta_k'$, where $\eta_k' =$ constant, and $g_k = h_k - T_k s_k$. The parameters for the SG EOS for the liquid and vapor phases of the species that may undergo transition are determined by imposing that the theoretical saturation curve defined by $g_1 = g_2$ matches the experimental one for the considered material [6]. The mixture pressure law is determined by the mixture energy relation $\mathcal{E} = \sum_{k=1}^{N} \alpha_k \mathcal{E}_k(p, \rho_k)$, where we have used the mechanical equilibrium conditions $p_k = p$, $\forall k = 1, \ldots, N$ in the phasic energy laws $\mathcal{E}_k(p_k, \rho_k)$.

Since here we will consider relaxation parameters either $= 0$ or $\to \infty$, a specification of the expression for the interface quantities p_{Ikj}, g_I is not needed. Nevertheless, let us remark that the definition of these interface quantities must be consistent with the second law of thermodynamics, which requires a nonnegative entropy production for the mixture. By writing the equation for the mixture entropy and by following the arguments in [3], one can infer the following sufficient consistency conditions: $p_{Ikj} \in [\min(p_k, p_j), \max(p_k, p_j)]$, and $g_I \in [\min(g_1, g_2), \max(g_1, g_2)]$.

The model (1) is hyperbolic, and the associated speed of sound c_f (non-equilibrium or frozen sound speed) is

$$c_f = \sqrt{\sum_{k=1}^{N} Y_k c_k^2}, \tag{4}$$

where c_k is the speed of sound of phase k, which can be expressed as $c_k = \sqrt{\Gamma_k h_k + \chi_k}$, where $h_k = (\mathcal{E}_k + p_k)/\rho_k$ is the specific enthalpy of phase k, $\Gamma_k = (\partial p_k / \partial \mathcal{E}_k)_{\rho_k}$, and $\chi_k = (\partial p_k / \partial \rho_k)_{\mathcal{E}_k}$.

2.1 Hierarchy of Multiphase Relaxed Models and Speed of Sound

In the considered limit of instantaneous mechanical relaxation $\mu_{kj} \equiv \mu \to \infty$, the model system (1) reduces to a hyperbolic single-velocity single-pressure model which is a generalization of the five-equation two-phase flow model of Kapila et

al. [5]. The reduced pressure equilibrium model can be derived by means of asymptotic techniques. Denoting with p the equilibrium pressure, we obtain the following relaxed system, composed of $2N + d$ equations:

$$\partial_t \alpha_1 + \mathbf{u}\cdot\nabla\alpha_1 = K_1 \nabla\cdot\mathbf{u} + \frac{\Gamma_1}{\rho_1 c_1^2}\sum_{j=2}^{N} Q_{1j} - \alpha_1 \frac{\rho c_p^2}{\rho_1 c_1^2}\sum_{\substack{j,i=1\\ i>j}}^{N} Q_{ji}\left(\frac{\Gamma_j}{\rho_j c_j^2} - \frac{\Gamma_i}{\rho_i c_i^2}\right) + \frac{\rho c_p^2}{\rho_1 c_1^2}\left((\Gamma_1(g_1 - h_1) + c_1^2)\sum_{\substack{j=1\\ j\neq k}}^{N}\frac{\alpha_j}{\rho_j c_j^2} + (\Gamma_2(g_1 - h_2) + c_2^2)\frac{\alpha_1}{\rho_2 c_2^2}\right)\mathcal{M}, \tag{5a}$$

$$\partial_t \alpha_k + \mathbf{u}\cdot\nabla\alpha_k = K_k \nabla\cdot\mathbf{u} + \frac{\Gamma_k}{\rho_k c_k^2}\sum_{\substack{j=1\\ j\neq k}}^{N} Q_{kj} - \alpha_k \frac{\rho c_p^2}{\rho_k c_k^2}\sum_{\substack{j,i=1\\ i>j}}^{N} Q_{ji}\left(\frac{\Gamma_j}{\rho_j c_j^2} - \frac{\Gamma_i}{\rho_i c_i^2}\right) + \rho c_p^2 \frac{\alpha_k}{\rho_k c_k^2}\left(\frac{\Gamma_2(g_1 - h_2) + c_2^2}{\rho_2 c_2^2} - \frac{\Gamma_1(g_1 - h_1) + c_1^2}{\rho_1 c_1^2}\right)\mathcal{M}, \quad k = 3, \ldots, N, \tag{5b}$$

$$\partial_t(\alpha_1\rho_1) + \nabla\cdot(\alpha_1\rho_1\mathbf{u}) = \mathcal{M}, \tag{5c}$$

$$\partial_t(\alpha_2\rho_2) + \nabla\cdot(\alpha_2\rho_2\mathbf{u}) = -\mathcal{M}, \tag{5d}$$

$$\partial_t(\alpha_k\rho_k) + \nabla\cdot(\alpha_k\rho_k\mathbf{u}) = 0, \quad k = 3, \ldots, N, \tag{5e}$$

$$\partial_t(\rho\mathbf{u}) + \nabla\cdot(\rho\mathbf{u}\otimes\mathbf{u} + p\mathbb{I}) = 0, \tag{5f}$$

$$\partial_t E + \nabla\cdot((E + p)\mathbf{u}) = 0, \tag{5g}$$

where

$$K_k = \rho c_p^2 \alpha_k \sum_{\substack{j=1\\ j\neq k}}^{N} \alpha_j \left(\frac{1}{\rho_k c_k^2} - \frac{1}{\rho_j c_j^2}\right) = \alpha_k\left(\frac{\rho c_p^2}{\rho_k c_k^2} - 1\right). \tag{6}$$

In the relations above, we have introduced the pressure equilibrium speed of sound c_p (a generalization of Wood's sound speed), defined by

$$c_p = \left(\rho\sum_{k=1}^{N}\frac{\alpha_k}{\rho_k c_k^2}\right)^{-\frac{1}{2}}. \tag{7}$$

Let us note that the source terms in the volume fraction Eqs. (5a), (5b) result from the asymptotic limit of instantaneous pressure relaxation.

More generally, a hierarchy of hyperbolic multiphase flow models can be established based on the assumptions on equilibria attained by different combinations of instantaneous relaxation processes. In particular, we study here the expression of the speed of sound for the relaxed models in the hierarchy, similar to [3, 4]. We can derive the following results, valid for any equation of state, whose full demonstration will be detailed elsewhere, together with the derivation of (5). First, assuming instantaneous mechanical equilibrium $\mu_{jk} \equiv \mu \to +\infty$ for all the phases and thermal equilibrium $\vartheta_{kj} \equiv \vartheta \to +\infty$ for M phases, $2 \leq M \leq N$, we obtain a hyperbolic relaxed system of $2N - M + 1 + d$ equations characterized by the speed of sound $c_{pT,M}$, defined by

$$\frac{1}{{c_{pT,M}}^2} = \frac{1}{{c_p}^2} + \frac{\rho T}{\sum_{k=1}^{M} C_{pk}}\sum_{k=1}^{M-1} C_{pk}\sum_{j=k+1}^{M} C_{pj}\left(\frac{\Gamma_j}{\rho_j c_j^2} - \frac{\Gamma_k}{\rho_k c_k^2}\right)^2, \tag{8}$$

where T denotes the equilibrium temperature, $C_{pk} = \alpha_k \rho_k \kappa_{pk}$, $\kappa_{pk} = (\partial h_k / \partial T_k)_{p_k}$ (specific heat at constant pressure), and we recall $\Gamma_k = (\partial p_k / \partial \mathcal{E}_k)_{\rho_k}$. If additionally we assume instantaneous chemical relaxation between the liquid and vapor phases 1 and 2, $\nu \to +\infty$, we obtain a hyperbolic relaxed system of $2(N - M + 1) + d$ equations characterized by a speed of sound $c_{pTg,M}$, defined by

$$\frac{1}{{c_{pTg,M}}^2} = \frac{1}{{c_{pT,M}}^2} + \frac{\rho T}{\sum_{k=1}^{M} C_{pk}} \left(\sum_{k=1}^{M} \frac{\Gamma_k C_{pk}}{\rho_k c_k^2} - \frac{1}{T} \left(\frac{\mathrm{d}T}{\mathrm{d}p} \right)_{\mathrm{sat}} \sum_{k=1}^{M} C_{pk} \right)^2 , \tag{9}$$

where we have introduced the derivatives $(\mathrm{d}T/\mathrm{d}p)_{\mathrm{sat}}$ evaluated on the liquid–vapor saturation curve. Analogously to the two-phase case [3], it is easy to observe that subcharacteristic conditions hold; namely, the speed of sound of the N-phase mixture is reduced whenever an additional equilibrium assumption is introduced: $c_{pTg} \equiv c_{pTg,N} \leq c_{pTg,M}$, $c_{pT} \equiv c_{pT,N} \leq c_{pT,M}$, and $c_{pTg} < c_{pT} < c_p < c_{\mathrm{f}}$.

Remark. In [11], an additional term of the form $\mathcal{M}/\rho_{\mathrm{I}}$ was written in the volume fraction equation of the six-equation two-phase model, with ρ_{I} representing an interface density. Similar to [3], this term is not included in the present multiphase model (1). The purpose of the term $\mathcal{M}/\rho_{\mathrm{I}}$ in [11] was to indicate the influence of the mass transfer process on the evolution of the volume fraction. Nonetheless, the rigorous derivation of the pressure-relaxed model (5) from the system (1) reveals that indeed mass transfer terms affect α_k via the pressure relaxation process, as we observe from the contribution of $\mathcal{M}$ appearing in (5a), (5b). Note that neglecting the term $\mathcal{M}/\rho_{\mathrm{I}}$ in the six-equation model of [11] does not affect the numerical model and the numerical results presented there, since $\nu = 0$ or $\nu \to \infty$, and the numerical procedure for treating instantaneous chemical relaxation consists in imposing directly algebraic thermodynamic equilibrium conditions.

3 Numerical Method

We focus now on the numerical approximation of the multiphase system (1), which we can write in compact vectorial form as

$$\partial_t q + \nabla \cdot \mathcal{F}(q) + \varsigma(q, \nabla q) = \psi_\mu(q) + \psi_\vartheta(q) + \psi_\nu(q) , \tag{10}$$

where $q = [\alpha_1, \alpha_3, \ldots, \alpha_N, \alpha_1\rho_1, \ldots, \alpha_N\rho_N, \rho\mathbf{u}, \alpha_1 E_1, \ldots, \alpha_N E_N]^{\mathrm{T}} \in \mathbb{R}^{3N-1+d}$ is the vector of the unknowns, $\mathcal{F}(q)$ represents the conservative portion of the system, and $\varsigma(q, \nabla q)$ is the non-conservative term. The source terms ψ_μ, ψ_ϑ, and ψ_ν in the system above contain mechanical, thermal, and chemical relaxation terms, respectively. To numerically solve the system (10), we use the same techniques that we have developed for the two-phase model in [11]. A fractional step method is employed, where we alternate between the solution of the homogeneous system

$\partial_t q + \nabla \cdot \mathcal{F}(q) + \varsigma(q, \nabla q) = 0$ and the solution of a sequence of systems of ordinary differential equations (ODEs) that take into account the relaxation source terms ψ_μ, ψ_ϑ, and ψ_v. As in [11], the resulting method is mixture-energy-consistent, in the sense that (i) it guarantees conservation at the discrete level of the mixture total energy; (ii) it guarantees consistency by construction of the values of the relaxed states with the mixture pressure law. The method has been implemented by using the libraries of the CLAWPACK software [10].

3.1 Solution of the Homogeneous System

To solve the hyperbolic homogeneous portion of (10), we employ the wave propagation algorithms of [8, 9], which are a class of Godunov-type finite volume methods to approximate hyperbolic systems of partial differential equations. We shall consider here for simplicity the one-dimensional case in the x direction, and we refer the reader to [9] for a comprehensive presentation of these numerical schemes. We assume a grid with cells of uniform size Δx, and we denote with Q_i^n the approximate solution of the system at the ith cell and at time t^n, $i \in \mathbb{Z}$, $n \in \mathbb{N}$. The second-order wave propagation algorithm has the form

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}(\mathcal{A}^+ \Delta Q_{i-1/2} + \mathcal{A}^- \Delta Q_{i+1/2}) - \frac{\Delta t}{\Delta x}(\tilde{F}_{i+1/2} - \tilde{F}_{i-1/2})\,. \quad (11)$$

Here $\mathcal{A}^\mp \Delta Q_{i+1/2}$ are the so-called fluctuations arising from Riemann problems at cell interfaces $(i + 1/2)$ between cells i and $(i + 1)$, and $\tilde{F}_{i+1/2}$ are correction terms for (formal) second-order accuracy. To define the fluctuations, a Riemann solver must be provided. For the present work, we have developed a numerical scheme in one and two spatial dimensions for the three-phase case, $N = 3$, by adopting a HLLC-type Riemann solver analogous to the one that we have presented in [11] for the two-phase case. This solver guarantees conservation of the partial densities $\alpha_k \rho_k$, the mixture momentum $\rho \mathbf{u}$, and the mixture total energy $E = \sum_{k=1}^N \alpha_k E_k$. This simple HLLC-type solver omits the discretization of the non-conservative terms Υ_k in the phasic energy equations. We refer to [11] for a discussion on this point and the rationale for this approach. We just remark here that for the two-phase case we have done comparisons of this HLLC-type solver with Riemann solvers that take into account the non-conservative terms Υ_k , including a Roe-type solver [11, 12] and a new Suliciu-type solver [1], and no relevant differences were observed in the results. Details on the Suliciu-type solver will be reported elsewhere.

3.2 Relaxation Steps

Similar to [7, 11], the numerical relaxation procedures to handle infinitely fast transfer processes are based on the idea of imposing directly equilibrium conditions to obtain a simple system of algebraic equations to be solved in each relaxation sub-step.

3.2.1 Mechanical Relaxation

We consider the solution of the system $\partial_t q = \psi_\mu(q)$ in the limit $\mu_{kj} \equiv \mu \to \infty$. We denote with superscript 0 the quantities at initial time, which come from the solution of the homogeneous system, and with superscript $*$ the quantities at final time, which are the quantities at mechanical equilibrium. First, we easily see that the exact solution of the system of ODEs gives $(\alpha_k \rho_k)^* = (\alpha_k \rho_k)^0$, $k = 1, \ldots, N$, and $(\rho \mathbf{u})^* = (\rho \mathbf{u})^0$, $E^* = E^0$, hence $\mathbf{u}^* = \mathbf{u}^0$ and $\mathcal{E}^* = \mathcal{E}^0$. We then integrate the equations for the phasic total energies by approximating the interface pressures p_{Ikj} with their values at equilibrium $p^*_{Ikj} = p^*$. This gives N equations of the form

$$(\alpha_k E_k)^* - (\alpha_k E_k)^0 = (\alpha_k \mathcal{E}_k)^* - (\alpha_k \mathcal{E}_k)^0 = -p^*(\alpha_k^* - \alpha_k^0), \quad k = 1, 2, \ldots, N\,. \tag{12}$$

Imposing the pressure equilibrium conditions $p_k = p^*, \forall k = 1, \ldots, N$, at final time the phasic internal energies are then expressed as $\mathcal{E}_k^* = \mathcal{E}_k(p^*, (\alpha_k \rho_k)^0 / \alpha_k^*)$. With these relations, system (12) and the constraint $\sum_{k=1}^N \alpha_k = 1$ give $N + 1$ equations for the unknowns α_k^*, $k = 1, \ldots, N$, and p^*. For the particular case of the SG EOS, the problem can be reduced to the solution of a polynomial equation of degree N for the equilibrium pressure p^*. Furthermore, for the case studied here with three phases, $N = 3$, and two gaseous phases governed by a SG EOS with $\varpi_k = 0$ (see Eq. (3)), the polynomial equation of degree 3 for p^* reduces to a quadratic equation, whose physically admissible solution is easily found.

3.2.2 Thermal Relaxation

If thermal relaxation terms are also activated, then we consider the solution of a system of the form $\partial_t q = \psi_\mu(q) + \psi_\vartheta(q)$, with $\mu_{kj} \equiv \mu \to \infty$ for all phase pairs, and $\vartheta_{kj} \equiv \vartheta \to \infty$ for some desired pairs (k, j). Let us assume instantaneous thermal equilibrium for M phases, $2 \leq M \leq N$, in addition to mechanical equilibrium for all phases. We will denote equilibrium values with the superscript $**$. Then, similar to the case of pressure relaxation, we can write $(\alpha_k \rho_k)^{**} = (\alpha_k \rho_k)^0$, $k = 1, \ldots, N$, $(\rho \mathbf{u})^{**} = (\rho \mathbf{u})^0$, $E^{**} = E^0$, and $\mathcal{E}^{**} = \mathcal{E}^0$. Moreover, we write $N - M$ equations of the form (12) with $(\cdot)^0$ replaced by $(\cdot)^*$ and $(\cdot)^*$ replaced by $(\cdot)^{**}$, the mechanical equilibrium conditions $p_k^{**} = p^{**}, \forall k = 1, \ldots, N$, and the thermal equilibrium conditions $T_k^{**} = T^{**}$ for M phases. All these relations give a system of algebraic equations to be solved for the equilibrium values α_k^{**}, p^{**}. As for the mechanical

relaxation step, the solution of this system of algebraic equations can be reduced to the solution of a polynomial equation of degree N for the pressure p^{**} when the SG EOS is adopted. The problem reduces further to the solution of a quadratic equation for the case $N = 3$ with two gaseous phases governed by SG pressure laws with $\varpi_k = 0$.

3.2.3 Thermo-Chemical Relaxation

If thermo-chemical relaxation is activated for the species that may undergo liquid–vapor transition, then we need to solve a system of ODEs of the form $\partial_t q = \psi_\mu(q) + \psi_\vartheta(q) + \psi_\nu(q)$, with $\mu_{kj} \equiv \mu \to \infty$ for all phase pairs, $\vartheta_{kj} \equiv \vartheta \to \infty$ for some phase pairs (k, j), and $\nu \to +\infty$ for the phase pair $(1, 2)$. Let us assume instantaneous thermal equilibrium for M phases, including at least the phases 1 and 2. We denote the quantities at thermodynamic equilibrium with the superscript $\oplus$. First, we can write $\rho^\oplus = \rho^0$, $(\rho\mathbf{u})^\oplus = (\rho\mathbf{u})^0$, $E^\oplus = E^0$, and $\mathcal{E}^\oplus = \mathcal{E}^0$. Moreover, we write $N - M$ equations of the form (12) with $(\cdot)^0$ replaced by $(\cdot)^{**}$ and $(\cdot)^*$ replaced by $(\cdot)^\oplus$, the mechanical equilibrium conditions $p_k^\oplus = p^\oplus$, $\forall k = 1, \ldots, N$, the thermal equilibrium conditions $T_k^\oplus = T^\oplus$ for M phases, and the chemical equilibrium condition $g_1^\oplus = g_2^\oplus$. This set of algebraic equations can be solved for the values of the equilibrium pressure $p^\oplus$, the equilibrium volume fractions $\alpha_k^\oplus$, and the equilibrium densities $\rho_k^\oplus$. For the case of the SG EOS considered here, we use a solution procedure similar to the two-phase case [11]. First, we reduce the set of algebraic conditions excluding the chemical equilibrium relation to the solution of a quadratic equation for the temperature as a function of the equilibrium pressure, $T^\oplus = T^\oplus(p^\oplus)$. Then, the expression of $T^\oplus(p^\oplus)$ is introduced into the equilibrium condition $g_1^\oplus = g_2^\oplus$. This gives an equation for $p^\oplus$, which is solved by Newton's iterative method.

4 Numerical Experiments

We now present some numerical experiments for three-phase flows involving the liquid and vapor phases of water and a third non-condensable phase. The parameters of the SG EOS for water are those used in [11] (we use hereafter the subscripts l and v for liquid and vapor, respectively): $\gamma_\mathrm{l} = 2.35$, $\gamma_\mathrm{v} = 1.43$, $\eta_\mathrm{l} = -1167 \times 10^3$ J/kg, $\eta_\mathrm{v} = 2030 \times 10^3$ J/kg, $\varpi_\mathrm{l} = 10^9$ Pa, $\varpi_\mathrm{v} = 0$ Pa, $\kappa_{v\mathrm{l}} = 1816$ J/(Kg · K), $\kappa_{v\mathrm{v}} = 1040$ J/(Kg · K), $\eta'_\mathrm{l} = 0$ J/(Kg · K), $\eta'_\mathrm{v} = -23.4 \times 10^3$ J/(Kg · K).

4.1 Three-Phase Water Cavitation Tube

We perform a test that is similar to the two-phase cavitation tube experiment presented in [11, 14]. We consider a tube filled initially with liquid water with a uni-

formly distributed small amount of water vapor $\alpha_{wv} = 10^{-2}$ and a small amount of air (non-condensable gas) $\alpha_g = 10^{-1}$. Air is modeled as an ideal gas with $\gamma_g = 1.4$ ($\eta_g = 0$ J/kg, $\varpi_g = 0$ Pa). The initial pressure is $p_0 = 10^5$ Pa, and the initial densities correspond to the temperature $T_0 = 354$ K. A velocity discontinuity is set at the initial time at the middle of the tube, with $u_0 = -20$ m/s on the left and $u_0 = 20$ m/s on the right. We use 3000 grid cells over the interval $[0, 1]$, and CFL = 0.5. We perform the simulation with different levels of activation of instantaneous relaxation processes: (i) only mechanical relaxation (p-relaxation); (ii) mechanical relaxation for all the three phases and thermal relaxation for the liquid–vapor pair only ($pT(\mathrm{lv})$-relaxation); (iii) mechanical and thermal relaxation for all the phases (pT-relaxation); (iv) mechanical relaxation for all the phases and thermal and chemical relaxation for the liquid-vapor pair ($pT(\mathrm{lv})g$-relaxation); (v) mechanical and thermal relaxation for all the phases and chemical relaxation for the liquid-vapor pair (pTg-relaxation). Second-order results are displayed in Fig. 1 for the pressure, the velocity, the total gaseous volume fraction $\alpha_{wv} + \alpha_g$, and the vapor mass fraction. In all the cases, we observe two rarefactions propagating in opposite directions that produce a pressure decrease in the middle region of the tube, and, correspondingly, an increase of the total gaseous component. For the cases with activation of mass transfer, i.e., $pT(\mathrm{lv})g$- and pTg-relaxation, two evaporation waves develop, causing an increase of the vapor mass fraction in the middle region. Note that in these cases the pressure decreases in the cavitation zone until the saturation value is reached, whereas the pressure reaches much lower values here if mass transfer is not activated. By inspecting the results, we observe that the speed of the leading edges of the two rarefactions decreases for any additional instantaneous thermal equilibrium process that we activate in the computation, consistently with the sub-characteristic property demonstrated theoretically for the hierarchy of relaxed models in Sect. 2.1. Let us note that chemical relaxation is not active here around the rarefaction fronts since mass transfer in this test is activated under the metastability condition $T_{\mathrm{liquid}} > T_{\mathrm{sat}}(p)$.

4.2 Underwater Explosion Close to a Rigid Surface

In this test, we simulate a cylindrical underwater explosion (UNDEX) close to a rigid surface. Following [15], we consider an initial bubble of highly pressurized gas (combustion products) surrounded by liquid water and located near an upper flat wall. Three fluid components are involved in this problem: liquid water, water vapor, and combustion gases. The domain is $[-0.6, 0.6] \times [-0.7, 0]\,\mathrm{m}^2$, and the bubble initially is located at $(x_b, y_b) = (0, -0.22)$ m, and it has radius $r_b = 0.05$ m. Inside the bubble, we set initially a pressure $p = 8290 \times 10^5$ Pa, a gas density $\rho_g = 1400\,\mathrm{kg/m}^3$, and volume fractions $\alpha_{wl} = \alpha_{wv} = 10^{-8}$ for the water phases. Outside the bubble, we set $p = 10^5$ Pa, $T = 303$ K, and the volume fractions $\alpha_{wv} = 10^{-4}$ and $\alpha_g = 10^{-7}$, for water vapor and gas, respectively. An ideal gas law is used for the combustion gases, with $\gamma_g = 2$. In this test, thermal and chemical relaxation are activated for the liquid–vapor water pair only. This explosion problem is characterized by a complex

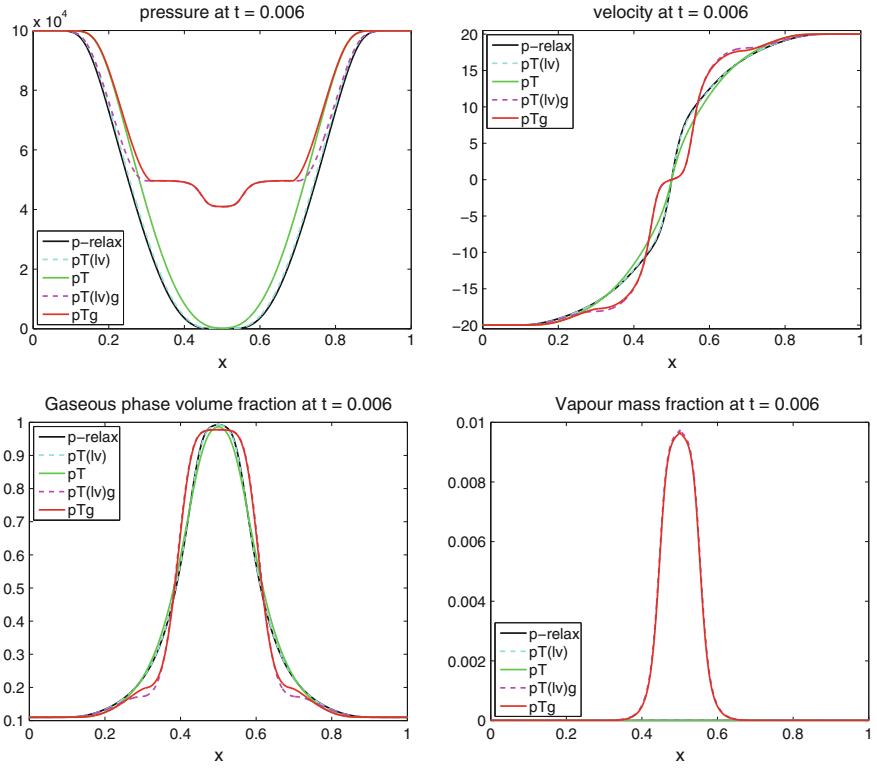


Fig. 1 Numerical results for the pressure, velocity, total gas volume fraction, and vapor mass fraction for the water cavitation tube test

pattern of shocks and rarefaction waves [15], and the likely occurrence of creation and collapse of vapor cavities in the liquid region close to the wall, due to the strong rarefactions and subsequent recompression. We show in Fig. 2 pseudo-color plots of the pressure at two different times. At $t = 0.2$ ms (upper left plot), the circular shock created by the explosion has reflected from the wall; at time $t = 0.35$ ms (lower left plot), a low-pressure cavitation region has developed close to the surface. The pressure and water vapor mass fraction histories in time at the point $(0, 0)$ at the center of the wall are also displayed in the two plots on the right of Fig. 2. We clearly observe the pressure peak corresponding to the instant at which the circular shock hits the wall, the drop of the pressure and consequent growth of a vapor region in this zone, which eventually disappears due to the recompression at later times. In the literature, these type of UNDEX problems are typically simulated by simpler single-fluid models [15], or by two-phase flow models [2] that are only able to describe mechanical cavitation processes, that is growth/collapse of gas cavities due to pressure variations, with no liquid-vapor transition. In contrast, our three-phase

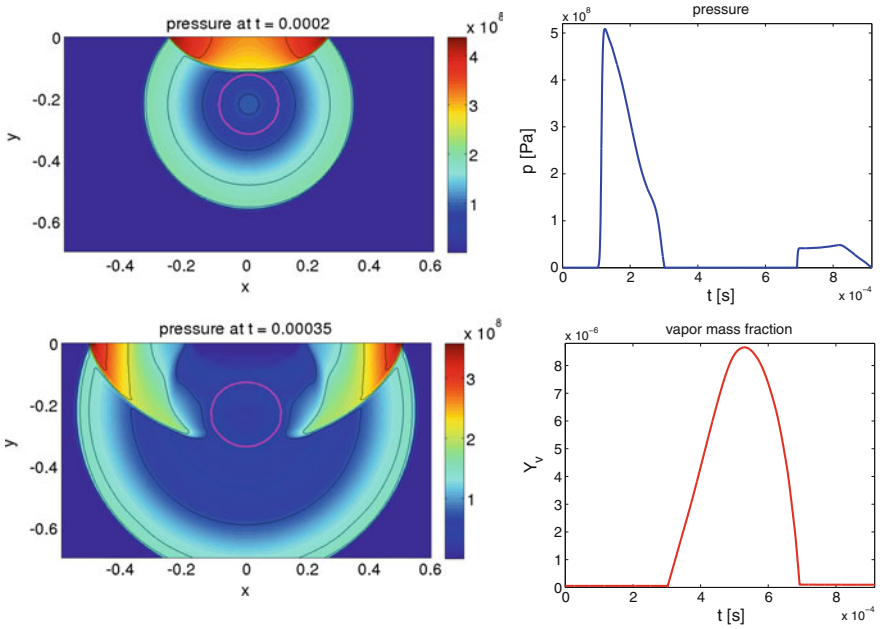


Fig. 2 Numerical results for the UNDEX experiment. Left: pressure field at time $t = 0.2\,\text{ms}$ (top) and $t = 0.35\,\text{ms}$ (bottom). The thick solid circle line indicates the water/bubble interface. Right: pressure history (top) and vapor mass fraction history (bottom) at the point $(0, 0)$ at the center of the wall

flow model allows a more accurate description of the thermodynamics of cavitation processes, which involve liquid–vapor phase change.

5 Conclusions

We have presented a numerical model for multiphase compressible flows involving the liquid and vapor phases of one species and a third inert gaseous phase. The model includes mechanical, thermal, and chemical relaxation processes. The multiphase equations are solved by a mixture-energy-consistent finite volume wave propagation method combined with simple and robust procedures for the stiff relaxation terms. Numerical results show the efficiency of the presented method in modeling complex wave patterns with thermal and mass transfer processes. An analytical study of the characteristic speeds of the hierarchy of relaxed models associated with the parent relaxation model has been also presented. The presented model is an extension of the two-phase flow model that we have introduced in [11]. This novel extension allows us the simulation of problems where the dynamical appearance of vapor cavities and evaporation fronts in a liquid is coupled to the dynamics of a third non-condensable

gaseous component governed by its own equation of state. An example of application illustrated in the present work is the simulation of an underwater explosion close to a rigid wall, where highly pressurized combustion gases (non-condensable phase) trigger cavitation processes in a liquid. Another application example can be found in [13].

Acknowledgements The authors gratefully acknowledge support from the French DGA Grant N. 2012.60.0011.00.470.75.01 for M. Pelanti, the Norwegian RCN Grant N. 234126/E30 for M. Pelanti and T. Flåtten, the Taiwanese NSC Grant N. 99-2115-M-002-005-MY2 for K.-M. Shyue.

References

1. F. Bouchut, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws, and Well-Balanced Schemes for Sources* (Birkhäuser, 2004)
2. A. Daramizadeh, M.R. Ansari, Numerical simulation of underwater explosion near air-water free surface using a five-equation reduced model. Ocean Eng. **110**, 25–35 (2015)
3. T. Flåtten, H. Lund, Relaxation two-phase models and the subcharacteristic condition. Math. Models Methods Appl. Sci. **21**, 2379–2407 (2011)
4. T. Flåtten, A. Morin, S.T. Munkejord, Wave propagation in multicomponent flow models. SIAM J. Appl. Math. **8**, 2861–2882 (2010)
5. A.K. Kapila, R. Menikoff, J.B. Bdzil, S.F. Son, D.S. Stewart, Two-phase modeling of deflagration-to-detonation transition in granular materials: reduced equations. Phys. Fluids **13**, 3002–3024 (2001)
6. O. Le Métayer, J. Massoni, R. Saurel, Elaborating equations of state of a liquid and its vapor for two-phase flow models. Int. J. Therm. Sci. **43**, 265–276 (2004)
7. O. Le Métayer, J. Massoni, R. Saurel, Dynamic relaxation processes in compressible multiphase flows. Application to evaporation phenomena. ESAIM Proc. **40**, 103–123 (2013)
8. R.J. LeVeque, Wave propagation algorithms for multi-dimensional hyperbolic systems. J. Comput. Phys. **131**, 327–353 (1997)
9. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems* (Cambridge University Press, Cambridge, 2002)
10. LeVeque, R.J.: clawpack. http://www.clawpack.org
11. M. Pelanti, K.-M. Shyue, A mixture-energy-consistent six-equation two-phase numerical model for fluids with interfaces, cavitation and evaporation waves. J. Comput. Phys. **259**, 331–357 (2014)
12. M. Pelanti, K.-M. Shyue, A mixture-energy-consistent numerical approximation of a two-phase flow model for fluids with interfaces and cavitation, in *Hyperbolic Problems: Theory, Numerics, Applications, Proceedings of 14th International Conference on Hyperbolic Problems, AIMS* (2014), pp. 839–846
13. F. Petitpas, J. Massoni, R. Saurel, E. Lapebie, L. Munier, Diffuse interface models for high speed cavitating underwater systems. Int. J. Multiphase Flows **35**, 747–759 (2009)
14. R. Saurel, F. Petitpas, R. Abgrall, Modelling phase transition in metastable liquids: application to cavitating and flashing flows. J. Fluid Mech. **607**, 313–350 (2008)
15. W.F. Xie, T.G. Liu, B.C. Khoo, Application of a one-fluid model for large scale homogeneous unsteady cavitation: the modified Schmidt model. Comput. Fluids **35**, 1177–1192 (2006)

Feedback Stabilization of a Linear Fluid–Membrane System with Time Delay

Gilbert Peralta

Abstract A coupled parabolic–hyperbolic system of partial differential equations modelling the interaction of a fluid and a membrane is considered. The model is reformulated as an abstract Cauchy problem and thereby constructing a semigroup for the evolution. This is done by eliminating the pressure. The system is stabilized through a feedback force applied to the membrane incorporating a time delay. The spectral properties and stability are considered under suitable conditions on the fluid viscosity, damping coefficient and delay coefficient.

Keywords Fluid-Membrane system · Stability · Feedback law · Delay

1 Introduction

Let us consider a sufficiently smooth bounded domain Ω in two- or three-dimensional space. Denote by Γ the boundary of the fluid domain Ω and $\Gamma = \overline{\Gamma}_0 \cup \overline{\Gamma}_1$ where Γ_0 and Γ_1 are nonempty open subsets of Γ, both with positive surface measure. On the boundary Γ_1 we have a solid wall while on Γ_0 we have a membrane. Let Σ_0 be the boundary of Γ_0. A linear model describing the above situation is given by the following coupled Navier–Stokes-wave system

$$\begin{cases} u_t - \mu \Delta u + \nabla p = 0, & \text{in } (0,\infty) \times \Omega, \\ \operatorname{div} u = 0, & \text{in } (0,\infty) \times \Omega, \\ u = 0, & \text{on } (0,\infty) \times \Gamma_1, \\ u = w_t \nu, & \text{on } (0,\infty) \times \Gamma_0, \\ w_{tt} - \Delta w = p - \mu \nu \cdot \partial_\nu u + F, & \text{in } (0,\infty) \times \Gamma_0, \\ w = 0, & \text{in } (0,\infty) \times \Sigma_0, \\ \int_{\Gamma_s} w \, \mathrm{d}x = 0, & \text{in } (0,\infty), \end{cases} \tag{1}$$

G. Peralta (✉)
Department of Mathematics and Computer Science, University of the Philippines Baguio, Governor Pack Road, Baguio 2600, Philippines
e-mail: grperalta@up.edu.ph

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_33

supplied with the initial conditions $u(0,x) = u_0(x)$ in Ω and $w(0,x) = w_0(x)$, $w_t(0,x) = w_1(x)$ in Γ_0. In (1), u and p are the velocity field and pressure for the fluid, w is the transversal displacement of the membrane, and F is the feedback force. Moreover, μ is the fluid viscosity and ν is the unit normal outward to Ω. In this paper, we consider small but rapid oscillations, under which we can assume that the domain occupied by the membrane is fixed. Unlike in the Navier–Stokes equation with no-slip boundary condition where the pressure is determined up to a constant, the pressure in (1) is unique due to the Neumann-type boundary condition on Γ_0. In fact, for sufficiently smooth solutions, p satisfies an elliptic problem with mixed Neumann–Robin boundary conditions.

Without the feedback force F, the above system is stable due to the diffusion of the fluid component. This dissipative mechanism, through the interface boundary condition, produces a dissipation for the membrane component. On the other hand, to stabilize the system faster one could add a dissipative mechanism for the membrane by introducing a feedback force. One of the common interior feedback force for the wave equation is using the velocity. This feedback may not be felt instantaneously by the evolution; that is, delay may take place. This consideration gives us the following form of the linear feedback law

$$F(t,x) = -\alpha w_t(t,x) - \beta w_t(t-\tau,x),$$

for $t > 0$ and $x \in \Gamma_0$, where $\alpha, \beta \geq 0$. The constant $\tau > 0$ represents the extent of delay, while the constants α and β represent the strengths of damping and delay, respectively. To have a well-posed system, one must incorporate an initial history

$$w(\theta,x) = z_0(x) \text{ in } (-\tau,0) \times \Gamma_0.$$

It is well known that delay induces a transport phenomenon in the system creating oscillations that may lead into instability; see [10] and the references therein. In the absence of delay, models similar to (1) where the membrane is replaced by a plate has been studied in [4, 5, 9]. Typically, if the damping factor dominates the delay factor, then the system will be stable, i.e. as if delay is not present, however, with a possible slower decay rate. If such terms are equal then the system may not be stable; see [10]. If there are other dissipative mechanisms in the system then we may obtain stability under appropriate conditions, for instance, viscoelasticity in wave equations in [7] and fluid viscosity in a fluid–structure system in [11]. In this paper, we shall also see that viscosity plays a role in deriving sufficient conditions for exponential stability.

The plan of this paper is as follows. In Sect. 2, we introduce generalized trace results that are needed in the elimination of the pressure in the semigroup formulation. In Sect. 3, we write (1) as an abstract Cauchy problem in a suitable state space and prove that it generates a contraction semigroup under a suitable assumption on α, β and μ. The spectral properties and uniform exponential stability of the semigroup will be discussed in Sects. 4 and 5, respectively.

2 Generalized Traces for Some Graph Spaces

Let $\Sigma \subset \Gamma$ be sufficiently smooth. For $s = m + \sigma$ where m is a nonnegative integer and $\sigma \in (0, 1)$, let $H^s_{00}(\Sigma) = \{w \in H^s(\Sigma) : \|w\|_{s,\Sigma} < \infty\}$ where

$$\|w\|^2_{s,\Sigma} = \|u\|^2_{H^s(\Sigma)} + \sum_{|\alpha|=m} \int_\Sigma \frac{|D^\alpha u(x)|^2}{d(x, \partial\Sigma)^{2\sigma}}\, \mathrm{d}x$$

and $d(x, \partial\Sigma)$ denotes the distance of x from $\partial\Sigma$. For each nonnegative integer m, $C^\infty_0(\Sigma)$ is dense in $H^{m+1/2}_{00}(\Sigma)$ and we have $(H^{m+1/2}_{00}(\Sigma))' = H^{-m-1/2}(\Sigma)$, see [8] for more details.

Consider the Hilbert space $L^2_{\mathrm{div}}(\Omega) = \{u \in L^2(\Omega) : \operatorname{div} u \in L^2(\Omega)\}$ with the graph norm. Recall that elements of $L^2_{\mathrm{div}}(\Omega)$ admit generalized normal traces $u \cdot \nu|_\Sigma$ and the corresponding mapping is a continuous linear operator from $L^2_{\mathrm{div}}(\Omega)$ into $H^{-1/2}(\Sigma)$. Moreover, for every $\varphi \in H^1(\Omega)$ with trace in $H^{1/2}_{00}(\Sigma)$ we have

$$\langle u \cdot \nu|_\Sigma, \varphi\rangle = \int_\Omega (\operatorname{div} u)\varphi\, \mathrm{d}x + \int_\Omega u \cdot \nabla\varphi\, \mathrm{d}x.$$

For the fluid component we shall use the function spaces

$$H = \{u \in L^2(\Omega) : \operatorname{div} u = 0 \text{ in } \Omega,\ u \cdot \nu|_{\Gamma_1} = 0\},$$
$$V = \{u \in H^1(\Omega) : \operatorname{div} u = 0 \text{ in } \Omega,\ u|_{\Gamma_1} = 0\}.$$

Recall that the trace maps $\gamma_0 : H^1(\Omega) \to H^{1/2}(\Gamma)$ and $\gamma_1 : H^2(\Omega) \to H^{3/2}(\Gamma) \times H^{1/2}(\Gamma)$ defined by $\gamma_0 u = u|_\Gamma$ and $\gamma_1 u = (u|_\Gamma, \partial_\nu u|_\Gamma)$ are surjective bounded linear operators. It follows that the operators $\gamma_0\gamma_0^*$ and $\gamma_1\gamma_1^*$ are strictly positive definite and thus invertible. Given $\varphi \in H^{1/2}_{00}(\Sigma)$, we extend it by zero to the whole boundary Σ to obtain an element in $H^{1/2}(\Gamma)$, which we shall still denote by φ. We do a similar construction for $\psi \in H^{3/2}_{00}(\Sigma)$. Consider the lifting operators $\ell : H^{3/2}_{00}(\Sigma) \times H^{1/2}_{00}(\Sigma) \to H^2(\Omega)$ and $\kappa : H^{1/2}_{00}(\Sigma) \to H^1(\Omega)$ given by

$$\ell(\psi, \varphi) = \gamma_1^*(\gamma_1\gamma_1^*)^{-1}(\psi, \varphi), \qquad \kappa\varphi = \gamma_0^*(\gamma_0\gamma_0^*)^{-1}\varphi.$$

Let ℓ_1 and ℓ_2 be the coordinate functions of ℓ, that is, $\ell_1\psi = \ell(\psi, 0)$ and $\ell_2\varphi = \ell(0, \varphi)$. It follows that ℓ_1, ℓ_2 and κ are bounded linear operators.

Let $\mathscr{D} = \{p \in H^1(\Omega) : \Delta p \in L^2(\Omega)\}$ and $\mathscr{W} = \{\pi \in L^2(\Omega) : \Delta\pi \in H^{-1}(\Omega)\}$ be equipped with the corresponding graph norms. Given $\pi \in \mathscr{W}$ and $p \in \mathscr{D}$, we define $\pi|_\Sigma$ and $\partial_\nu p|_\Sigma$ by

$$\langle \pi|_\Sigma, \varphi\rangle = \int_\Omega \pi\, \Delta(\ell_2\varphi)\, \mathrm{d}x - \langle \Delta\pi, \ell_2\varphi\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}$$
$$\langle \partial_\nu p|_\Sigma, \psi\rangle = \int_\Omega (\Delta p)\ell_1\psi\, \mathrm{d}x + \int_\Omega (\nabla p)\cdot\nabla(\ell_1\psi)\, \mathrm{d}x \tag{2}$$

for every $\varphi \in H_{00}^{1/2}(\Sigma)$ and $\psi \in H_{00}^{3/2}(\Sigma)$. From the definition and the properties of the above extension operators, one can immediately see that $\pi \mapsto \pi|_\Sigma \in \mathscr{L}(\mathscr{W}, H^{-1/2}(\Sigma))$ and $p \mapsto \partial_\nu p|_\Sigma \in \mathscr{L}(\mathscr{D}, H^{-3/2}(\Sigma))$. If $\pi \in H^1(\Omega)$ and $p \in H^2(\Omega)$ then these traces coincide with the usual traces. This remark follows immediately from the above definitions and Green's identities.

Now consider the subspace $\mathscr{Y} = \{\pi \in L^2(\Omega) : \Delta\pi \in L^2(\Omega)\}$ of $\mathscr{W}$ with the associated graph norm. Given $\pi \in \mathscr{Y}$, define $\pi|_\Sigma$ and $\partial_\nu\pi|_\Sigma$ as follows

$$\langle \pi|_\Sigma, \varphi\rangle = \int_\Omega \pi\, \Delta(\ell_2\varphi)\, \mathrm{d}x - \int_\Omega (\Delta\pi)\ell_2\varphi\, \mathrm{d}x$$
$$\langle \partial_\nu\pi|_\Sigma, \psi\rangle = \int_\Omega (\Delta\pi)\ell_1\psi\, \mathrm{d}x - \int_\Omega \pi\, \Delta(\ell_1\psi)\, \mathrm{d}x$$

for every $\varphi \in H_{00}^{1/2}(\Sigma)$ and $\psi \in H_{00}^{3/2}(\Sigma)$. Again these traces are bounded, more precisely, $\pi \mapsto \pi|_\Sigma \in \mathscr{L}(\mathscr{Y}, H^{-1/2}(\Sigma))$ and $\pi \mapsto \partial_\nu\pi|_\Sigma \in \mathscr{L}(\mathscr{Y}, H^{-3/2}(\Sigma))$. Notice that the definition of $\pi|_\Sigma$ is the same whether it is viewed as an element of $\mathscr{W}$ or $\mathscr{Y}$. Likewise, if $p \in H^1(\Omega) \cap \mathscr{Y} \subset \mathscr{D}$ then the definition of $\partial_\nu p|_\Sigma$ coincides with the earlier formulation (2).

Let us consider the graph space $G = \{(u, p) \in V \times L^2(\Omega) : -\mu\Delta u + \nabla p \in L^2_{\mathrm{div}}(\Omega)\}$ endowed with the graph norm. Notice that $\Delta p = \mathrm{div}(\nabla p - \mu\Delta u) \in L^2(\Omega)$ so that $p \in \mathscr{Y}$ and hence it admits traces such that

$$\|p\|_{H^{-1/2}(\Sigma)} + \|\partial_\nu p\|_{H^{-3/2}(\Sigma)} \le C(\|p\|_{L^2(\Omega)} + \|\,\mathrm{div}(\nabla p - \mu\Delta u)\|_{L^2(\Omega)}).$$

For $(u, p) \in G$, we define the following

$$\langle \mu\partial_\nu u|_\Sigma, \varphi\rangle = \langle p\nu|_\Sigma, \varphi\rangle - \mu\int_\Omega \nabla u\cdot\nabla(\kappa\varphi)\, \mathrm{d}x + \int_\Omega p\,\mathrm{div}(\kappa\varphi)\, \mathrm{d}x$$
$$+ \int_\Omega (-\mu\Delta u + \nabla p)\cdot\kappa\varphi\, \mathrm{d}x$$
$$\langle \mu\Delta u\cdot\nu|_\Sigma, \psi\rangle = \langle \partial_\nu p|_\Sigma, \psi\rangle + \int_\Omega (-\mu\Delta u + \nabla p)\cdot\nabla(\ell_1\psi)\, \mathrm{d}x$$

for every $\varphi \in H_{00}^{1/2}(\Sigma)$ and $\psi \in H_{00}^{3/2}(\Sigma)$. Again, one can see immediately that these generalized traces are bounded; that is, $(u, p) \mapsto \partial_\nu u|_\Sigma \in \mathscr{L}(G, H^{-1/2}(\Sigma))$ and $(u, p) \mapsto \Delta u\cdot\nu \in \mathscr{L}(G, H^{-3/2}(\Sigma))$. In fact, we have

$$\|\partial_\nu u\|_{H^{-1/2}(\Sigma)} \le C(\|p\|_{H^{-1/2}(\Sigma)} + \|u\|_V + \|p\|_{L^2(\Omega)} + \|\mu\Delta u - \nabla p\|_{L^2(\Omega)})$$
$$\|\Delta u \cdot \nu\|_{H^{-3/2}(\Sigma)} \le C(\|\mu\Delta u - \nabla p\|_H + \|\partial_\nu p\|_{H^{-3/2}(\Sigma)}).$$

From the above discussion note that $-\mu\Delta u + \nabla p$ admits a generalized normal trace on Σ. In the case $\operatorname{div}(-\mu\Delta u + \nabla p) = 0$, it follows from the divergence theorem that

$$(-\mu\Delta u + \nabla p) \cdot \nu|_\Sigma = \mu\Delta u \cdot \nu|_\Sigma - \partial_\nu p|_\Sigma.$$

In particular, we have the following generalized integration by parts formula

$$\int_\Omega (\mu\Delta u - \nabla p) f \, \mathrm{d}x = \langle \mu\partial_\nu u - p\nu|_\Sigma, f\rangle - \mu \int_\Omega \nabla u \cdot \nabla f \, \mathrm{d}x$$

for every $(u, p) \in G$ and $f \in V$. We refer to [2, 13] for similar discussions.

3 Abstract Formulation and Well-Posedness of the System

The coupled system (1) will be expressed as an evolution equation in a suitable state space. Using the divergence theorem, it can be seen that we need to factor the constants in the space for the states associated with the membrane. Let

$$X = \{(u, w, v, z) \in H \times \widehat{H}_0^1(\Gamma_0) \times \widehat{L}^2(\Gamma_0) \times L^2(-\tau, 0; \widehat{L}^2(\Gamma_0)) : u \cdot \nu = v \text{ in } \Gamma_0\},$$

where $\widehat{L}^2(\Gamma_0) = \{w \in L^2(\Gamma_0) : \int_{\Gamma_0} w \, \mathrm{d}x = 0\}$ and $\widehat{H}_0^1(\Gamma_0) = H^1(\Gamma_0) \cap \widehat{L}^2(\Gamma_0)$, be equipped with the norm

$$\|(u, w, v, z)\|_X^2 = \int_\Omega |u|^2 \, \mathrm{d}x + \int_{\Gamma_0} |\nabla w|^2 + |v|^2 \, \mathrm{d}x + \beta \int_{-\tau}^0 \int_{\Gamma_0} |z|^2 \, \mathrm{d}x \, \mathrm{d}\theta.$$

Following [1], we eliminate p in the system by rewriting it as an elliptic problem with boundary data involving the fluid velocity and the displacement of the membrane. Define the mixed Neumann–Robin map $M : H^{-3/2}(\Gamma_1) \times H^{-3/2}(\Gamma_0) \to L^2(\Omega)$ according to

$$\pi = M(\varphi, \psi) \Longleftrightarrow \begin{cases} \Delta\pi = 0 & \text{in } \Omega, \\ \partial_\nu \pi = \varphi & \text{on } \Gamma_1, \\ \partial_\nu \pi + \pi = \psi & \text{on } \Gamma_0. \end{cases}$$

For smooth solutions we can see that p satisfies the boundary value problem

$$\begin{cases} \Delta p = 0 & \text{in } \Omega, \\ \partial_\nu p = \mu \Delta u \cdot \nu & \text{on } \Gamma_1, \\ \partial_\nu p + p = -\Delta w + \alpha v + \beta z(-\tau) + \mu v \cdot \partial_\nu u + \mu \Delta u \cdot \nu & \text{on } \Gamma_0. \end{cases}$$

Hence, we can represent p in terms of the map M as follows

$$p = L(u, w, v, z) := M(\mu \Delta u \cdot \nu, -\Delta w + \alpha v + \beta z(-\tau) + \mu v \cdot \partial_\nu u + \mu \Delta u \cdot \nu).$$

To keep track of the retarded term in (1), let us introduce the delay variable $z(t, \theta, x) = w_t(t + \theta, x)$, which satisfies the following transport equation in $(-\tau, 0)$ with parameter $x \in \Gamma_0$

$$\begin{cases} z_t(t, \theta, x) - z_\theta(t, \theta, x) = 0, & \text{in } (0, \infty) \times (-\tau, 0) \times \Gamma_0, \\ z(t, 0, x) = w_t(t, x), & \text{in } (0, \infty) \times \Gamma_0, \\ z(0, \theta, x) = z_0(\theta, x), & \text{in } (-\tau, 0) \times \Gamma_0. \end{cases} \tag{3}$$

The fluid–membrane system (1) can now be rewritten as an evolution equation in X

$$\frac{d}{dt}(u, w, v, z) = A(u, w, v, z)$$

where $A : D(A) \to X$ is the linear operator defined by

$$A(u, w, v, z) := (\mu \Delta u - \nabla p, v, \Delta w - \alpha v - \beta z(-\tau) + p - \mu v \cdot \partial_\nu u, \partial_\theta z)$$

with domain $D(A)$ consisting of all elements $(u, w, v, z) \in X$ such that $u \in V$, $v \in \widehat{H}^1_0(\Gamma_0)$, $z \in H^1(-\tau, 0; \widehat{L}^2(\Gamma_0))$, $u = v\nu$ on Γ_0, $z_{|\theta=0} = v$ on Γ_0, $\mu \Delta u - \nabla p \in H$ and $\Delta w - \alpha v - \beta z(-\tau) + p - \mu v \cdot \partial_\nu u \in \widehat{L}^2(\Gamma_0)$, where $p = L(u, w, v, z) \in L^2(\Omega)$.

Let C_P be the constant in the following inequality obtained from trace theory and the Poincaré inequality

$$\int_{\Gamma_0} |u \cdot \nu|^2 \, \mathrm{d}x \le C_P \int_\Omega |\nabla u|^2 \, \mathrm{d}x, \qquad \forall\, u \in V. \tag{4}$$

Theorem 1. *If $\alpha + \frac{\mu}{C_P} \ge \beta$, where C_P is the constant in (4), then A is the generator of a strongly continuous semigroup of contractions on X.*

Proof. We apply the Lumer–Phillips Theorem and hence we must show that A is m-dissipative. Given $X_0 = (u, w, v, z) \in D(A)$, by applying generalized Green's identity for the membrane component, divergence theorem to the fluid component and Cauchy–Schwarz inequality we have

$$\begin{aligned}\mathrm{Re}(AX_0, X_0)_X &= -\mu\int_\Omega |\nabla u|^2\,\mathrm{d}x - \left(\alpha - \frac{|\beta|}{2}\right)\int_{\Gamma_0}|v|^2\,\mathrm{d}x - |\beta|\int_{\Gamma_0} z(-\tau)v\,\mathrm{d}x \\ &\quad -\frac{|\beta|}{2}\int_{\Gamma_0}|z(-\tau)|^2\,\mathrm{d}x \;\le\; -\left(\alpha - |\beta| + \frac{\mu}{C_P}\right)\int_{\Gamma_0}|v|^2\,\mathrm{d}x, \quad (5)\end{aligned}$$

establishing the dissipativity of A.

To prove maximality, it is enough to prove that $0 \in \rho(A)$, where $\rho(A)$ denotes the resolvent set of A. In order to show this, we need to find $(u, w, v, z) \in D(A)$ such that $A(u, w, v, z) = (f, g, h, \zeta)$ for a given $(f, g, h, \zeta) \in X$ and $\|(u, w, v, z)\|_X \le C\|(f, g, h, \zeta)\|_X$ for some constant $C > 0$ independent of (u, w, v, z) and (f, g, h, ζ). The equation to solve is equivalent to $v = g$, $z_\theta = \zeta$, $z_{|\theta=0} = v$, the Stokes problem

$$\begin{cases} -\mu\Delta u + \nabla p = -f, & \text{in } \Omega, \\ \operatorname{div} u = 0, & \text{in } \Omega, \\ u = 0, & \text{on } \Gamma_1, \\ u = gv, & \text{on } \Gamma_0, \end{cases} \quad (6)$$

and the elliptic equation with homogeneous Dirichlet condition

$$\begin{cases} -\Delta w = -\alpha g - \beta z(-\tau) + p - \mu v\cdot\partial_v u - h, & \text{in } \Gamma_0, \\ w = 0, & \text{on } \Sigma_0. \end{cases} \quad (7)$$

We can see immediately that the delay variable is given by $z(\theta) = v - \int_\theta^0 \zeta(\vartheta)\,\mathrm{d}\vartheta$ from which we have $z \in H^1(-\tau, 0; \widehat{L}^2(\Gamma_0))$ and

$$\|z\|_{L^2(-\tau,0;L^2(\Gamma_0))} + \|z(-\tau)\|_{L^2(\Gamma_0)} \le C_\tau(\|g\|_{L^2(\Gamma_0)} + \|\zeta\|_{L^2(-\tau,0;L^2(\Gamma_0))}). \quad (8)$$

The Stokes equation (6) admits a solution pair $(u, \tilde{p}) \in V \times L^2(\Omega)$; see, for instance [12]. Given a constant p^*, the pair (u, p) where $p = \tilde{p} + p^*$ is also a solution pair and we have

$$\|u\|_V + \|p\|_{L^2(\Omega)} \le C(\|f\|_H + \|g\|_{H_0^1(\Gamma_0)}),$$

and consequently we have the following trace estimate

$$\|p\|_{H^{-1/2}(\Gamma_0)} + \|\partial_v u\|_{H^{-1/2}(\Gamma_0)} \le C(\|f\|_H + \|g\|_{H_0^1(\Gamma_0)}). \quad (9)$$

Since the right-hand side for the elliptic equation (7) lies in $H^{-1}(\Gamma_0)$, by standard elliptic theory we have a solution $w \in H^1(\Gamma_0)$ and from (8) and (9) it is not hard to see that $\|w\|_{H_0^1(\Gamma_0)} \le C\|(f, g, h, \zeta)\|_X$ for some constant $C > 0$.

The final step is to choose the constant p^* in such a way that w has average zero. Let $\psi \in H_0^1(\Gamma_0)$ be the solution of the Poisson equation $-\Delta\psi = 1$ on Γ_0 with boundary condition $\psi = 0$ on Σ_0. A straightforward calculation yields that $w \in \widehat{L}^2(\Gamma_0)$ if and only if

$$p^* = \|\nabla\psi\|^{-2}_{L^2(\Gamma_0)}\left((\alpha+\beta)\int_{\Gamma_0} v\psi\,\mathrm{d}x - \beta\int_{-\tau}^{0}\int_{\Gamma_0}\zeta(\vartheta)\psi\,\mathrm{d}x\,\mathrm{d}\vartheta - \langle\tilde{p} - \mu\nu\cdot\partial_\nu u, \psi\rangle\right).$$

Finally one can check that $p = L(u, w, v, z)$ and $\|(u, w, v, z)\|_X \leq C\|(f, g, h, \zeta)\|_X$.

4 Spectral Properties

First, let us present the adjoint of the generator A. To describe the said operator, we consider the isomorphism $J : X \to X$

$$J(f, g, h, \zeta(\theta)) = (-f, g, -h, z(-\theta - \tau)).$$

Theorem 2. *The X-adjoint $A^* : D(A^*) \to X$ of the closed operator A is given by*

$$A^*(f, g, h, \zeta) = (\mu\Delta f - \nabla\pi, -h, -\Delta g - \alpha h + \beta\zeta(0) + \pi - \mu\nu\cdot\partial_\nu f, -\partial_\theta\zeta)$$

with domain $D(A^*)$ comprising of all elements $(f, g, h, \zeta) \in X$ such that $f \in V$, $h \in \widehat{H}^1_0(\Gamma_0)$, $\zeta \in H^1(-\tau, 0; \widehat{L}^2(\Gamma_0))$, $f = h\nu$ on Γ_0, $\zeta(-\tau) = -h$ on Γ_0, $\mu\Delta f - \nabla\pi \in H$ and $-\Delta g - \alpha h + \beta\zeta(0) + \pi - \mu\nu\cdot\partial_\nu f \in \widehat{L}^2(\Gamma_0)$ where $\pi = -LJ(f, g, h, \zeta)$.

Proof. The proof is similar to [11, Theorem 2.7] and therefore we omit it here.

In the following, we shall show that the spectrum of A consists of eigenvalues only, except possibly on the negative real axis. This will be done by rewriting the resolvent equation in variational form on a suitable space and then applying the Fredholm alternative and Lax–Milgram Lemma. For this direction, we introduce the following function spaces

$$W_0 = H \times \widehat{L}^2(\Gamma_0), \qquad W_1 = \{(u, v) \in V \times \widehat{H}^1_0(\Gamma_0) : u = v\nu \text{ on } \Gamma_0\}.$$

The embedding $W_1 \subset W_0$ is compact, dense and continuous.

Given a nonzero complex number λ and $Y = (f, g, h, \varphi) \in X$ we define the sesquilinear form $a_\lambda : W_1 \times W_1 \to \mathbb{C}$

$$a_\lambda((u, v), (\phi, \psi)) = \lambda\int_\Omega u\cdot\phi\,\mathrm{d}x + \mu\int_\Omega \nabla u\cdot\nabla\phi\,\mathrm{d}x + q(\lambda)\int_{\Gamma_0} v\psi\,\mathrm{d}x + \frac{1}{\lambda}\int_{\Gamma_0}\nabla v\cdot\nabla\psi\,\mathrm{d}x$$

where $q(\lambda) = \lambda + \alpha + \beta e^{-\lambda\tau}$ and the antilinear form $F_{Y,\lambda} : W_1 \to \mathbb{C}$ by

$$F_{Y,\lambda}(\phi,\psi)=\int_\Omega f\cdot\phi\,\mathrm{d}x-\frac{1}{\lambda}\int_{\Gamma_0}\nabla g\cdot\nabla\psi\,\mathrm{d}x+\int_{\Gamma_0}h\psi\,\mathrm{d}x$$
$$-\beta\int_{-\tau}^{0}\int_{\Gamma_0}e^{-\lambda(\theta+\tau)}\varphi(\theta)\psi\,\mathrm{d}x\,\mathrm{d}\theta.$$

Theorem 3. *Let $\sigma(A)$ and $\sigma_p(A)$ be the spectrum and point spectrum of A, respectively. If $\alpha+\frac{\mu}{C_P}\geq|\beta|$ then $\sigma(A)\cap(\mathbb{C}\setminus(-\infty,0])=\sigma_p(A)$ and $\sigma(A^*)\cap(\mathbb{C}\setminus(-\infty,0])=\sigma_p(A^*)$.*

Proof. Let $\lambda\in\mathbb{C}\setminus(-\infty,0]$. For a given $Y=(f,g,h,\varphi)\in X$, suppose that there exists $(u,w,v,z)\in D(A)$ such that

$$(\lambda I-A)(u,w,v,z)=(f,g,h,\zeta). \tag{10}$$

This equation is equivalent to the condition that $\lambda w-v=g$, $\lambda z-\partial_\theta z=h$, $z(0)=v$, u satisfies the Stokes equation

$$\begin{cases}\lambda u-\mu\Delta u+\nabla p=f, & \text{in } \Omega,\\ \operatorname{div}u=0, & \text{in } \Omega,\\ u=0, & \text{on } \Gamma_1,\\ u=v\nu, & \text{on } \Gamma_0,\end{cases} \tag{11}$$

and (v,w) satisfies the boundary value problem

$$\begin{cases}(\lambda+\alpha)v-\Delta w=-\beta z(-\tau)-p+\mu\nu\cdot\partial_\nu u+h, & \text{in } \Gamma_0,\\ w=0, & \text{on } \Sigma_0.\end{cases} \tag{12}$$

Applying the variation of parameters formula to the equation for z, we obtain

$$z(\theta)=e^{\lambda\theta}v+\int_\theta^0 e^{\lambda(\theta-\vartheta)}\varphi(\vartheta)\,\mathrm{d}\vartheta. \tag{13}$$

Using this and the fact that $w=\frac{1}{\lambda}(v+g)$ we can see that the variational form of the elliptic equation (12) is given by

$$q(\lambda)\int_{\Gamma_0}v\psi\,\mathrm{d}x\;+\frac{1}{\lambda}\int_{\Gamma_0}\nabla v\cdot\nabla\psi\,\mathrm{d}x=-\frac{1}{\lambda}\int_{\Gamma_0}\nabla g\cdot\nabla\psi\,\mathrm{d}x+\int_{\Gamma_0}h\psi\,\mathrm{d}x$$
$$-\beta\int_{-\tau}^{0}\int_{\Gamma_0}e^{-\lambda(\theta+\tau)}\varphi(\theta)\psi\,\mathrm{d}x\,\mathrm{d}\theta-\langle\mu\partial_\nu u-pv,\psi\nu\rangle \tag{14}$$

for every $\psi\in H_0^1(\Gamma_0)$. Also, the weak form of the Stokes equation (11) is given by

$$\lambda \int_\Omega u \cdot \phi \, \mathrm{d}x + \mu \int_\Omega \nabla u \cdot \nabla \phi \, \mathrm{d}x = \langle \mu \partial_\nu u - p\nu, \phi \rangle + \int_\Omega f \cdot \phi \, \mathrm{d}x \tag{15}$$

for every $\phi \in V$. Therefore if $(\phi, \psi) \in W_1$, taking the sum of (14) and (15) so that the duality pairing vanishes, we obtain the variational equation

$$a_\lambda((u, v), (\phi, \psi)) = F_{Y,\lambda}(\phi, \psi), \qquad \forall\, (\phi, \psi) \in W_1. \tag{16}$$

Conversely suppose that the variational Eq. (16) is satisfied. Define z and w as above. Choosing $\psi = 0$ we can see that u satisfies the Stokes equation (11) in the sense of distributions. Using Green's identity the elliptic equation (12) holds in the distributional sense as well. We choose $p = \tilde{p} + p^*$ where

$$p^* = \langle \mu \nu \partial_\nu u - \tilde{p} - \beta z(-\tau) - (\lambda + \alpha)v, \psi_0 \rangle$$

and $\{\psi_0\}$ is a basis of $\{\psi \in H_0^1(\Gamma_0) : \Delta u \text{ is constant}\}$, which has dimension 1, and is the orthogonal complement of $\widehat{H}_0^1(\Gamma_0)$ in $H_0^1(\Gamma_0)$. Split the sesquilinear form a_λ as $a_\lambda = a_{0,\lambda} + a_{1,\lambda}$ where the sesquilinear forms $a_{i,\lambda} : W_i \times W_i \to \mathbb{C}$ for $i = 0, 1$ are given by

$$a_{1,\lambda}((u, v), (\phi, \psi)) = \mu \int_\Omega \nabla u \cdot \nabla \phi \, \mathrm{d}x + \frac{1}{\lambda} \int_{\Gamma_0} \nabla v \cdot \nabla \psi \, \mathrm{d}x$$
$$a_{0,\lambda}((u, v), (\phi, \psi)) = \lambda \int_\Omega u \cdot \phi \, \mathrm{d}x + q(\lambda) \int_{\Gamma_0} v\psi \, \mathrm{d}x.$$

The form $a_{1,\lambda}$ is W_1-coercive provided that $\operatorname{Im} \lambda \neq 0$ and $a_{0,\lambda}$ is bounded. From the Lax–Milgram–Fredholm Lemma (see [6]) we obtain the desired result. The corresponding result for the adjoint can be done in a similar way.

5 Uniform Exponential Stability

In this section we prove that the energy of the solutions for the fluid–membrane interaction model decays to zero exponentially under the condition $\alpha + \frac{\mu}{C_P} > \beta$. The result will be shown using the Lyapunov method. The success of this method to the system (1) relies on the following theorem in [5].

Theorem 4. *Let S be the Stokes map defined in the following way*

$$u = Sv \iff \begin{cases} -\mu \Delta u + \nabla p = 0, & \text{in } \Omega, \\ \operatorname{div} u = 0, & \text{in } \Omega, \\ u = 0, & \text{on } \Gamma_1, \\ u = v\nu, & \text{on } \Gamma_0. \end{cases}$$

Then it holds that $S \in \mathscr{L}(\widehat{L}^2(\Gamma_0), H^{1/2}(\Omega) \cap H) \cap \mathscr{L}(\widehat{H}^1_0(\Gamma_0), H^{3/2}(\Omega) \cap H)$.

Theorem 5. *Suppose that* $\alpha + \frac{\mu}{C_P} > \beta$. *The semigroup generated by A is uniformly exponentially stable; that is, there exist* $\sigma > 0$ *and* $M \geq 1$ *such that* $\|e^{tA}X_0\|_X \leq Me^{-\sigma t}\|X_0\|_X$ *for every* $X_0 \in X$ *and* $t \geq 0$.

Proof. By a standard density argument it is enough to consider initial data in the domain of A. For this purpose, let $Y(t) = (u(t), v(t), w(t), z(t)) = e^{tA}(u_0, w_0, v_0, z_0)$ where $(u_0, w_0, v_0, z_0) \in D(A)$. We define the Lyapunov functional L as follows

$$L(t) = \frac{1}{2}\|(u(t), v(t), w(t), z(t))\|_X^2 + \varepsilon_1 \int_{-\tau}^{0}\int_{\Gamma_0} e^{a\theta}|z(t,\theta)|^2 \,\mathrm{d}x\,\mathrm{d}\theta + \varepsilon_2 \int_\Omega u(t)\cdot Sw(t)\,\mathrm{d}x + \varepsilon_2 \int_{\Gamma_0} w(t)v(t)\,\mathrm{d}x.$$

The positive constants a, ε_1 and ε_2 will be chosen below. Note that for sufficiently small ε_1 and ε_2, the functional $L(t)$ and the energy $E(t) := \frac{1}{2}\|(u(t), v(t), w(t), z(t))\|_X^2$ are equivalent, that is, there exist constants $c_1, c_2 > 0$ independent of t such that $c_1E(t) \leq L(t) \leq c_2E(t)$ for every $t \geq 0$.

Revising the dissipativity estimate (5) we have

$$\frac{d}{dt}E(t) \leq -\varepsilon \int_\Omega |\nabla u(t)|^2\,\mathrm{d}x - \left(\alpha - \beta + \frac{\mu - \varepsilon}{C_P}\right)\int_{\Gamma_0}|v(t)|^2\,\mathrm{d}x \tag{17}$$

where $\varepsilon > 0$ is small enough so that $k := \alpha - \beta + \frac{\mu-\varepsilon}{C_P} > 0$. On the other hand, taking the derivative of the second term of L and then using the transport equation for z we have

$$\begin{aligned}\frac{d}{dt}\int_{-\tau}^{0}\int_{\Gamma_0} e^{a\theta}|z(t,\theta)|^2\,\mathrm{d}x\,\mathrm{d}\theta &= \int_{-\tau}^{0}\int_{\Gamma_0} e^{a\theta}\partial_\theta(|z(t,\theta)|^2)\,\mathrm{d}x\,\mathrm{d}\theta \\ &= \int_{\Gamma_0}(|v(t)|^2 - e^{-a\tau}|z(t,-\tau)|^2)\,\mathrm{d}x - a\int_{-\tau}^{0}\int_{\Gamma_0} e^{a\theta}|z(t,\theta)|^2\,\mathrm{d}x\,\mathrm{d}\theta. \end{aligned}\tag{18}$$

Taking the derivative of the third term of L and using the fact that $\operatorname{div} Sw = 0$, $Sw = 0$ on Γ_1 and $Sw = wv$ on Γ_0 we have

$$\begin{aligned}&\frac{d}{dt}\int_\Omega u(t)\cdot Sw(t)\,\mathrm{d}x \\ &= \int_\Omega (\mu\Delta u(t) - \nabla p(t))\cdot Sw(t)\,\mathrm{d}x + \int_\Omega u(t)\cdot Sv(t)\,\mathrm{d}x \\ &= -\mu\int_\Omega \nabla u(t)\cdot\nabla Sw(t)\,\mathrm{d}x + \langle \mu v\cdot\partial_v u(t) - p(t), w(t)\rangle + \int_\Omega u(t)\cdot Sv(t)\,\mathrm{d}x.\end{aligned}\tag{19}$$

From Theorem 4 and the Poincaré inequality, we have the estimates

$$\left|\mu \int_\Omega \nabla u(t) \cdot \nabla S w(t)\,\mathrm{d}x\right| \le C_{\mu,\varepsilon_3} \int_\Omega |\nabla u(t)|^2\,\mathrm{d}x + \varepsilon_3 \int_{\Gamma_0} |\nabla w(t)|^2\,\mathrm{d}x \tag{20}$$

$$\left|\int_\Omega u(t) \cdot S v(t)\,\mathrm{d}x\right| \le C_\mu \int_\Omega |\nabla u(t)|^2\,\mathrm{d}x + C \int_{\Gamma_0} |v(t)|^2\,\mathrm{d}x. \tag{21}$$

Let C_{Γ_0} be the Poincaré constant corresponding to the domain Γ_0. Then we have

$$\begin{aligned}
&\langle \mu v \cdot \partial_\nu u(t) - p(t), w(t)\rangle + \frac{d}{dt}\int_{\Gamma_0} v(t)w(t)\,\mathrm{d}x - \int_{\Gamma_0} |v(t)|^2\,\mathrm{d}x \\
&= -\int_{\Gamma_0} |\nabla w(t)|^2\,\mathrm{d}x - \int_{\Gamma_0} (\alpha v(t) - \beta z(t,-\tau))w(t)\,\mathrm{d}x \\
&\le -(1-\varepsilon_3 C_{\Gamma_0})\int_{\Gamma_0} |\nabla w(t)|^2\,\mathrm{d}x - C_{\alpha,\beta,\varepsilon_3}\int_{\Gamma_0} (|v(t)|^2 + |z(t,-\tau)|^2)\,\mathrm{d}x.
\end{aligned} \tag{22}$$

Therefore, if we choose the positive constants ε_i for $i = 1, 2, 3$ in such a way that $\varepsilon - \varepsilon_2(C_\mu + C_{\mu,\varepsilon_3}) > 0, k - \varepsilon_1 - (1 + C + C_{\alpha,\beta,\varepsilon_3})\varepsilon_2 > 0, \varepsilon_1 e^{-a\tau} - C_{\alpha,\beta,\varepsilon_3}\varepsilon_2 > 0$ and $1 - \varepsilon_3(1 + C_{\Gamma_0}) > 0$, then from (17) to (22) we can see that there exists a positive constant $C > 0$ such that $L'(t) \le -CL(t)$. Using the equivalence of L and E, we obtain the desired result.

Acknowledgements The author is supported by the Philippine Commission on Higher Education (CHED) and by the Ernst-Mach Grant of the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH).

References

1. G. Avalos, R. Triggiani, The coupled PDE system arising in fluid-structure interaction, Part I: Explicit semigroup generator and its spectral properties. Contemp. Math. **440**, 15–54 (2007)
2. G. Avalos, R. Triggiani, Semigroup wellposedness in the energy space of a parabolic-hyperbolic coupled Stokes-Lamé PDE system of fluid-structure interaction. Discr. Contin. Dyn. Syst. **2**, 417–447 (2009)
3. G. Avalos, R. Triggiani, Fluid structure interaction with and without internal dissipation of the structure : A contrast study in stability. Evol. Equ. Control Theory **2**, 563–598 (2013)
4. I. Chuesov, A global attractor for a fluid-plate interaction model accounting only for longitudinal deformations of the plate. Math. Methods Appl. Sci. **34**, 1801–1812 (2011)
5. I. Chuesov, I. Ryzhkova, A global attractor for a fluid-plate interaction model. Commun. Pure Appl. Anal. **12**(4), 1635–1656 (2013)
6. W. Desch, Fašangová, E., Milota J., Propst, G.: Stabilization through viscoelastic boundary damping: a semigroup approach. Semigroup Forum **80**, 405–415 (2010)
7. M. Kirane, B. Said-Houari, Existence and asymptotic stability of a viscoelastic wave equation with delay. Z. Angew. Math. Phys. **62**, 1065–1082 (2011)
8. J.L. Lions, E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, vol. 1 (Springer, New York, 1972)
9. J.L. Lions, E. Zuazua, Approximate controllability of a hydro-elastic coupled system. ESAIM Control Optim. Calc. Var. **1**, 1–15 (1995)

10. S. Nicaise, C. Pignotti, Stability and instability results of the wave equation with a delay term in boundary or internal feedbacks. SIAM J. Control Optim. **45**, 1561–1585 (2006)
11. G. Peralta, A fluid-structure interaction model with interior damping and delay in the structure. Z. Angew. Math. Phys. **67**, 1–20 (2016)
12. R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis* (AMS Chelsea Publishing, Providence, Rhode Island, 2001)
13. M. Tucsnak, G. Weiss, *Observation and Control for Operator Semigroups* (Birkhäuser, Basel, 2009)

A Unified Hyperbolic Formulation for Viscous Fluids and Elastoplastic Solids

Michael Dumbser, Ilya Peshkov and Evgeniy Romenski

Abstract We discuss a unified flow theory which in a single system of hyperbolic partial differential equations (PDEs) can describe the two main branches of continuum mechanics, fluid dynamics and solid dynamics. The fundamental difference from the classical continuum models, such as the Navier–Stokes, for example, is that the finite length scale of the continuum particles is not ignored but kept in the model in order to semi-explicitly describe the essence of any flows, that is the process of continuum particles rearrangements. To allow the continuum particle rearrangements, we admit the deformability of particle which is described by the distortion field. The ability of media to flow is characterized by the strain dissipation time which is a characteristic time necessary for a continuum particle to rearrange with one of its neighboring particles. It is shown that the continuum particle length scale is intimately connected with the dissipation time. The governing equations are represented by a system of first-order hyperbolic PDEs with source terms modeling the dissipation due to particle rearrangements. Numerical examples justifying the reliability of the proposed approach are demonstrated.

Keywords Unified flow theory · Hyperbolic equations · Viscous fluids Elastoplasticity

M. Dumbser
Department of Civil, Environmental and Mechanical Engineering, University of Trento, Via Mesiano 77, 38123 Trento, Italy
e-mail: michael.dumbser@unitn.it

I. Peshkov (✉)
Institut de Mathématiques de Toulouse, Université Toulouse III, 31062 Toulouse, France
e-mail: peshenator@gmail.com

I. Peshkov · E. Romenski
Sobolev Institute of Mathematics, 4 Acad. Koptyug Avenue, 630090 Novosibirsk, Russia

E. Romenski
Novosibirsk State University, 2 Pirogova Str., 630090 Novosibirsk, Russia
e-mail: evrom@math.nsc.ru

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_34

1 Introduction

This paper contains an extended abstract of the talk given at the XVI International Conference on Hyperbolic Problems Theory, Numerics, Applications (HYP2016), Aachen (Germany), August 1–5, 2016. The talk was dedicated to the unified hyperbolic formulation of fluid and solid dynamics recently proposed in [1, 2]. In particular, the emphasis was done on the discussion of the *physical model* underlying the mathematical formulation. To emphasize how important such a physical interpretation of the mathematical model is, we recall that the equations which constitute the model were proposed many years ago, back to 1970th, by Godunov and Romenski in [3, 4] for modeling of large elastoplastic deformations in metals, and the equations were used until recently only in the solid dynamics context by several authors, e.g., [5–16] to cite just a few. Moreover, similar equations and even an idea to apply them to modeling of fluids were proposed by Besseling in [17], but unfortunately it has never been appreciated in the fluid dynamics context nor by Besseling itself neither by the others. Perhaps, one of the reason for that the hyperbolic Godunov–Romenski equations was not even thought to be used in the fluid dynamics context is an *exceptional role* of the parabolic Navier–Stokes–Fourier (NSF) equations in the fluid dynamics. For example, it is believed that any mathematical model aiming to describe viscous flows has to *literally* coincide with the NSF equations in the diffusion regime. This should be understood as that the second-order parabolic terms should appear *explicitly* in the PDEs, and they are a fundamental *hallmark* of the diffusion in the mathematical description. For instance, the well-known first-order hyperbolic extension of the NSF equations, the Maxwell–Cattaneo equations

$$\dot{X} = -\frac{1}{\lambda}\left(X - X^{\mathrm{NSF}}\right), \tag{1}$$

relax to the NSF equations as the relaxation parameter $\lambda \to 0$. Here, X is a dissipative quantity in the Maxwell–Cattaneo approach, while X^{NSF} is the value of X obtained in the framework of the NSF theory, the upper dot denotes a time derivative. This, in particular, results in that some characteristic speeds of the Maxwell–Cattaneo equations *unphysically* tend to infinity as $\lambda \to 0$. From the other side, as it is shown in [1, 2], there are no physical reasons saying that the diffusion processes should be exclusively modeled by the second-order parabolic equations, and a radically different first-order hyperbolic description which is not based on the steady-state laws such as Newton's law of viscosity or Fourier law of heat conduction is possible.

Perhaps, the right question in this context is that after more than one hundred year history of successful use of the NSF theory, do we need at all another transport theory different from the classical parabolic approach? From a practical point of view, the answer is not clear yet, but from a physical viewpoint the answer is obviously positive. Indeed, the heart of the NSF equations, Newton's law of viscosity and Fourier's law of heat conduction, are the *phenomenological* laws and thus should be substituted by more physically meaningful laws. We thus would like to emphasize an important role of the physical model in that it helped us to dare to propose an alternative physically

based description of viscous dissipation. Eventually, it is necessary to note that our hyperbolic unified approach is now well established after an extensive comparison with the NSF theory in [2]. Moreover, the model was recently extended in [18] in order to include the interaction of matter with the electromagnetic field where we also provided an extensive comparison of the extended model with the ideal MHD and parabolic viscous resistive MHD equations.

2 Physical Model

Despite we oppose our model to the classical continuum models such as the NSF equations, we underline that the proposed approach entirely relies on the conventional postulates of continuum mechanics and thermodynamics. The main difference though is that we do not assume some simplifications which are implied in the classical theories. Namely, the key difference is that the *continuum particles* are not treated as *scaleless mathematical points* but are considered as the finite volumes of a small but *finite scale* ℓ. Recall that the notion of the continuum particle is central in any continuum theory. This notion relies on the longtime observations suggesting that for the macroscopic description of the dynamics of matter (gas, liquid, or solid), the very detailed information about the molecular motion is irrelevant but the dynamics of *ensembles of molecules* instead should be considered as a dynamics of new entities of a particle nature. Of course, such particles have not to exist forever but only during a finite time which shall be considered later as an important measure of *fluidity*. Thus, in our approach, the continuum is represented by a system of finite scale particles (finite volumes) covering the space occupied by the media without gaps; see Fig. 1.

Once one admits or, rather to say, does not ignore that the continuum particles have a finite scale ℓ, the description of the flow becomes straightforward because the essence of any flow phenomena in any system of finite scale particles is the process of *rearrangements* of these particles.[1] Thus, the central task of our approach is *to find a mathematical framework to express the dynamics of the continuum particles*. Further, as depicted in Fig. 1, there is no *free volume* between the continuum particles, and hence to allow the particle rearrangements, we have to admit the deformability of the particles; otherwise (if they would be rigid volumes), they cannot rearrange and the flow is impossible. Thus, in our approximation, the continuum particles are the structureless (homogeneous) "soft" deformable particles. The ability of the particles to deform can be characterized by the ability to transmit the *transversal perturbations*, which in turn can be characterized by the shear sound speed, denoted here by c_s. As

[1] From the other hand, in the context of the scaleless particles of classical continuum mechanics it is impossible to define the rearrangements because the notion of a *neighboring continuum particle* becomes *indefinite*, and thus what remains is not to describe the flow itself but rather to *mimic* some indirect flow indicators, such as stress–strain-rate relations, etc. Such a mimic strategy is of course admissible in the engineering problems, but it is unable to give a meaningful explanation to the physical phenomena.

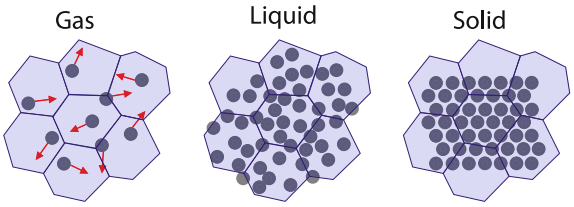


Fig. 1 Sketch of the continuum particles (honeycomb-like cells). If the scale of the continuum particles is not ignored, then the continuum representation of all three states of matter is identical. The circles represent the real molecules

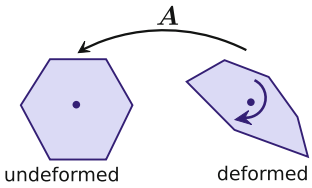


Fig. 2 Sketch for the distortion field $\boldsymbol{A}$. It maps a particle from a current deformed state to the undeformed stress free state

for the measure of the deformation of particles, we use the distortion field $\boldsymbol{A} = [A_{ij}]$ which maps a particle from the current deformed state to the undeformed state; see Fig. 2.

Furthermore, the ability of the particles to rearrange, or to change their neighbors, can be characterized by a time τ which is the characteristic time necessary for a given particle to rearrange with one of its neighboring particles. Because we keep the finite scale of the continuum particles in the physical model, it is then obvious that such particles cannot rearrange instantaneously because of the *causality principle*, and hence the time τ is also finite. The time τ is a continuum interpretation of the seminal idea of the so-called *particle settled life time* of Frenkel [19], who applied it to describe the ability of liquids to flow, see also recent promising experimental and theoretical advances [20–24] confirming and further developing Frenkel's ideas.

Another non-trivial, and probably the most important, consequence of the finiteness of the particle length scale is that because the particles cannot rearrange instantaneously, there is a *relative* motion between the neighboring particles; see Fig. 3. Such a relative motion assumes the existence of a slip between neighboring particles. In turn, the transversal perturbations that carry the information about the deformation of the continuum particles cannot propagate across such slip planes without a loss of information. This results in that the distortion field is *incompatible*,[2] or not integrable [1, 4, 25]. Such a loss of information is represented by a dissipation term in the time evolution for the distortion field which "dissipates" shear deformation stored in $\boldsymbol{A}$. This term is proportional to $1/\tau$, and thus time τ is also referred to as the characteristic strain dissipation time in our papers [1, 2, 18].

[2]The incompatibility condition for $\boldsymbol{A}$ is $\boldsymbol{B} := \mathrm{curl}(\boldsymbol{A}) \neq 0$, where $\boldsymbol{B}$ is a so-called Burgers tensor which is interpreted as the number density of the slips (defects) between continuum particles. The term $\mathrm{curl}(\boldsymbol{A})$ also emerges in the time evolution for $\boldsymbol{A}$.

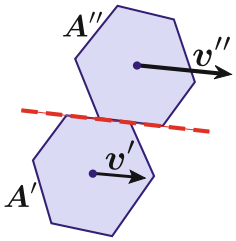


Fig. 3 Sketch for the particle rearrangements. Because the continuum particles are finite, there is a relative velocity $\boldsymbol{w} = \boldsymbol{v}'' - \boldsymbol{v}'$ between neighboring particles. While the longitudinal (pressure) perturbations can propagate across the slip plane, the transversal perturbations can not propagate without a loss of information, and thus the distortions $\boldsymbol{A}'$ and $\boldsymbol{A}''$ are *incompatible*

At this point, it is necessary to emphasize that the *representation of the continuum by a system of finite volumes* is what actually *unifies* all the three states of mater, gaseous, liquid, and solid, because now, the problem of the continuum particle dynamics (finite volumes) is *essentially a geometrical problem* (deformation problem). Such a geometrical reformulation is insensible to the *content* of the continuum particles.

It is also clear from the above discussion that the main approximation of our physical model is the treatment of the continuum particles as structureless homogeneous elastic volumes. However, as it is shown in [1, 2] such an approach is a very precise approximation as long as the characteristic wave length λ of the mechanical perturbations is larger than the particle length scale ℓ. Moreover, as it is proven in [2], the knowledge of only the continuum particle dynamics is sufficient to build a unified flow theory for gases and liquids which incorporates the Newtonian behavior of viscous fluids as a particular case. On the contrary, the *exceptionally different* molecular dynamics of gases and liquids suggests that not the molecular dynamics is responsible for the *mathematical form* of the transport laws (identical in both cases) but a dynamics at a larger scale, which we believe is the scale of the continuum particles. Thus, the knowledge of the length scale ℓ is extremely important for understanding of the limits of applicability of our physical model, and as will be shown later with the dispersion analysis, the continuum particle length scale is of the order of $\tau\, c_s$, i.e.

$$\ell \sim \tau\, c_s. \tag{2}$$

If one needs to deal with a problem solution to which strongly depends on the dynamics at a scale for which $\lambda \sim \ell$ or even $\lambda < \ell$, then it is necessary to enlarge the model by providing a more accurate description of the perturbation propagation inside of the continuum particles.

3 Mathematical Model

The governing PDEs are formulated for the following volume average quantities

$$(\boldsymbol{m}, \boldsymbol{A}, \rho, \sigma), \tag{3}$$

where $\boldsymbol{m} = [m_i] = \rho \boldsymbol{v}$ is the momentum density, ρ is the mass density, $\boldsymbol{v} = [v_i]$ is the velocity vector, $\boldsymbol{A} = [A_{ij}]$ is the distortion field, and $\sigma = \rho s$ is the entropy density, while s is the specific entropy. Also, an exceptional role is played by the total energy density potential

$$\mathscr{E} = \mathscr{E}(\boldsymbol{m}, \boldsymbol{A}, \rho, \sigma) \tag{4}$$

which plays the role of a generating potential as discussed in details in [2].

The system of governing PDEs can be written as

$$\frac{\partial m_i}{\partial t} + \frac{\partial \left(m_i v_k + [m_l \mathscr{E}_{m_l} + \rho \mathscr{E}_\rho + \sigma \mathscr{E}_\sigma - \mathscr{E}] \delta_{ik} + A_{li} \mathscr{E}_{A_{lk}} \right)}{\partial x_k} = 0, \tag{5a}$$

$$\frac{\partial A_{ik}}{\partial t} + \frac{\partial (A_{il} v_l)}{\partial x_k} + v_j \left(\frac{\partial A_{ik}}{\partial x_j} - \frac{\partial A_{ij}}{\partial x_k} \right) = - \frac{\mathscr{E}_{A_{ik}}}{\theta}, \tag{5b}$$

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho v_k)}{\partial x_k} = 0, \tag{5c}$$

$$\frac{\partial \sigma}{\partial t} + \frac{\partial (\sigma v_k)}{\partial x_k} = \frac{1}{\mathscr{E}_\sigma \theta} \mathscr{E}_{A_{ij}} \mathscr{E}_{A_{ij}} \geq 0, \tag{5d}$$

while the energy conservation law reads as

$$\frac{\partial \mathscr{E}}{\partial t} + \frac{\partial}{\partial x_k} \left(\mathscr{E} v_k + v_n \left([m_l \mathscr{E}_{m_l} + \rho \mathscr{E}_\rho + \sigma \mathscr{E}_\sigma - \mathscr{E}] \delta_{nk} + A_{ln} \mathscr{E}_{A_{lk}} \right) \right) = 0. \tag{6}$$

As in all our previous papers [1, 2, 18], the notations such as $\mathscr{E}_\rho, \mathscr{E}_{m_i}, \mathscr{E}_{A_{ij}}, \mathscr{E}_\sigma$ are used to denote the partial derivatives $\partial \mathscr{E} / \partial \rho, \partial \mathscr{E} / \partial m_i$. Thus, to specify all the terms in the equations, that is to close the system, one needs to specify the total energy $\mathscr{E}$. Also, $\theta \sim \tau$ is a relaxation parameter which will be specified later. These two scalar functions, $\mathscr{E}$ and θ, are the only degrees of freedom in the model formulation. For example, the non-dissipative part of the PDEs, i.e., all the differential terms which are collected on the left-hand side, is *complete* in the sense that no differential terms can be added or removed and the only possibility to modify something is to change the potential $\mathscr{E}$. The dissipative part of the equations is the only algebraic source

terms on the right-hand side which depend on the specification of the energy and the dissipation parameter θ.

The non-advective momentum flux

$$\Sigma_{ik} = -[m_l \mathscr{E}_{m_l} + \rho \mathscr{E}_\rho + \sigma \mathscr{E}_\sigma - \mathscr{E}]\delta_{ik} - A_{li}\mathscr{E}_{A_{lk}} \tag{7}$$

is the stress tensor. Its form is completely defined by the structure of the time evolution equations while its further specification depends solely on the choice of the energy $\mathscr{E}$. Here, the scalar $p = m_l \mathscr{E}_{m_l} + \rho \mathscr{E}_\rho + \sigma \mathscr{E}_\sigma - \mathscr{E}$ can be refereed to as the pressure which coincides with the classical hydrodynamic pressure for equilibrium flows. Indeed, if one introduces the specific total energy density E as $\mathscr{E} = \rho E$, for which the following decomposition is usually assumed

$$E = E^1(\rho, s, \boldsymbol{A}) + \frac{1}{2} v_i v_i \tag{8}$$

then $p = m_l \mathscr{E}_{m_l} + \rho \mathscr{E}_\rho + \sigma \mathscr{E}_\sigma - \mathscr{E} = \rho^2 E^1_\rho$, exactly as in our previous paper [1]. The last term in (7) represents the viscous stresses or elastic stresses in the case of solid dynamics.

For the further specification of the total energy potential $\mathscr{E}$, we note that there are three scales involved in the physical model formulation described in Introduction. Namely, the molecular scale, called here *microscale*; the scale of the continuum particles, called here *mesoscale*; and the flow scale, or observable *macroscale*. We thus assume that E is a sum of three terms each of which represents the amount of energy stored on the corresponding scale

$$E = E^{\mathrm{mi}}(\rho, s) + E^{\mathrm{me}}(\rho, s, \boldsymbol{A}) + E^{\mathrm{ma}}(\boldsymbol{v}). \tag{9}$$

The terms E^{mi} and E^{ma} are conventional. They are the kinetic energy $E^{\mathrm{ma}}(\boldsymbol{v}) = \frac{1}{2} v_i v_i$, which represents the part of the total energy stored in the macroscale, and an internal energy $E^{\mathrm{mi}}(\rho, s)$ represents the kinetic energy of the molecular motion. In [2, 18], we used an ideal gas equation or stiffened gas equation of state for E^{mi} to model gases or liquids and solids, respectively

For the mesoscopic, or *non-equilibrium*, part of the total energy, we shall use a quadratic form

$$E^{\mathrm{me}} = \frac{c_s^2}{4} G^{\mathrm{TF}}_{ij} G^{\mathrm{TF}}_{ij}, \tag{10}$$

where $G^{\mathrm{TF}}_{ij} = G_{ij} - G_{ii}/3$ is the deviator of the tensor $G_{ij} = A_{ki}A_{kj}$, c_s is the characteristic velocity of propagation of transversal perturbations, we call it here *shear sound velocity*. In general, c_s is a function of ρ and s.

With such a specification of the term E^{me}, the explicit form of the viscous/elastic stress (the last term in (7)) which we denote by σ_{ik} is

$$\sigma_{ik} = \rho c_s^2 G_{il} G^{\mathrm{TF}}_{lk}. \tag{11}$$

The mesoscopic energy E^{me} also defines [2] the dissipation terms as

$$\frac{\mathscr{E}_{\boldsymbol{A}}}{\theta} = \frac{3}{\tau}|\boldsymbol{A}|^{\frac{5}{3}}\boldsymbol{A}\boldsymbol{G}^{\mathrm{TF}}, \tag{12}$$

where we use $\theta = \tau c_s^2/3|\boldsymbol{A}|^{\frac{5}{3}}$ for θ and $|\boldsymbol{A}|$ to denote the determinant of $\boldsymbol{A}$. In general, τ is a function of the state variables $\tau = \tau(\rho, s, \boldsymbol{A})$ while for Newtonian fluids, it can be taken to be constant as shown in [2] through a formal asymptotic analysis. In particular, the dependence of τ on $\boldsymbol{A}$ defines the non-Newtonian properties of fluids or controls the transition from elastic to plastic regime in solids [4, 9, 10, 25]; see also numerical examples in the following Sect. 4.

4 Numerical Results

In this section, we demonstrate that the proposed model can be applied to modeling of non-equilibirum effects in gases as well as to modeling of viscous fluid flows and elastoplastic deformation in metals.

4.1 Non-equilibrium Sound Wave Propagation in a Viscous Gas

We first study the propagation of plane acoustic waves of an angular frequency ω in a viscous gas. As it is well known, the presence of the dissipative process gives rise to the phenomena called *dispersion* when the wave phase speed V depends on the frequency of the wave, $V = V(\omega)$. This dependency is defined by the dispersion relation for a given model. The dispersion relation for the proposed hyperbolic model can be found in [1] in Sect. 2.2.2. The phase velocity $V(\omega)$ and the attenuation factor for Eqs. (5), (9), and (10) are presented in Fig. 4 for Helium.

As can be seen in Fig. 4 (left), at a frequency $\omega^* = 2\pi/\tau$ the dispersion almost disappears and the phase velocity $V(\omega)$ tends to a constant value $c_\infty = \sqrt{c_0^2 + 4c_s^2/3}$ (see also [1]) called a high-frequency limit for the sound speed. This experimentally defined value can be used to estimate the shear sound speed c_s, and subsequently to estimate the dissipation time τ from the relation $\eta = \frac{\rho}{6}c_s^2\tau$ for the shear viscosity η.

The dispersion disappearance of $V(\omega)$ is fully conditioned by the physical model underlying the mathematical formulation. Indeed, because the continuum particles have the finite scale ℓ, the behavior of $V(\omega)$ should change when the wavelength λ becomes comparable with the particle size, $\lambda = \lambda^* \sim \ell$. We thus can use this fact to estimate the particle length scale ℓ as $\ell \sim \tau c_s$. Indeed,

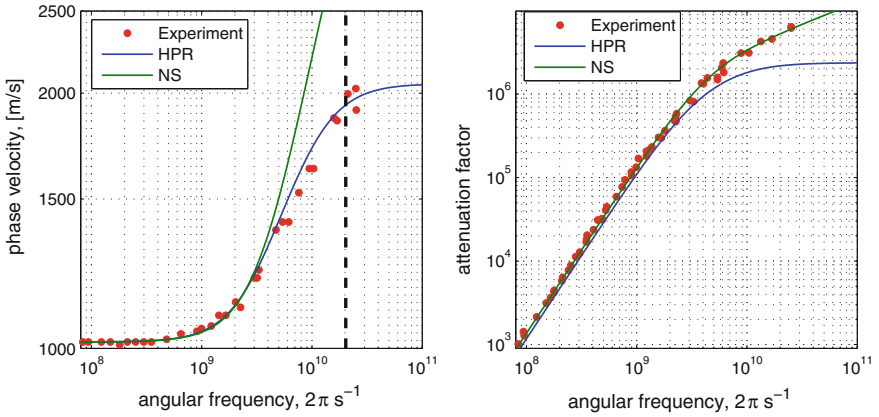


Fig. 4 Sound wave dispersion in Helium. Comparison with the experimental data (red dots) from [28]. The blue solid lines correspond to the hyperbolic model, and the green solid lines correspond to the Navier–Stokes–Fourier model. The vertical black dashed line in the left figure corresponds to the frequency $\omega^* = 2\pi/\tau \approx 2 \cdot 10^{10}\, 2\pi\ \mathrm{s}^{-1}$ and to the wavelength $\lambda \approx 10^{-7}$ m

$$\lambda^* = \frac{V(\omega^*)}{\omega^*} \approx \frac{c_\infty}{\omega^*} \sim \frac{c_s}{2\pi/\tau} \sim \tau c_s. \tag{13}$$

Thus, for the experimental data presented in Fig. 4, the continuum particle length scale can be estimated as $\ell \approx 4.6 \cdot 10^{-7}$ m. For this, we took $c_\infty = 2\,052$ m/s, shear viscosity $\eta = 2 \cdot 10^{-5}$ Pa·s, and mass density $\rho = 0.16\,\mathrm{kg/m^3}$ which gives us $c_s = \sqrt{(c_\infty^2 - c_0^2)3/4} \approx 1\,543$ m/s and $\tau = 6\eta/(\rho c_s^2) \approx 3 \cdot 10^{-10}$ s.

Eventually, we note that there is a certain discrepancy in the attenuation factor visible in Fig. 4(right) if compared with the experimental data, while the Navier–Stokes–Fourier model (see Chap. 11 of [26] for the dispersion relation for the NSF equations) shows an excellent agreement. First of all, one should note that, at high values of ω, there may be a contribution to the absorption arising from diffusion in the piezoelectric receiver, as pointed out by Woods and Troughton [27] (see also discussion in Chap. 11 of book [26]), so that the experimental result for the absorption factor should be considered as an upper limit to the actual value. Secondly, so far, we ignore such important processes as heat transfer and volume relaxation which of course should increase the dissipation. This will be studied elsewhere.

4.2 *Viscous Fluids and Elastoplastic Solids*

In order to demonstrate the ability of the proposed unified hyperbolic model to deal with *radically different* behaviors of matter such as flows of viscous gases and elastoplastic deformation in solids, we consider two 2D problems. These examples

merely serve to demonstrate the diversity of regimes allowed to be captured by the model while the numerical schemes we use in this paper are not very accurate such as those used in our recent papers [2, 16, 18] where much more accurate results were obtained with the use of advanced high-order arbitrary high-order derivatives (ADER) [29], discontinuous Galerkin and finite-volume schemes, moving mesh, and adaptive mesh refinement techniques. An extensive comparison against the parabolic theories like the Navier–Stokes–Fourier equations and resistive MHD model is also provided in [2, 18].

The key parameter controlling the transition between the fluid-like and solid-like behavior is the dissipation time τ. As discussed in the introduction section and in [1, 2], for the elastic solids, the continuum particles do not rearrange and hence time τ is infinite, while for viscous fluids $0 < \tau < \infty$. For elastoplastic solids, time τ depends on the yield strength and rapidly but continuously changes from an infinite value (in fact from a sufficiently large value) to a finite value in the plastic regime, in which the continuum particles do rearrange.

In the first example, a gravity-driven Rayleigh–Taylor instability in a viscous gas confined in a rectangular domain with no-slip boundary conditions is simulated. The domain is a box $(x, y) \in [0, 1/3] \times [0, 1]$ which was discretized with a Cartesian mesh consisting of 200×600 cells, gravitational field is directed vertically downward and has a magnitude $g = 0.1$, the initial conditions are: $\boldsymbol{v} = 0$, the density is taken 2 if $y > 0.5 + 0.01\cos(6\pi x)$ and 1 otherwise, the ideal gas equation of state is used for the internal energy E^{mi} (see (9)) with the ratio of specific heats $\gamma = 1.4$, the pressure is set to $1/\gamma$ everywhere, and the shear sound speed c_s was set to $c_s = c_0 = 1$. For the whole domain, we set the shear viscosity to 10^{-5} Pa·s and the dissipation time to $\tau = 6\eta/c_s^2 \approx 6 \cdot 10^{-5}$ s. Figure 5 depicts several time instants of the simulation. The fluid-like motion (formation of the vortexes) is clearly identified.

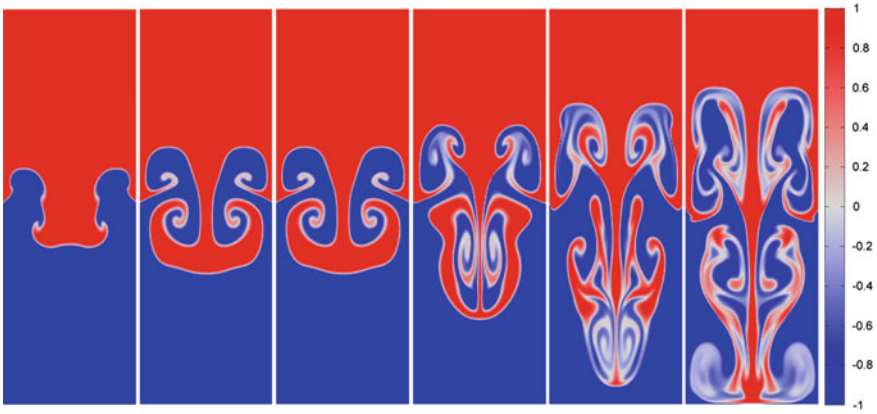


Fig. 5 Rayleigh–Taylor instability in a viscous gas modeled with the proposed hyperbolic model with the equation of state (9)–(10) and $\tau = \text{const} > 0$. The heavier gas is colored in red while the lighter gas is colored in blue. A Cartesian mesh of 200×600 cells and no-slip boundary conditions were used

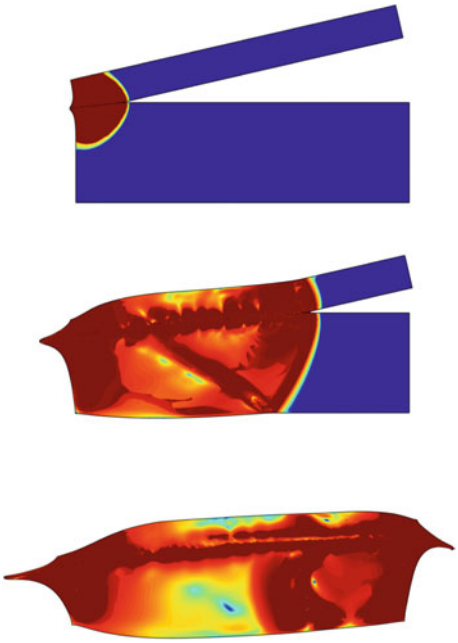


Fig. 6 An oblique high-velocity impact of two solid plates. Three instants of time are shown. The colors represent the norm of the stress tensor deviator. The dark red corresponds to 0.04 GPa and indicates the zones of plastic deformation, while all other colors correspond to elastic deformation, the blue corresponds to 0 GPa

We also note that the gas sticks to the walls as no-slip boundary conditions is used. For this simulation, we use CLAWPACK software [30] designed specifically for hyperbolic PDEs. The numerical fluxes are obtained via the solution of an approximate Riemann problem which was solved using the eigenvalue decomposition of the Jacobi matrix for the fluxes. The second-order wave propagation algorithm of CLAWPACK with the "minmod" limiter was used.

In the second numerical example, we consider an oblique high-velocity collision of two solid plates presented in Fig. 6. This example is motivated by the explosion welding process [31]. The initial angle between the plates is 13°, and the lower plate is at rest while the upper plate has the velocity 1500 m/s normal to the bottom face. It is assumed that there is no slip on the contact interface between the bodies. The upper plate has dimensions 1×0.1 cm and is discretized with a 600×60 Cartesian mesh while the lower plate has dimensions 1×0.3 and is discretized with a 600×180 Cartesian mesh. The both plates have the same material parameters which were taken as follows. The mass density is 7.9 kg/m^3, the longitudinal sound speed is 6700 m/s, the shear sound speed is 3 150 m/s, the dissipation time was taken as $\tau = \tau_0(\sigma_0/\sigma)^n$ where $\tau_0 = 0.1$ s, σ_Y is a parameter which controls the transition from elastic to plastic regime and was set to 0.0056 GPa, while σ is the second norm of the deviator of the stress tensor. The power law index n was set to $n = 10$. For such parameters, the effective yield strength appears to be $\approx$0.04 GPa. The sound speeds were taken as 6700 and 3150 m/s for longitudinal and transversal sound speed, respectively. The equation of state is given in [10]. This numerical example was performed several

years ago, and model (5) was written in the Lagrangian coordinates which are well suited for the solid dynamics problems; see details in [10]. However, recently, the Eulerian equations (5) were also implemented in an arbitrary Lagrangian Eulerian (ALE) code [16] which also opens new possibilities for more efficient simulation of elastoplastic solids experiencing large deformations. Two independent Cartesian meshes were used in this simulation, and the equations were solved with a standard first-order Godunov scheme with an acoustic Riemann solver; see [10]. The contact boundary requires a specific treatment. For this purpose, the contact cells have to be detected and the numerical flux on the contact interface is obtained with the same Riemann solver that used for the internal cells.

Acknowledgements I.P. have received funding from ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02 and a partial support from the Russian Foundation for Basic Research (grant number 16-31-00146). E.R. acknowledges a partial support by the Program N15 of the Presidium of RAS, project 121 and the Russian Foundation for Basic Research (grant number 16-29-15131). M.D. have received funding from the European Union's Horizon 2020 Research and Innovation Programme under the project *ExaHyPE*, grant agreement number no. 671698 (call FETHPC-1-2014).

References

1. I. Peshkov, E. Romenski, A hyperbolic model for viscous Newtonian flows. Contin. Mech. Thermodyn. **28**(1–2), 85–104 (2016)
2. M. Dumbser, I. Peshkov, E. Romenski, O. Zanotti, High order ADER schemes for a unified first order hyperbolic formulation of continuum mechanics: viscous heat-conducting fluids and elastic solids. J. Comput. Phys. **314**:824–862 (2016), http://www.sciencedirect.com/science/article/pii/S0021999116000693
3. S.K. Godunov, E.I. Romenskii, Nonstationary equations of nonlinear elasticity theory in Eulerian coordinates. J. Appl. Mech. Techn. Phys. **13**(6), 868–884 (1972)
4. S.K. Godunov, Elements of Mechanics of Continuous Media. Nauka
5. E.I. Romenskii, Hyperbolic equations of Maxwell's nonlinear model of elastoplastic heat-conducting media. Sib. Math. J. **30**(4), 606–625 (1989)
6. L.A. Merzhievsky, A.D. Resnyansky, The role of numerical simulation in the study of high-velocity impact. Int. J. Impact Eng. **17**(4), 559–570 (1995)
7. A.D. Resnyansky, DYNA-modelling of the high-velocity impact problems with a split-element algorithm. Int. J. Impact Eng. **27**(7), 709–727 (2002)
8. S.L. Gavrilyuk, N. Favrie, R. Saurel, Modelling wave dynamics of compressible elastic materials. J. Comput. Phys. **227**, 2941–2969 (2008)
9. P.T. Barton, D. Drikakis, E.I. Romenski, An Eulerian finite-volume scheme for large elastoplastic deformations in solids. Int. J. Numer. Methods Eng. **81**(4), 453–484 (2010)
10. S.K. Godunov, I.M. Peshkov, Thermodynamically consistent nonlinear model of elastoplastic maxwell medium. Comput. Math. Math. Phys. **50**(8), 1409–1426 (2010)
11. N. Favrie, S.L. Gavrilyuk, R. Saurel, Solid-fluid diffuse interface model in cases of extreme deformations. J. Comput. Phys. **228**(16), 6037–6077 (2009)
12. A.D. Resnyansky, N.K. Bourne, J.C.F. Millett, E.N. Brown, Constitutive modeling of shock response of polytetrafluoroethylene. J. Appl. Phys. **110**(3), 33530 (2011)
13. P.T. Barton, R. Deiterding, D. Meiron, D. Pullin, Eulerian adaptive finite-difference method for high-velocity impact and penetration problems. J. Comput. Phys. **240**, 76–99 (2013)

14. S. Ndanou, N. Favrie, S. Gavrilyuk, Criterion of hyperbolicity in hyperelasticity in the case of the stored energy in separable form. J. Elast. **115**(1), 1–25 (2014)
15. I. Peshkov, M. Grmela, E. Romenski, Irreversible mechanics and thermodynamics of two-phase continua experiencing stress-induced solid–fluid transitions. Contin. Mech. Thermodyn. **27**(6), 905–940 (2015)
16. W. Boscheri, M. Dumbser, R. Loubère, Cell centered direct Arbitrary-Lagrangian-Eulerian ADER-WENO finite volume schemes for nonlinear hyperelasticity. Comput. Fluids. **134–135**:111–129 (2016), http://linkinghub.elsevier.com/retrieve/pii/S004579301630144X
17. J.F. Besseling, A thermodynamic approach to rheology, in *Irreversible Aspects of Continuum Mechanics and Transfer of Physical Characteristics in Moving Fluids. IUTAM Symposia* ed. by H. Parkus, L.I. Sedov (Springer, Vienna 1968) pp. 16–53
18. M. Dumbser, I. Peshkov, E. Romenski, O. Zanotti, High order ADER schemes for a unified first order hyperbolic formulation of Newtonian continuum mechanics coupled with electrodynamics. J. Comput. Phys. (2017) In Press, http://www.sciencedirect.com/science/article/pii/S0021999117305284
19. J. Frenkel, *Kinetic Theory of Liquids* (Dover, 1955)
20. V.V. Brazhkin, Y.D. Fomin, A.G. Lyapin, V.N. Ryzhov, K. Trachenko, Two liquid states of matter: a dynamic line on a phase diagram. Phys. Rev. E. **85**(3), 31203 (2012)
21. D. Bolmatov, V.V. Brazhkin, K. Trachenko, Thermodynamic behaviour of supercritical matter. Nat. Commun. **4** (2013)
22. D. Bolmatov, M. Zhernenkov, D. Zav'yalov, S. Stoupin, Y.Q. Cai, A. Cunsolo, Revealing the mechanism of the viscous-to-elastic crossover in liquids. J. Phys. Chem. Lett. **6**(15), 3048–3053 (2015)
23. D. Bolmatov, M. Zav'yalov, M. Zhernenkov, E.T. Musaev, Y.Q. Cai, Unified phonon-based approach to the thermodynamics of solid, liquid and gas states. Ann. Phys. **363**:221–242 (2015). https://doi.org/10.1016/j.aop.2015.09.018
24. D. Bolmatov, M. Zhernenkov, D. Zav'yalov, S. Stoupin, A. Cunsolo, Y.Q. Cai, Thermally triggered phononic gaps in liquids at THz scale. Sci. Rep. **6**(November 2015):19469 (2016), http://www.nature.com/articles/srep19469
25. S.K. Godunov, E.I. Romenskii, *Elements of Continuum Mechanics and Conservation Laws* (Kluwer Academic/Plenum Publishers, 2003)
26. D. Jou, J. Casas-Vázquez, G. Lebon, *Extended Irreversible Thermodynamics* (Springer, Dordrecht, Berlin, Heidelberg, 2010). https://doi.org/10.1007/978-90-481-3074-0
27. L.C. Woods, H. Troughton, Transport processes in dilute gases over the whole range of Knudsen numbers. Part 2. Ultrasonic sound waves. J. Fluid Mech. **100**(02):321–331 (1980), http://www.journals.cambridge.org/abstract_S0022112080001176
28. M. Greenspan, Propagation of sound in five monatomic gases. J. Acoust. Soc. Am. **28**(4), 644–648 (1956)
29. E.F. Toro, V.A. Titarev, Derivative Riemann solvers for systems of conservation laws and ADER methods. J. Comput. Phys. **212**(1), 150–165 (2006)
30. Clawpack Development Team. Clapack software (2014), http://www.clawpack.org
31. S.K. Godunov, A.A. Deribas, A.V. Zabrodin, N.S. Kozin, Hydrodynamic effects in colliding solids. J. Comput. Phys. **5**(3), 517–539 (1970)

On the Transverse Diffusion of Beams of Photons in Radiation Therapy

S. Brull, B. Dubroca, M. Frank and T. Pichard

Abstract Typical external radiotherapy treatments consist in emitting beams of energetic photons targeting the tumor cells. Those photons are transported through the medium and interact with it. Such interactions affect the motion of the photons but they are typically weakly deflected which is not well modeled by standard numerical methods. The present work deals with the transport of photons in water. The motion of those particles is modeled by an entropy-based moment model, i.e., the M_1 model. The main difficulty when constructing numerical approaches for photon beam modeling emerges from the significant difference of magnitude between the diffusion effects in the forward and transverse directions. A numerical method for the M_1 equations is proposed with a special focus on the numerical diffusion effects.

Keywords Entropy-based moment model · Photon beam modeling
Transverse diffusion

1 Introduction

Radiotherapy treatments consist in emitting radiations to target cancer cells. Such radiations deposit energy in the medium, so-called dose, which is responsible for biological effects (see, e.g., [18]). Radiations can be seen as beams of energetic particles traveling through a medium. Here, the motion of photons modeled by a linear Boltzmann equation is focused on. Solving directly such kinetic equations requires

S. Brull
IMB, Université de Bordeaux, 351 cours de la libération, 33400 Talence, France

B. Dubroca
CELIA, Université de Bordeaux, 351 cours de la libération, 33400 Talence, France

M. Frank
MathCCES, RWTH Aachen University, Schinkelstr. 2, 52062 Aachen, Germany

T. Pichard (✉)
LJLL, Université Pierre et Marie Curie, 4 Place Jussieu,75004 Paris, France
e-mail: pichard@ann.jussieu.fr

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_35

high computational powers. As an alternative, the method of moments is used, leading to the so-called M_1 model. At the numerical level, such a model is cheaper than kinetic ones. However, moment equations require particular considerations because they are nonlinear and their solution is constrained by a realizability condition (specified below).

The present work is a follow-up to [5, 15, 16] which is devoted to adapt the numerical scheme presented in [15] to accurately model beams of photons. Such beams travel almost straight in a human-sized medium. The main difficulty emerges from the difference of magnitude of the diffusion effects in the forward direction and in the direction normal to the beam. Standard numerical methods typically overestimate the transverse diffusion which affects the accuracy of the results.

In the next section, the motion of photons is modeled, through kinetic and M_1 models. A standard numerical method is presented in Sect. 3 and tested in Sect. 4. The problem of the transverse diffusion is presented and solved in Sect. 5. The last section is devoted to conclusion.

2 Photon Transport Models

For the sake of simplicity, only the motion of the photon is studied. The photons are assumed to collide only with atoms of the background medium.

2.1 A Kinetic Model

At the kinetic level, the motion of the photons is modeled by the fluence ψ of photons, which satisfies the following linear Boltzmann equation

$$\Omega.\nabla_x \psi(\varepsilon, x, \Omega) = \int_{\varepsilon}^{\varepsilon_{\max}} \int_{S^2} \sigma(\varepsilon', \varepsilon, \Omega'.\Omega)\psi(\varepsilon', x, \Omega')d\Omega' d\varepsilon' - \sigma_T(\varepsilon)\psi(\varepsilon, x, \Omega), \tag{1}$$

where ψ depends on energy $\varepsilon \in [\varepsilon_{\min}, \varepsilon_{\max}]$, position $x \in Z \subset \mathbb{R}^3$, and direction of flight $\Omega \in S^2$. The physical parameters σ and σ_T are called respectively differential and total cross sections, and they are chosen to model Compton collisions [4] as this effect is predominant in the considered energy range. Other effects may be considered for further applications.

In this equation, the ε variable is considered similarly as a numerical time, and due to the energy integral in (1), such equation is solved backward in energy, from a maximum energy $\varepsilon_{\max}$ to a minimum one $\varepsilon_{\min}$.

Discretizing directly this equation is computationally too expensive for application in medical centers. For this purpose, the method of moments is applied.

2.2 The M_1 Model

The method of moments consists in studying angular moments, i.e., weighted integrals of ψ according to the variable Ω, instead of the fluence itself. Those moments depend on less variables and are therefore typically cheaper at the computational level. The moments of ψ of order up to two are defined by

$$\psi^0 = \int_{S^2} \psi(\Omega) d\Omega, \quad \psi^1 = \int_{S^2} \Omega \psi(\Omega) d\Omega, \quad \psi^2 = \int_{S^2} \Omega \otimes \Omega \psi(\Omega) d\Omega. \tag{2}$$

According to (1), the moments of ψ satisfy the following equations

$$\nabla_x . \psi^1(\varepsilon, x) = \int_{\varepsilon}^{\varepsilon_{\max}} \sigma^0(\varepsilon', \varepsilon) \psi^0(\varepsilon', x) d\varepsilon' - \sigma_T(\varepsilon) \psi^0(\varepsilon, x), \tag{3a}$$

$$\nabla_x . \psi^2(\varepsilon, x) = \int_{\varepsilon}^{\varepsilon_{\max}} \sigma^1(\varepsilon', \varepsilon) \psi^1(\varepsilon', x) d\varepsilon' - \sigma_T(\varepsilon) \psi^1(\varepsilon, x), \tag{3b}$$

$$\sigma^0(\varepsilon', \varepsilon) = 2\pi \int_{-1}^{+1} \sigma(\varepsilon', \varepsilon, \mu) d\mu, \qquad \sigma^1(\varepsilon', \varepsilon) = 2\pi \int_{-1}^{+1} \mu \sigma(\varepsilon', \varepsilon, \mu) d\mu.$$

The system (3) has more unknowns than equations. In order to solve such an undetermined system, one typically closes it by expressing the moment ψ^2 as a function of ψ^0 and ψ^1. For the present application, the entropy-based closure [13] was preferred as it provides desirable properties (hyperbolicity, entropy decay, correct modeling of beams). This closure, leading to the so-called M_1 closure, consists in defining ψ^2 as the second-order moment of the ansatz ψ_{M_1} minimizing Boltzmann entropy under the following constraints

$$\psi^2 = \int_{S^2} \Omega \otimes \Omega \psi_{M_1}(\Omega) d\Omega, \tag{4}$$

$$\psi_{M_1} = \underset{f \in \mathscr{C}(\psi^0, \psi^1)}{\operatorname{argmin}} \; f \log f - f, \tag{5}$$

$$\mathscr{C}(\psi^0, \psi^1) = \left\{ f \in L^1(S^2), \quad f \geq 0, \int_{S^2} f(\Omega) d\Omega = \psi^0, \int_{S^2} \Omega f(\Omega) d\Omega = \psi^1 \right\}.$$

The ansatz ψ_{M_1} can be proved to have the form [3, 9, 12, 17]

$$\psi_{M_1} = \exp\left(\boldsymbol{\Lambda} . \mathbf{m}(\Omega)\right), \tag{6}$$

where $\mathbf{m}(\Omega) = (1, \Omega_1, \Omega_2, \Omega_3)^T$ and $\Lambda \in \mathbb{R}^4$. However, the minimization problem (5) has a solution if and only if the set $\mathscr{C}(\psi^0, \psi^1)$ is non-empty.

Proposition 1. ([10]) *The problem (5) has a solution if and only if the moments* (ψ^0, ψ^1) *are in the realizability domain* $\mathscr{R}_{\mathbf{m}}$ *characterized by*

$$\mathscr{R}_{\mathbf{m}} = \left\{(\psi^0, \psi^1) \in \mathbb{R} \times \mathbb{R}^3, \quad s.t. \quad \psi^0 > |\psi^1|\right\}. \tag{7}$$

For writing purposes, one rewrites (3) under the form

$$\nabla_x . F(\mathbf{\Psi})(\varepsilon, x) = \int_{\varepsilon}^{\varepsilon_{\max}} \Sigma(\varepsilon)\mathbf{\Psi}(\varepsilon, x)d\varepsilon - \sigma_T(\varepsilon)\mathbf{\Psi}(\varepsilon, x), \tag{8}$$

$$\mathbf{\Psi} = (\psi^0, \psi^1)^T \equiv \int_{S^2} \mathbf{m}(\Omega)\psi_{M_1}(\Omega)d\Omega, \qquad \Sigma = \begin{pmatrix} \sigma^0 & 0_{\mathbb{R}^3}^T \\ 0_{\mathbb{R}^3} & \sigma^1 Id \end{pmatrix},$$
$$F = (\psi^1, \psi^2)^T \equiv \int_{S^2} \Omega \otimes \mathbf{m}(\Omega)\psi_{M_1}(\Omega)d\Omega.$$

Writing the ansatz ψ_{M_1} under the form (6), one proves that the fluxes F are those of a symmetric hyberbolic equation [7, 12].

3 Numerical Approach

In order to handle the nonlinearity in (8), the relaxation approach proposed in [16] and based on the previous work of [1, 14] is used.

3.1 Relaxation Method

The relaxation approximation leads to studying linear equations instead of (8). Let us chose J directions of relaxation $\lambda_i \in \mathbb{R}^N$ and equilibrium functions $\mathbf{M}_i(\mathbf{\Psi})$ commonly called Maxwellians. With those relaxation parameters, define the relaxed equations for (8)

$$\lambda_i . \nabla_x \mathbf{f}_i^\tau(\varepsilon, x) - \left(\int_{\varepsilon}^{\varepsilon_{\max}} \Sigma(\varepsilon)\mathbf{f}_i^\tau(\varepsilon, x)d\varepsilon - \sigma_T(\varepsilon)\mathbf{f}_i^\tau(\varepsilon, x)\right) = \frac{\mathbf{M}_i\left(\sum_{i=1}^{J} \mathbf{f}_i^\tau\right) - \mathbf{f}_i^\tau}{\tau} \tag{9}$$

In [1, 14] , the authors showed for similar equations that

$$\lim_{\tau \to 0} \sum_{i=1}^{J} \mathbf{f}_i^\tau = \mathbf{\Psi},$$

where the $\mathbf{f}_j^\tau$ solve (9), and under the conditions

$$\forall n \in S^2, \qquad Sp(\partial_{\mathbf{\Psi}} \mathbf{F}_n(\mathbf{\Psi})) \subset \left[\min_{i=1,\dots,J} \lambda_j.n, \max_{i=1,\dots,J} \lambda_j.n \right], \tag{10a}$$

$$\sum_{i=1}^{J} \mathbf{M}_i(\mathbf{\Psi}) = \mathbf{\Psi}, \qquad \sum_{i=1}^{J} \lambda_i \otimes \mathbf{M}_i(\mathbf{\Psi}) = F(\mathbf{\Psi}), \tag{10b}$$

where $\mathbf{F}_n = (\psi^1.n, \psi^2 n)$. For the present applications, we also require that the Maxwellians $\mathbf{M}_i : \mathscr{R}_{\mathbf{m}} \to \mathscr{R}_{\mathbf{m}}$ are realizable.

As a first approach, one may use the following two propositions to define relaxation parameters.

Proposition 2. ([2]) *The eigenvalues of the Jacobian of the fluxes are bounded by 1*

$$\forall n \in S^2, \qquad Sp(\partial_{\mathbf{\Psi}} \mathbf{F}_n(\mathbf{\Psi})) \subset [-1, 1].$$

Proposition 3. ([16]) *The following vector is realizable*

$$\forall n \in S^2, \quad \forall \mathbf{\Psi} \in \mathscr{R}_{\mathbf{m}}, \qquad \mathbf{\Psi} + \mathbf{F}_n(\mathbf{\Psi}) \in \mathscr{R}_{\mathbf{m}}.$$

As a first approach, we propose to chose $2N$ directions (N being the number of space dimensions) of relaxations and to define

$$\lambda_i = Ne_i, \quad \lambda_{i+N} = -Ne_i, \quad \mathbf{M}_i = \frac{\mathbf{\Psi} + \mathbf{F}_{e_i}(\mathbf{\Psi})}{N}, \quad \mathbf{M}_{i+N} = \frac{\mathbf{\Psi} - \mathbf{F}_{e_i}(\mathbf{\Psi})}{N}. \tag{11}$$

One verifies using the last two propositions that those parameters verify (10).

3.2 A Numerical Scheme for 2D Equations

In the following, we focus on a 2D problem ($N = 2$) and the notations are adapted to 2D equations. However, the method can straightforwardly be extended to 3D problems. The superscript n refers to the energy step ε^n, and the subscripts l and m refer to the position $x_{l,m}$, respectively, according to the first and second Cartesian axes e_1 and e_2. A numerical scheme for (8) is obtained using a splitting method on (9).

1. At the entry ε^n of each energy cell, the values of $\mathbf{f}_i^n$ are initialized at the values of the associated Maxwellians $\mathbf{M}_i(\mathbf{\Psi}^n)$.
2. Then one solves the homogeneous relaxed equation

$$\lambda_i.\nabla_x \mathbf{f}_i(\varepsilon, x) - \left(\int_{\varepsilon}^{\varepsilon_{\max}} \Sigma(\varepsilon) \mathbf{f}_i(\varepsilon, x) d\varepsilon - \sigma_T(\varepsilon) \mathbf{f}_i(\varepsilon, x) \right) = 0 \tag{12}$$

on the interval $[\varepsilon^{n+1}, \varepsilon^n]$, i.e. one computes $\mathbf{f}_i^{n+1}$.

3. Finally, the influence of the relaxation term is added, which corresponds, when $\tau \to 0$, to fixing the new value

$$\boldsymbol{\Psi}^{n+1} = \sum_{i=1}^{J} \mathbf{f}_i^{n+1}. \tag{13}$$

One only needs to construct a numerical scheme for (12). Using a simple upwind discretization for the spatial flux and a quadrature for the integral term together with (13) leads to define the following scheme for $\boldsymbol{\Psi}$

$$\frac{\mathbf{F_1}^n_{l+\frac{1}{2},m} - \mathbf{F_1}^n_{l-\frac{1}{2},m}}{\Delta x} + \frac{\mathbf{F_2}^n_{l,m+\frac{1}{2}} - \mathbf{F_2}^n_{l,m-\frac{1}{2}}}{\Delta y} - \left(\sum_{n'=1}^{n} \Sigma^{n',n} \boldsymbol{\Psi}^{n'}_{l,m} \Delta\varepsilon^{n'} - \sigma_T^n \boldsymbol{\Psi}^n_{l,m} \right) = 0, \tag{14}$$

where the fluxes are of HLL [8] type

$$\mathbf{F_1}^n_{l+\frac{1}{2},m} = \frac{1}{2}\left[\mathbf{F_1}(\boldsymbol{\Psi}^n_{l+1,m}) + \mathbf{F_1}(\boldsymbol{\Psi}^n_{l,m}) + (\boldsymbol{\Psi}^n_{l+1,m} - \boldsymbol{\Psi}^n_{l,m})\right],$$
$$\mathbf{F_2}^n_{l,m+\frac{1}{2}} = \frac{1}{2}\left[\mathbf{F_2}(\boldsymbol{\Psi}^n_{l,m+1}) + \mathbf{F_2}(\boldsymbol{\Psi}^n_{l,m}) + (\boldsymbol{\Psi}^n_{l,m+1} - \boldsymbol{\Psi}^n_{l,m})\right].$$

Recall that Eq. (12) is solved backwardly in energy. An iterative solver was proposed in [15] to compute $\boldsymbol{\Psi}^n_{l,m}$ at each iteration, and a complete analysis of this scheme is postponed to a future paper.

4 A Numerical Experiment

This test case corresponds to injecting a beam of photons in a 2D medium. The size of the medium is 2 cm $\times$ 10 cm, and the beam is 0.5 cm large and composed of 500 keV photons. This is modeled by the following kinetic boundary condition

$$\psi(x, \varepsilon, \Omega) = 10^{10} \exp\left(-\alpha_\varepsilon (\varepsilon - \varepsilon_0)^2\right) \exp\left(-\alpha_\mu (\Omega_1 - 1)^2\right) \mathbf{1}_B(x), \text{ for } n.\Omega < 0,$$
$$B = \left\{ x = (x_1, x_2), \qquad x_1 = 0, \qquad x_2 \in [0.75 \text{ cm}, 1.25 \text{ cm}] \right\}.$$

where $\mathbf{1}_B$ is the indicator function in the set B, n is the outgoing normal, with $\varepsilon_0 = 500$ keV, and the constants $\alpha_\varepsilon = 20000$ and $\alpha_\mu = 10000$ are chosen arbitrarily large to model a beam. At the discrete level for the moment models, we fix

$$\boldsymbol{\Psi}^n_{0,m} = 10^{10} \exp\left(-\alpha_\varepsilon \left(\varepsilon^n - \varepsilon_0\right)^2\right) \int_{S^2} \mathbf{m}(\Omega) \exp\left(-\alpha_\mu (\Omega_1 - 1)^2\right) d\Omega \mathbf{1}_B(x_{l,m}),$$
$$\boldsymbol{\Psi}^n_{l,0} = \boldsymbol{\Psi}^n_{l_{\max},m} = \boldsymbol{\Psi}^n_{l,m_{\max}} = 0_{\mathbb{R}^4}.$$

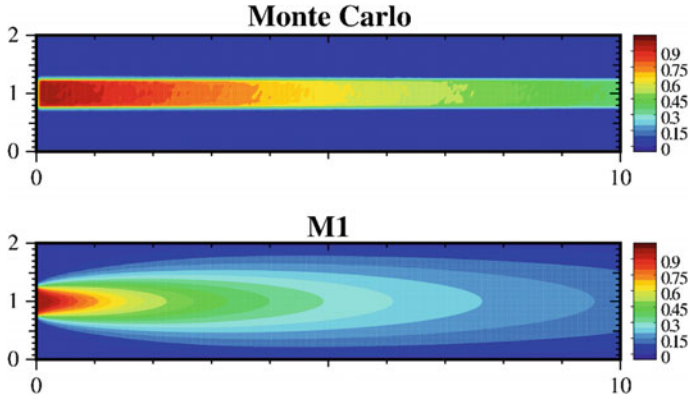


Fig. 1 Density ρ obtained with the Monte Carlo solver (top) and the M_1 model (below)

The density of particles ρ is given by the formula

$$\rho(x) = \int_{\varepsilon_{\min}}^{\varepsilon_{\max}} \psi^0(x, \varepsilon) d\varepsilon.$$

It is represented in Fig. 1 normalized by the maximum density $\rho_{\max}$, computed using the scheme (14) on the M_1 equation and compared to a reference Monte Carlo result.

One observes on the Monte Carlo results in Fig. 1 that the photons travel through the medium almost in straightline. They are only rarely scattered and the beam remains sharp deep in the medium. However, the M_1 results with the scheme (14) are significantly different. One may observe a beam shape; however, this beam is very diffused especially in the direction transverse to the beam. This effect is actually a numerical artifact due to the relaxation parameters chosen (11), and the next section is devoted to correcting it.

5 A Correction to Accurately Model Transverse Diffusion

In practice, the relaxation method can be used under the stability condition (10) on the relaxation speeds. However, the relaxation method is known to be overdiffusive when the relaxation speeds λ_j are too large. In practice, the larger is the λ_i the more stable is the resulting scheme, but the more diffusive it is. This can be verified by reproducing the computations of [1, 14]. The present correction uses similar ideas as the ones proposed in [2, 6].

5.1 Bounds on the Eigenvalues of the Jacobian of the Flux

The relaxation speeds were chosen to be of norm $|\lambda_j| = N$ which was enough to satisfy (10) according to Proposition 2. However, those bounds are too large when $\boldsymbol{\Psi}$ is the moment vector of a beam. Consider a beam of direction e_1 modeled by

$$\boldsymbol{\Psi} = \int_{S^2} \mathbf{m}(\Omega) \exp(-\alpha_\mu(\Omega_1 - 1))d\Omega.$$

The spectral radius of the Jacobian of the flux $\mathbf{F_2}$ transverse to the direction of the beam is zero. Indeed, using the even and odd character of the following functions, one finds that for all $V \in \mathbb{R}^4$

$$V^T \partial_{\boldsymbol{\Lambda}} \mathbf{F_2}(\boldsymbol{\Psi}) V = \int_{S^2} \Omega_2 (V.\mathbf{m}(\Omega))^2 \exp(-\alpha_2(\Omega_1 - 1))d\Omega = 0,$$
$$V^T \partial_{\boldsymbol{\Lambda}} \boldsymbol{\Psi} V = \int_{S^2} (V.\mathbf{m}(\Omega))^2 \exp(-\alpha_\mu(\Omega_1 - 1))d\Omega > 0.$$

This means that the eigenvalues of $\partial_{\boldsymbol{\Lambda}} \mathbf{F_2}(\boldsymbol{\Psi})$ are all zeros while $\partial_{\boldsymbol{\Lambda}} \boldsymbol{\Psi}$ is strictly positive. Therefore, the eigenvalues of the Jacobian $\partial_{\boldsymbol{\Psi}} \mathbf{F_2} = \partial_{\boldsymbol{\Lambda}} \mathbf{F_2}(\boldsymbol{\Psi})(\partial_{\boldsymbol{\Lambda}} \boldsymbol{\Psi})^{-1}$ of the transverse flux are all zero.

Those eigenvalues can actually be computed analytically (as the Jacobian of the flux is a 4×4 matrix, those are the roots of a quartic). For the present numerical purpose, we only compute bounds on those eigenvalues that are easily computable and implementable. For writing consideration, those computations are gathered in Appendix. In the rest of this paper, the minimum and maximum of $Sp(\partial_{\boldsymbol{\Psi}} \mathbf{F_n})$ are called b_- and b_+ and are given functions of the direction n of the flux $\mathbf{F_n}$ and of the normalized first-order moment

$$N^1 = \frac{\psi^1}{\psi^0}.$$

5.2 The Modified Relaxation Parameters

Based on those bounds, we propose to modify the relaxation parameters (11) into

$$\lambda_1 = (b_1, 0)\,, \quad \lambda_2 = (b_2, 0)\,, \quad \lambda_3 = (0, b_3)\,, \quad \lambda_4 = (0, b_4)\,, \tag{15a}$$

$$\mathbf{M_1} = \frac{|b_1|}{|b_1| + |b_2|}\left(\frac{\boldsymbol{\Psi}}{2} + \frac{\mathbf{F_1}}{|b_1|}\right), \quad \mathbf{M_2} = \frac{|b_2|}{|b_1| + |b_2|}\left(\frac{\boldsymbol{\Psi}}{2} + \frac{\mathbf{F_1}}{|b_2|}\right), \tag{15b}$$
$$\mathbf{M_3} = \frac{|b_3|}{|b_3| + |b_4|}\left(\frac{\boldsymbol{\Psi}}{2} + \frac{\mathbf{F_2}}{|b_3|}\right), \quad \mathbf{M_4} = \frac{|b_4|}{|b_3| + |b_4|}\left(\frac{\boldsymbol{\Psi}}{2} + \frac{\mathbf{F_2}}{|b_4|}\right).$$

The coefficients b_i can still be chosen such that the parameters (15) satisfy (10) and such that they are smaller than those in (11).

Recall that we also required the Maxwellians $\mathbf{M_i} \in \mathscr{R}_\mathbf{m}$ to be realizable. In practice, this leads to an additional requirement on the bounds b_1, b_2, b_3, and b_4. For the M_1 model, those requirements can easily be computed using (7), for $\mathbf{M_1}$ it reads

$$\left(\frac{1}{2} + \frac{N_1^1}{|b_1|}\right)^2 > \left|\frac{1}{2}N^1 + \frac{N^2.e_1}{|b_1|}\right|^2.$$

Solving this quadratic inequality leads to chose b_1 such that

$$|b_1| > \max\left(0, b_{\min}\left(N^1, e_1\right)\right), \qquad b_{\min}(N^1, n) := \frac{-\beta_n - \sqrt{\beta_n^2 - \alpha\gamma_n}}{\alpha},$$
$$\alpha = \frac{1 - |N^1|^2}{4}, \quad \beta_n = \frac{1}{2}\left(N^1.n - N^1.(N^2 n)\right), \quad \gamma_n = (N^1.n)^2 - |N^2 n|^2.$$

Similar computations hold for b_2, b_3, and b_4 which lead to fixing the bounds

$$\begin{aligned}
b_1(\mathbf{\Psi}) &= \max\left(10^{-8}, \quad b_+(N^1, e_1), \quad b_{\min}(N^1, e_1)\right),\\
b_2(\mathbf{\Psi}) &= \min\left(-10^{-8}, \quad b_-(N^1, e_1), \quad -b_{\min}(N^1, -e_1)\right),\\
b_3(\mathbf{\Psi}) &= \max\left(10^{-8}, \quad b_+(N^1, e_2), \quad b_{\min}(N^1, e_2)\right),\\
b_4(\mathbf{\Psi}) &= \min\left(-10^{-8}, \quad b_-(N^1, e_2), \quad -b_{\min}(N^1, -e_2)\right),
\end{aligned}$$

where the constants $\pm 10^{-8}$ are chosen arbitrarily low to avoid divisions by zero and e_1 and e_2 are the Cartesian axes.

5.3 The New Numerical Scheme

Using the relaxation parameters (15) to construct a scheme for (3) leads to rewrite the fluxes of the form ([6])

$$\mathbf{F}_{l+\frac{1}{2},m}^{n+1} = \frac{1}{|b_{1,l+\frac{1}{2},m}^{n+1}| + |b_{2,l+\frac{1}{2},m}^{n+1}|}\Big[|b_{2,l+\frac{1}{2},m}^{n+1}|\mathbf{F_1}(\mathbf{\Psi}_{l+1,m}^{n+1}) + |b_{1,l+\frac{1}{2},m}^{n+1}|\mathbf{F_1}(\mathbf{\Psi}_{l,m}^{n+1}) \tag{16a}$$
$$+|b_{1,l+\frac{1}{2},m}^{n+1} b_{2,l+\frac{1}{2},m}^{n+1}|(\mathbf{\Psi}_{l+1,m}^{n+1} - \mathbf{\Psi}_{l,m}^{n+1})\Big],$$

$$\mathbf{F}^{n+1}_{l,m+\frac{1}{2}} = \frac{1}{|b^{n+1}_{3,l,m+\frac{1}{2}}| + |b^{n+1}_{4,l,m+\frac{1}{2}}|}\Big[|b^{n+1}_{4,l,m+\frac{1}{2}}|\mathbf{F_2}(\boldsymbol{\Psi}^{n+1}_{l,m+1}) + |b^{n+1}_{3,l,m+\frac{1}{2}}|\mathbf{F_2}(\boldsymbol{\Psi}^{n+1}_{l,m}) + |b^{n+1}_{3,l,m+\frac{1}{2}} b^{n+1}_{4,l,m+\frac{1}{2}}|(\boldsymbol{\Psi}^{n+1}_{l,m+1} - \boldsymbol{\Psi}^{n+1}_{l,m})\Big], \quad (16b)$$

$$b^{n+1}_{1,l+\frac{1}{2},m} = \max\left(b_1(\boldsymbol{\Psi}^{n+1}_{l,m}), b_1(\boldsymbol{\Psi}^{n+1}_{l+1,m})\right), \quad b^{n+1}_{2,l+\frac{1}{2},m} = \min\left(b_2(\boldsymbol{\Psi}^{n+1}_{l,m}), b_2(\boldsymbol{\Psi}^{n+1}_{l+1,m})\right),$$

$$b^{n+1}_{3,l,m+\frac{1}{2}} = \max\left(b_3(\boldsymbol{\Psi}^{n+1}_{l,m}), b_3(\boldsymbol{\Psi}^{n+1}_{l+1,m})\right), \quad b^{n+1}_{4,l,m+\frac{1}{2}} = \min\left(b_4(\boldsymbol{\Psi}^{n+1}_{l,m}), b_4(\boldsymbol{\Psi}^{n+1}_{l+1,m})\right)$$

in the scheme (14). The numerical fluxes are now defined locally as a function of the unknowns and the fluxes which allows to better capture the diffusion effects. One may verify that the coefficients $|b^{n+1}_{1,l+\frac{1}{2},m} b^{n+1}_{2,l+\frac{1}{2},m}|$ before the terms $(\boldsymbol{\Psi}^{n+1}_{l+1,m} - \boldsymbol{\Psi}^{n+1}_{l,m})$, responsible for the numerical diffusion, in the definition of the numerical fluxes (16) are lower than the one in the scheme (14).

5.4 Results with the Modified Scheme

Using this modified scheme on the test case of Sect. 4 provides the dose result in Fig. 2 with the computational times in Table 1.

The results with the modified relaxation parameters are much closer to the reference Monte Carlo results. The diffusion in the transverse direction is much lower than the one in Fig. 1, and the beam stays sharp through the medium.

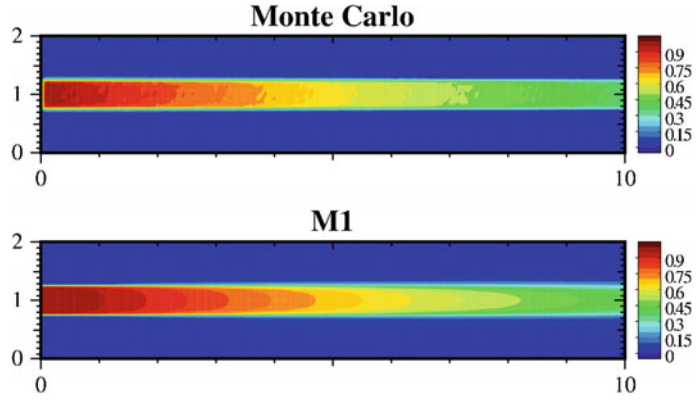


Fig. 2 Density ρ obtained with the Monte Carlo solver (top) and the M_1 model (below) with the modified relaxation parameters

Table 1 Computational times with the different numerical methods

Solver	Monte Carlo	M_1 solver	Modified M_1 solver
Computation times	14 h	49.78699 s	204.1239 s

The computational cost is considerably lower with the present numerical method compared to the reference Monte Carlo code. The computational time is although higher when using the new relaxation parameters (15) than when using the one in (11). This is actually due to the method used to compute $\mathbf{\Psi}_{l,m}^n$ from (14) or (16). The conditioning of problem (14) is simply better than the one of (16) which explains the difference of computational times.

6 Conclusion

A numerical method for the transport of photons in a human-sized medium was proposed. This method aims to solve M_1 equations. When modeling a beam of photons in such a medium, the main difficulties of those simulations are due to the fact that the diffusion phenomena in the direction of the beam and the ones normal to the beam have considerably different magnitude. Standard methods present a numerical diffusion which is considerably overestimated in the transverse direction. This effect affects the accuracy of the simulation. First a numerical scheme was proposed for solving the transport equation, then it was improved to accurately capture the diffusion effect in the transverse direction.

Acknowledgements The author would like to acknowledge K. Küpper and G. Birindelli for performing the Monte Carlo simulations used as reference results in this paper.

Appendix : Computation of Bounds on the Eigenvalues of the Jacobian of the Flux

Using rotational invariance and a normalization (see [11] for details), the closure (4) can also be rewritten under the form

$$\psi^2 = \psi^0 \left[\frac{1-\chi}{2} Id + \frac{3\chi - 1}{2} \frac{\psi^1 \otimes \psi^1}{|\psi^1|^2} \right], \tag{17}$$

where the Eddington factor χ is a scalar function of the scalar $|\psi^1|/\psi^0$.

Consider that ψ^1 is colinear to e_1 (otherwise just use a rotation to work in such a reference frame). Using the form (17) of the closure, the fluxes $\mathbf{F_n}$ in the direction $n \in S^2$ read

$$\boldsymbol{\Psi} = \left(\psi^0,\ \psi^1\right),$$
$$\forall n \in S^2, \qquad \mathbf{F_n}(\boldsymbol{\Psi}) = \left(\psi^1.n,\ \frac{\psi^0}{2}\left[(1-\chi)n + (3\chi-1)\frac{(\psi^1.n)\psi^1}{|\psi^1|^2}\right]\right),$$

Chose a reference frame such that $\psi^1 = \psi_1^1 e_1$ with $\psi_1^1 \geq 0$. In this reference frame, the spectrum of the Jacobian of the flux $\mathbf{F_1}$ along the direction $n = e_1$ (direction of the beam) and along $n = e_2$ (direction normal to the beam) read

$$Sp\left(\partial_{\boldsymbol{\Psi}}\left(\mathbf{F_1}(\boldsymbol{\Psi})\right)\right) = \left(\frac{3\chi-1}{2N_1^1},\ \frac{\chi' \pm \sqrt{\chi'^2 + 4(\chi - N_1^1\chi')}}{2}\right),$$
$$Sp\left(\partial_{\boldsymbol{\Psi}}\left(\mathbf{F_2}(\boldsymbol{\Psi})\right)\right) = \left(0,\ \pm\sqrt{\frac{1-\chi+N_1^1\chi' - \frac{3\chi-1}{2N_1^1}\chi'}{2}}\right).$$

Now, in order to come back to the computations in any reference frame, one can simply use a rotation R such that $R\psi^1 = \psi_1^1 e_1$. One has

$$\forall n \in S^2, \qquad \begin{aligned}\partial_{\boldsymbol{\Psi}}\left(\mathbf{F_n}(\boldsymbol{\Psi})\right) &= \partial_{(\psi^0, R\psi_1^1 e_1)}\left(\mathbf{F_n}(\psi^0, R\psi_1^1 e_1)\right) \\ &= R_2 \partial_{\boldsymbol{\Psi}} \mathbf{F}_{\mathbf{R^T n}}\left(\psi^0, |\psi^1| e_1\right) R_2^T,\end{aligned} \qquad R_2 = \begin{pmatrix} 1 & 0_{\mathbb{R}^3} \\ 0_{\mathbb{R}^3} & R \end{pmatrix}.$$

The spectrum of such a matrix can be bounded using the previous computations

$$\forall n \in S^2, \qquad Sp\left(\partial_{\boldsymbol{\Psi}}\left(\mathbf{F_n}(\boldsymbol{\Psi})\right)\right) \subset [b_-, b_+], \tag{18a}$$
$$b_-(N^1, n) = (1-\theta)\min S_t(|N^1|) + \theta \min S_n(|N^1|), \tag{18b}$$

$$b_+(N^1, n) = (1-\theta)\max S_t(|N^1|) + \theta \max S_n(|N^1|), \qquad \theta = \frac{N^1.n}{|N^1|}.$$

The exact bounds b_- and b_+ of $Sp\left(\partial_{\boldsymbol{\Psi}} \mathbf{F_n}(\boldsymbol{\Psi})\right)$ could be computed analytically as the eigenvalues of the 4×4 matrix $\partial_\Psi\left(\mathbf{F_n}(\Psi)\right)$. However, using such analytical formulae may introduce errors at the numerical level that may be non-negligible. Computing the bounds in (18) is easier, and they are sufficient for the present applications.

References

1. D. Aregba-Driollet, R. Natalini, Discrete kinetic schemes for multidimensional systems of conservation laws. SIAM J. Numer. Anal. **6**, 1973–2004 (2000)
2. C. Berthon, P. Charrier, B. Dubroca, An HLLC scheme to solve the M_1 model of radiative transfer in two space dimensions. J. Sci. Comput. **31**(3), 347–389 (2007)
3. J. Borwein, A. Lewis, Duality relationships for entropy-like minimization problems. SIAM J. Control Optim. **29**(2), 325–338 (1991)
4. C.M. Davisson, R.D. Evans, Gamma-ray absorption coefficients. Rev. Mod. Phys. (1952)

5. R. Duclous, B. Dubroca, M. Frank, A deterministic partial differential equation model for dose calculation in electron radiotherapy. Phys. Med. Biol. **55**, 3843–3857 (2010)
6. B. Einfeldt, On godunov-type methods for gas dynamics. SIAM J. Numer. Anal. **25**(2), 294–318 (1988)
7. K.O. Friedrichs, P.D. Lax, Systems of conservation equations with a convex extension. Proc. Natl. Acad. Sci. **68**(8), 1686–1688 (1971)
8. A. Harten, P. Lax, B. Van Leer, On upstream differencing and Gudonov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**(1), 35–61 (1983)
9. M. Junk, Maximum entropy for reduced moment problems. Math. Mod. Meth. Appl. S. **10**(1001–1028), 2000 (1998)
10. D. Kershaw, Flux limiting nature's own way. Technical report, Lawrence Livermore Laboratory (1976)
11. C.D. Levermore, Relating Eddington factors to flux limiters. J. Quant. Spectros. Radiat. Transf. **31**, 149–160 (1984)
12. C.D. Levermore, Moment closure hierarchies for kinetic theories. J. Stat. Phys. **83**(5–6), 1021–1065 (1996)
13. G.N. Minerbo, Maximum entropy Eddington factors. J. Quant. Spectros. Radiat. Transf. **20**, 541–545 (1978)
14. R. Natalini, A discrete kinetic approximation of entropy solutions to multidimensional scalar conservation laws. J. Differ. Equ. **148**(2), 292–317 (1998)
15. T. Pichard, G.W. Alldredge, S. Brull, B. Dubroca, M. Frank, An approximation of the M_2 closure: application to radiotherapy dose simulation. J. Sci. Comput. (to appear)
16. T. Pichard, D. Aregba-Driollet, S. Brull, B. Dubroca, M. Frank, Relaxation schemes for the M_1 model with space-dependent flux: application to radiotherapy dose calculation. Commun. Comput. Phys. **19**, 168–191 (2016)
17. J. Schneider, Entropic approximation in kinetic theory. ESAIM-Math. Model. Num. **38**(3), 541–561 (2004)
18. C.P. South, M. Partridge, P.M. Evans, A theoretical framework for prescribing radiotherapy dose distributions using patient-specific biological information. Med. Phys. **35**(10), 4599–4611 (2008)

Numerical Viscosity in Large Time Step HLL-Type Schemes

Marin Prebeg

Abstract We consider Large Time Step (LTS) methods, i.e., the explicit finite volume methods not limited by the Courant–Friedrichs–Lewy (CFL) condition. The original LTS method (LeVeque in SIAM J Numer Anal 22, 1985) was constructed as an extension of the Godunov scheme, and successive versions have been developed in the framework of Roe's approximate Riemann solver. Recently, Prebeg et al. (in ESAIM: M2AN, in press, 2017) developed the LTS extension of the HLL and HLLC schemes. We perform the modified equation analysis and demonstrate that for the appropriate choice of the wave velocity estimates, the LTS-HLL scheme yields entropy-satisfying solutions. We apply the LTS-HLL(C) schemes to the one-dimensional Euler equations and consider the Sod shock tube, double rarefaction, and Woodward–Colella blast-wave problem.

Keywords Large Time Step · HLL · Entropy violation · The Euler equations Hyperbolic conservation laws

1 Introduction

We consider the hyperbolic system of conservation laws:

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x = 0, \tag{1a}$$

$$\mathbf{U}(x, 0) = \mathbf{U}_0(x), \tag{1b}$$

where $\mathbf{U} \in \mathbb{R}^m$, $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$, $x \in \mathbb{R}$, and $t \in \mathbb{R}^+$. We are interested in solving (1) with an explicit finite volume method not limited by the Courant–Friedrichs–Lewy (CFL) condition.

M. Prebeg (✉)
Department of Energy and Process Engineering, Norwegian University of Science and Technology, Kolbjørn Hejes vei 2, 7491 Trondheim, Norway
e-mail: marin.prebeg@gmail.com

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_36

A class of such methods has been proposed by LeVeque [1–3]. Therein, the Godunov scheme was extended to the LTS-Godunov and LTS-Roe schemes and applied to the one-dimensional Euler equations. Most recent applications of these ideas include shallow water equations (Murillo, Morales-Hernández and co-workers [4–8] and Xu et al. [9]), three-dimensional Euler equations (Qian and Lee [10]), high-speed combustion waves (Tang et al. [11]), Maxwell's equations (Makwana and Chatterjee [12]), and two-phase flows (Lindqvist and Lund [13] and Prebeg et al. [14]). All the methods discussed above share the feature of starting from a Godunov- or Roe-type Riemann solver and extending it to the LTS framework. In addition to these applications, Lindqvist et al. [15] studied the TVD properties of LTS methods and introduced the LTS-Lax-Friedrichs scheme. Several authors [1, 3, 5, 9, 10, 13, 15] reported that the LTS-Roe scheme yields entropy-violating solutions even more often than the standard Roe scheme. Therein, this issue is solved by splitting the rarefaction wave into several expansion shocks [1, 3, 5, 9, 10] or by varying the time step [13, 15].

Prebeg et al. [16] developed the LTS extension of the Harten–Lax–van Leer (HLL) [17–19] and HLL–Contact (HLLC) [20] schemes and applied them to a one-dimensional Euler equations. They observed that the LTS-HLL(C) schemes with the wave velocity estimates according to Einfeldt [18] yield entropy-satisfying solutions. This observation motivates the present paper, which is structured as follows: In Sect. 2 we outline the problem and the numerical methods we will consider, most importantly the LTS-HLL(C) schemes; in Sect. 3, we discuss the entropy violation associated with the LTS methods and use the modified equation analysis to demonstrate that the LTS-HLL scheme with the choice of the wave velocities estimates according to Einfeldt [18] yields entropy-satisfying solutions; in Sect. 4, we perform numerical investigations, while in Sect. 5, we end with conclusions.

2 Preliminaries

We specify the particular hyperbolic conservation law we will investigate and outline the framework of the numerical methods we will use.

2.1 Problem Outline

As an example of (1), we consider the one-dimensional Euler equations, where

$$\mathbf{U} = (\rho, \rho u, E)^{\mathrm{T}}, \tag{2a}$$

$$\mathbf{F}(\mathbf{U}) = \left(\rho u, \rho u^2 + p, u(E + p)\right)^{\mathrm{T}}, \tag{2b}$$

where ρ, u, E, p denote the density, velocity, total energy density, and pressure, respectively. The system is closed by the definition of the total energy density, $E = \rho e + \rho u^2/2$, where e is the internal energy given by the equation of state as $e = p/(\rho(\gamma - 1))$. We use $\gamma = 1.4$ for air. We can also write (1) in a quasilinear form as

$$\mathbf{U}_t + \mathbf{A}(\mathbf{U})\mathbf{U}_x = 0, \qquad \mathbf{A}(\mathbf{U}) = \frac{\partial \mathbf{F}(\mathbf{U})}{\partial \mathbf{U}}. \tag{3}$$

We assume that the system of Eq. (3) is hyperbolic; i.e., the Jacobian matrix $\mathbf{A}$ has real eigenvalues and linearly independent eigenvectors.

2.2 Numerical Methods

We discretize (1) by the explicit Euler method in time and the finite volume method in space:

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{F}_{j+1/2}^n - \mathbf{F}_{j-1/2}^n\right), \tag{4}$$

where $\mathbf{U}_j^n$ is an approximation of the average of $\mathbf{U}$ in the cell j at time level n and $\mathbf{F}_{j+1/2}^n$ is a numerical approximation of the flux function at the cell interface $x_{j+1/2}$ at time level n. In standard (3-point) methods, the numerical flux depends only on the neighboring cell values and we may write the numerical fluxes in the numerical viscosity form:

$$\mathbf{F}_{j+1/2}^n = \frac{1}{2}\left(\mathbf{F}_j^n + \mathbf{F}_{j+1}^n\right) - \frac{1}{2}\mathbf{Q}_{j+1/2}^n\left(\mathbf{U}_{j+1}^n - \mathbf{U}_j^n\right), \tag{5}$$

where $\mathbf{F}_j^n = \mathbf{F}(\mathbf{U}_j^n)$ and $\mathbf{Q}_{j+1/2}^n$ is the numerical viscosity matrix. To simplify the notation, the time level n will be implicitly assumed in the absence of a temporal index. In the numerical viscosity framework (5), the HLL scheme is obtained by setting:

$$\mathbf{Q}_{\mathrm{HLL}} = \frac{S_{\mathrm{R}}^+ + S_{\mathrm{L}}^-}{S_{\mathrm{R}}^+ - S_{\mathrm{L}}^-}\hat{\mathbf{A}} - 2\frac{S_{\mathrm{L}}^- S_{\mathrm{R}}^+}{S_{\mathrm{R}}^+ - S_{\mathrm{L}}^-}\mathbf{I} \tag{6}$$

where $\hat{\mathbf{A}}$ is the Roe matrix [21], S_{R} and S_{L} are the wave velocity estimates, and the superscripts denote $S_{\mathrm{R}}^+ = \max(0, S_{\mathrm{R}})$ and $S_{\mathrm{L}}^- = \min(0, S_{\mathrm{L}})$. The choice of the wave velocity estimates will be addressed in Sect. 2.3. We note that $\mathbf{Q}$ can be diagonalized as

$$\mathbf{Q} = \hat{\mathbf{R}}\boldsymbol{\Omega}\hat{\mathbf{R}}^{-1}, \tag{7}$$

where $\hat{\mathbf{R}}$ is the matrix of the right eigenvectors of the Roe matrix, and $\boldsymbol{\Omega} = \mathrm{diag}\left(\omega^1, \ldots, \omega^m\right)$ is the matrix of the eigenvalues of $\mathbf{Q}$, where the superscript denotes

the particular characteristic field. Then, we may define the HLL scheme through the diagonal entries of $\boldsymbol{\Omega}$ as

$$\omega_{\mathrm{HLL}} = \frac{S_{\mathrm{R}}^{+}(\hat{\lambda} - S_{\mathrm{L}}^{-}) - S_{\mathrm{L}}^{-}(S_{\mathrm{R}}^{+} - \hat{\lambda})}{S_{\mathrm{R}}^{+} - S_{\mathrm{L}}^{-}}, \tag{8}$$

where $\hat{\lambda}$ are the eigenvalues of the Roe matrix $\hat{\mathbf{A}}$. For more details on the derivation of the HLL scheme, we refer to [17–19, 22].

For the 3-point method (5), the time step Δt is limited by the CFL condition:

$$C = \max_{p,j} |\lambda_j^p| \frac{\Delta t}{\Delta x} \leq 1, \tag{9}$$

where λ_j^p are the eigenvalues of the Jacobian matrix $\mathbf{A}(\mathbf{U}_j)$ in (3), and the superscript p denotes the particular characteristic field, $p = 1, \ldots, m$. We are interested in explicit methods not limited by the condition (9).

2.2.1 Large Time Step HLL Scheme

The natural LTS extension of the numerical viscosity formulation (5) is [15]

$$\mathbf{F}_{j+1/2} = \frac{1}{2}\left(\mathbf{F}_j + \mathbf{F}_{j+1}\right) - \frac{1}{2}\sum_{i=-\infty}^{\infty} \mathbf{Q}_{j+1/2+i}^{i}\left(\mathbf{U}_{j+1+i} - \mathbf{U}_{j+i}\right). \tag{10}$$

We note that (10) differs from [15] in the sense that we scale $\mathbf{Q}^i$ with $\Delta x/\Delta t$. By using the results from [16], we write the LTS-HLL scheme in the numerical viscosity form (10) by defining:

$$\mathbf{Q}_{j+1/2}^{i} = \left(\hat{\mathbf{R}}\boldsymbol{\Omega}^{i}\hat{\mathbf{R}}^{-1}\right)_{j+1/2}, \tag{11}$$

where the diagonal entries of $\boldsymbol{\Omega}$ are defined as

$$\omega_{\mathrm{HLL}}^{0} = \frac{S_{\mathrm{R}}^{+}(\hat{\lambda} - S_{\mathrm{L}}^{-}) - S_{\mathrm{L}}^{-}(S_{\mathrm{R}}^{+} - \hat{\lambda})}{S_{\mathrm{R}}^{+} - S_{\mathrm{L}}^{-}}, \tag{12a}$$

$$\omega_{\mathrm{HLL}}^{\mp i} = 2\frac{\hat{\lambda} - S_{\mathrm{L}}}{S_{\mathrm{R}} - S_{\mathrm{L}}} \max\left(0, \pm S_{\mathrm{R}} - i\frac{\Delta x}{\Delta t}\right) + 2\frac{S_{\mathrm{R}} - \hat{\lambda}}{S_{\mathrm{R}} - S_{\mathrm{L}}} \max\left(0, \pm S_{\mathrm{L}} - i\frac{\Delta x}{\Delta t}\right) \quad \text{for} \quad i > 0. \tag{12b}$$

We refer to [16] for the derivation of these formulae.

2.2.2 Large Time Step HLLC Scheme

The HLL scheme assumes a two-wave structure of the solution and leads to poor resolution of the contact discontinuity in the one-dimensional Euler equations (2). Toro et al. [20] introduced the HLLC solver where the missing contact wave is restored. Following [22], the main idea consists of assuming a three-wave structure of the solution, thus splitting the Riemann fan into two intermediate states:

$$\widetilde{\mathbf{U}}(x,t) = \begin{cases} \mathbf{U}_j & \text{if} \quad x < S_{\mathrm{L}} t, \\ \mathbf{U}_{\mathrm{L}}^{\mathrm{HLLC}} & \text{if} \quad S_{\mathrm{L}} t < x < S_{\mathrm{C}} t, \\ \mathbf{U}_{\mathrm{R}}^{\mathrm{HLLC}} & \text{if} \quad S_{\mathrm{C}} t < x < S_{\mathrm{R}} t, \\ \mathbf{U}_{j+1} & \text{if} \quad x > S_{\mathrm{R}} t, \end{cases} \tag{13}$$

where the intermediate states are

$$\mathbf{U}_{\mathrm{K}}^{\mathrm{HLLC}} = \rho_{\mathrm{K}} \left(\frac{S_{\mathrm{K}} - u_{\mathrm{K}}}{S_{\mathrm{K}} - S_{\mathrm{C}}} \right) \begin{bmatrix} 1 \\ S_{\mathrm{C}} \\ \frac{E_{\mathrm{K}}}{\rho_{\mathrm{K}}} + (S_{\mathrm{C}} - u_{\mathrm{K}}) \left(S_{\mathrm{C}} + \frac{p_{\mathrm{K}}}{\rho_{\mathrm{K}} (S_{\mathrm{K}} - u_{\mathrm{K}})} \right) \end{bmatrix}, \tag{14}$$

where index K denotes left (L) or right (R) state in (13). The contact discontinuity velocity is given by

$$S_{\mathrm{C}} = \frac{p_{\mathrm{R}} - p_{\mathrm{L}} + \rho_{\mathrm{L}} u_{\mathrm{L}} (S_{\mathrm{L}} - u_{\mathrm{L}}) - \rho_{\mathrm{R}} u_{\mathrm{R}} (S_{\mathrm{R}} - u_{\mathrm{R}})}{\rho_{\mathrm{L}} (S_{\mathrm{L}} - u_{\mathrm{L}}) - \rho_{\mathrm{R}} (S_{\mathrm{R}} - u_{\mathrm{R}})}. \tag{15}$$

For details on the derivation of these formulae, we refer to the book by Toro [22]. Herein, we present the LTS-HLLC scheme in the conservation form as derived in [16]. The numerical flux to be used in (4) is

$$\mathbf{F}_{j+1/2}^{\text{LTS-HLLC}} = \mathbf{F}_{j+1/2}^{0} + \sum_{i=1}^{\infty} \mathbf{F}_{j+1/2-i}^{-i} + \sum_{i=1}^{\infty} \mathbf{F}_{j+1/2+i}^{+i}, \tag{16}$$

where $\mathbf{F}_{j+1/2}^{0}$ is defined as

$$\mathbf{F}_{j+1/2}^{0} = \begin{cases} \mathbf{F}_j & \text{if} \quad 0 < S_{\mathrm{L}}, \\ \mathbf{F}_{\mathrm{L},j+1/2}^{\mathrm{HLLC}} & \text{if} \quad S_{\mathrm{L}} < 0 < S_{\mathrm{C}}, \\ \mathbf{F}_{\mathrm{R},j+1/2}^{\mathrm{HLLC}} & \text{if} \quad S_{\mathrm{C}} < 0 < S_{\mathrm{R}}, \\ \mathbf{F}_{j+1} & \text{if} \quad 0 > S_{\mathrm{R}}. \end{cases} \tag{17}$$

In the interesting case, $S_{\mathrm{L}} < 0 < S_{\mathrm{R}}$, the numerical flux function has the form:

$$\mathbf{F}_{\mathrm{L},j+1/2}^{\mathrm{HLLC}} = \mathbf{F}_j + S_{\mathrm{L}} \left(\mathbf{U}_{\mathrm{L},j+1/2}^{\mathrm{HLLC}} - \mathbf{U}_j \right), \tag{18}$$

$$\mathbf{F}_{\mathrm{R},j+1/2}^{\mathrm{HLLC}} = \mathbf{F}_{j+1} + S_{\mathrm{R}} \left(\mathbf{U}_{\mathrm{R},j+1/2}^{\mathrm{HLLC}} - \mathbf{U}_{j+1} \right). \tag{19}$$

The remaining terms in (16) are

$$\begin{aligned}\mathbf{F}^{-i}_{j+1/2-i} &= S^{-i}_{\mathrm{R},j+1/2-i}\left(\mathbf{U}^{\mathrm{HLLC}}_{\mathrm{R},j+1/2-i} - \mathbf{U}_{j+1-i}\right)\\ &+ S^{-i}_{\mathrm{C},j+1/2-i}\left(\mathbf{U}^{\mathrm{HLLC}}_{\mathrm{L},j+1/2-i} - \mathbf{U}^{\mathrm{HLLC}}_{\mathrm{R},j+1/2-i}\right)\\ &+ S^{-i}_{\mathrm{L},j+1/2-i}\left(\mathbf{U}_{j-i} - \mathbf{U}^{\mathrm{HLLC}}_{\mathrm{L},j+1/2-i}\right), \end{aligned} \tag{20}$$

$$\begin{aligned}\mathbf{F}^{+i}_{j+1/2+i} &= S^{+i}_{\mathrm{L},j+1/2+i}\left(\mathbf{U}^{\mathrm{HLLC}}_{\mathrm{L},j+1/2+i} - \mathbf{U}_{j+i}\right)\\ &+ S^{+i}_{\mathrm{C},j+1/2+i}\left(\mathbf{U}^{\mathrm{HLLC}}_{\mathrm{R},j+1/2+i} - \mathbf{U}^{\mathrm{HLLC}}_{\mathrm{L},j+1/2+i}\right)\\ &+ S^{+i}_{\mathrm{R},j+1/2+i}\left(\mathbf{U}_{j+1+i} - \mathbf{U}^{\mathrm{HLLC}}_{\mathrm{R},j+1/2+i}\right). \end{aligned} \tag{21}$$

Herein, the modified velocities are

$$S^{-i}_{[\mathrm{L,C,R}],\,j+1/2-i} = \max\left(S_{[\mathrm{L,C,R}],\,j+1/2-i} - i\frac{\Delta x}{\Delta t}, 0\right), \tag{22}$$

$$S^{+i}_{[\mathrm{L,C,R}],\,j+1/2+i} = \min\left(S_{[\mathrm{L,C,R}],\,j+1/2+i} + i\frac{\Delta x}{\Delta t}, 0\right). \tag{23}$$

We refer to [16] for the derivation of these formulae.

2.3 *Estimates for Wave Velocities S_{L} and S_{R}*

In the present paper, the choice of the wave velocity estimates is made according to Einfeldt [18]:

$$S_{\mathrm{L},j+1/2} = \min\left(\lambda^1(\mathbf{U}_j), \hat{\lambda}^1(\widehat{\mathbf{U}}_{j+1/2})\right), \tag{24a}$$

$$S_{\mathrm{R},j+1/2} = \max\left(\hat{\lambda}^3(\widehat{\mathbf{U}}_{j+1/2}), \lambda^3(\mathbf{U}_{j+1})\right), \tag{24b}$$

where $\widehat{\mathbf{U}}$ denotes the Roe average of conserved variables. The HLL scheme with (24) is usually denoted as the HLLE scheme. Einfeldt et al. [23] showed that the standard (3-point) HLLE scheme yields entropy-satisfying solutions and preserves positivity. Batten et al. [24] showed that the HLLC scheme [20] with (24) also preserves positivity. In the following section, we demonstrate that the LTS-HLLE scheme yields entropy-satisfying solutions.

3 Entropy Violation

A weak solution to a conservation law is not necessary unique [25, p. 217]. For the numerical scheme to select the physically relevant solution, we need to impose so-called *entropy conditions*. Entropy violation is most commonly associated and

discussed as it appears in the Roe scheme [21]. We start by following the same approach and consider the *numerical viscosity* interpretation of the entropy violation [25].

Consider a standard (3-point) Roe scheme written in the numerical viscosity formulation (5). The eigenvalues of the numerical viscosity matrix $\mathbf{Q}_{\text{Roe}}$ are given by

$$\omega_{\text{Roe}} = |\hat{\lambda}|. \tag{25}$$

In the transonic case, a particular eigenvalue ω^p_{Roe} $(p = 1, \ldots, m)$ may be close to zero, corresponding to no viscosity in the field p associated with the eigenvalue ω^p. We define the interface Courant number $C^p_{j+1/2} = \omega^p_{j+1/2}\Delta t/\Delta x$ and observe that if

$$C^p_{j+1/2} = 0, \tag{26}$$

we may expect an entropy violation in the particular field p. For the standard (3-point) method, these situations are well understood and we refer to [25] and references therein for a detailed discussion.

Lindqvist et al. [15] showed that for the LTS-Roe scheme, the entropy violation may also appear when

$$C^p_{j+1/2} = -i, \quad \forall i \in \mathbb{Z}. \tag{27}$$

To clarify this phenomenon and to show how it is avoided in the LTS-HLL scheme, we employ the modified equation analysis.

3.1 Modified Equation Analysis

For scalar conservation laws, Lindqvist et al. [15] showed that the LTS method (10) gives a second-order accurate approximation to the equation:

$$u_t + f(u)_x = \frac{1}{2}\Delta x \left[\frac{\Delta x}{\Delta t} \left(\sum_{i=1-k}^{k-1} \bar{Q}^i \frac{\Delta t}{\Delta x} - c^2 \right) u_x \right]_x, \tag{28}$$

where $\bar{Q}^i = Q^i(u, \ldots, u)$ is the numerical viscosity coefficient of the $(2k+1)$-point method, and $c = f'(u)\Delta t/\Delta x$. Therein, the expression:

$$D(u) = \sum_{i=1-k}^{k-1} \bar{Q}^i \frac{\Delta t}{\Delta x} - c^2, \tag{29}$$

is interpreted as the amount of numerical diffusion inherent to the scheme. In [15], $D(u)$ for the LTS-Roe scheme is determined as

$$D_{\text{LTS-Roe}} = (\lceil |c| \rceil - |c|)\,(1 + |c| - \lceil |c| \rceil)\,, \tag{30}$$

where $\lceil c \rceil = \min\{n \in \mathbb{Z} \,|\, n \geq c\}$ is the ceiling function. We may observe that D vanishes when (27) is satisfied. If the solution is supposed to be a rarefaction wave, this will lead to an entropy-violating expansion shock. We note that in [15], the modified equation (28) is defined for scalar conservation laws. Herein, we use it for systems of conservation laws by treating each characteristic field p separately.

Proposition 1. *The numerical diffusion D^p in the p-th characteristic field for the LTS-HLL scheme (11)–(12) is*

$$\begin{aligned} D^p_{\textit{LTS-HLL}} = &\frac{c - c_{\text{L}}}{c_{\text{R}} - c_{\text{L}}}\,(\lceil |c_{\text{R}}| \rceil - |c_{\text{R}}|)\,(1 + |c_{\text{R}}| - \lceil |c_{\text{R}}| \rceil) \\ &+ \frac{c_{\text{R}} - c}{c_{\text{R}} - c_{\text{L}}}\,(\lceil |c_{\text{L}}| \rceil - |c_{\text{L}}|)\,(1 + |c_{\text{L}}| - \lceil |c_{\text{L}}| \rceil) \\ &+ (c - c_{\text{L}})\,(c_{\text{R}} - c)\ , \end{aligned} \tag{31}$$

where $c_{\text{L}} = S_{\text{L}}\Delta t/\Delta x$, $c_{\text{R}} = S_{\text{R}}\Delta t/\Delta x$, and $c = \hat{\lambda}^p \Delta t/\Delta x$.

Proof. Use (12) in (29). □

Proposition 2. *If the exact solution in the p-th field is a rarefaction wave, i.e.,*

$$\lambda^p_j < \hat{\lambda}^p_{j+1/2} < \lambda^p_{j+1}, \tag{32}$$

the numerical diffusion D^p for the LTS-HLLE scheme satisfies

$$D^p_{\textit{LTS-HLLE}} > 0. \tag{33}$$

Proof. If (32) holds, Eq. (24) yields

$$S_{\text{L},j+1/2} < \hat{\lambda}^p_{j+1/2} < S_{\text{R},j+1/2}. \tag{34}$$

By using this in (31), we observe that

$$D^p_{\text{LTS-HLLE}} \geq (c - c_{\text{L}})\,(c_{\text{R}} - c) > 0. \tag{35}$$

□

Numerical investigations in the following section suggest that the above also applies to the LTS-HLLC scheme with the wave velocity estimates according to [18].

4 Results

In this section, we compare the LTS-HLL(C) schemes with their non-LTS counterparts and the LTS-Roe scheme. We note that all the results presented for LTS-HLL(C) schemes are obtained with the wave velocity estimates (24). Further, the input discretization parameters are the Courant number C and Δx. Then, the time step Δt is evaluated at each time step according to

$$\Delta t = \frac{C\Delta x}{\max\limits_{p,j} |\lambda_j^p|}. \tag{36}$$

4.1 Sod Shock Tube

We consider the Sod shock tube problem [26] with initial data:

$$\mathbf{U}(x,0) = \begin{cases} (1,0,2.5)^{\mathrm{T}} & \text{if} \quad x < 0, \\ (0.125,0,0.25)^{\mathrm{T}} & \text{if} \quad x > 0, \end{cases} \tag{37}$$

with the solution evaluated at $t = 0.4$ on a grid with 200 cells. Figure 1 shows the comparison between LTS methods. We observe that the LTS-HLL(C) schemes yield entropy-satisfying solutions, while the LTS-Roe scheme leads to an entropy violation at $x \approx -0.25$.

4.2 Double Rarefaction Problem

Next, we consider the double rarefaction test case which is often used as a benchmark test case for the positivity preserving. The initial data is

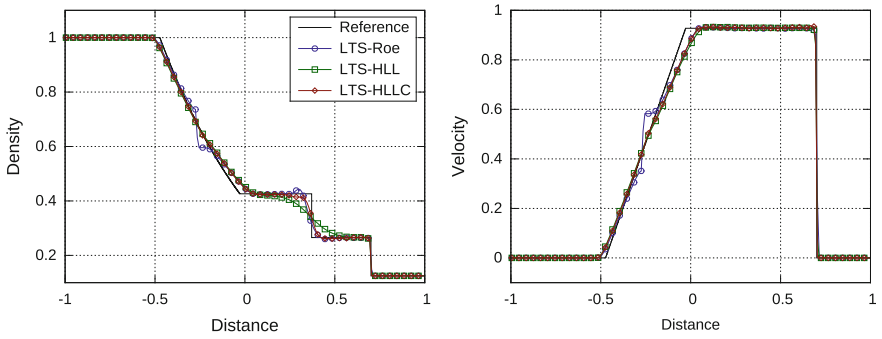


Fig. 1 Comparison between different LTS methods at $C = 3.5$ for problem (37)

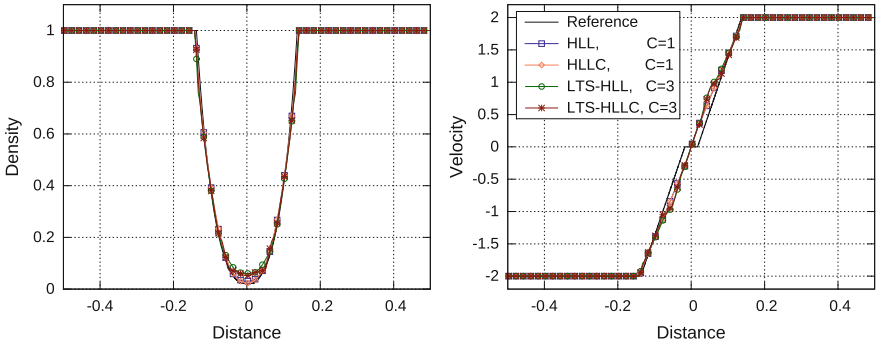


Fig. 2 Comparison between the standard HLL(C) and LTS-HLL(C) schemes for problem (38)

$$\mathbf{U}(x,0) = \begin{cases} (1,-2,1)^{\mathrm{T}} & \text{if} \quad x < 0, \\ (1,2,1)^{\mathrm{T}} & \text{if} \quad x > 0, \end{cases} \tag{38}$$

with the solution evaluated at $t = 0.05$ on a grid with 200 cells. Figure 2 shows that the LTS-HLL(C) schemes successfully handle the near-vacuum conditions. In addition, the accuracy is very close to that of the non-LTS methods.

4.3 Woodward–Colella Blast-Wave Problem

As the last test case, we consider the Woodward–Colella blast-wave problem [27]. The initial data is given by uniform density $\rho(x,0) = 1$, uniform velocity $u(x,0) = 0$, and two discontinuities in the pressure:

$$p(x,0) = \begin{cases} 1000 & \text{if} \quad 0 < x < 0.1, \\ 0.01 & \text{if} \quad 0.1 < x < 0.9, \\ 100 & \text{if} \quad 0.9 < x < 1, \end{cases} \tag{39}$$

with the solution evaluated at $t = 0.038$ on a grid with 1000 cells. The reference solution was obtained by the Roe scheme with the superbee wave limiter on the grid with 16000 cells. The boundary walls at $x = 0$ and $x = 1$ are treated as reflective boundary conditions. In Fig. 3, we can see that all LTS methods correctly capture positions of shocks and contact discontinuities. In the density plot, we observe that both the LTS-Roe and the LTS-HLLC are much more accurate than the standard HLLC scheme.

However, the LTS-Roe scheme produces an entropy violation at $x \approx 0.69$, while LTS-HLL(C) schemes do not. This can be seen in Fig. 4 where we zoomed in the area of interest in the plot for the velocity.

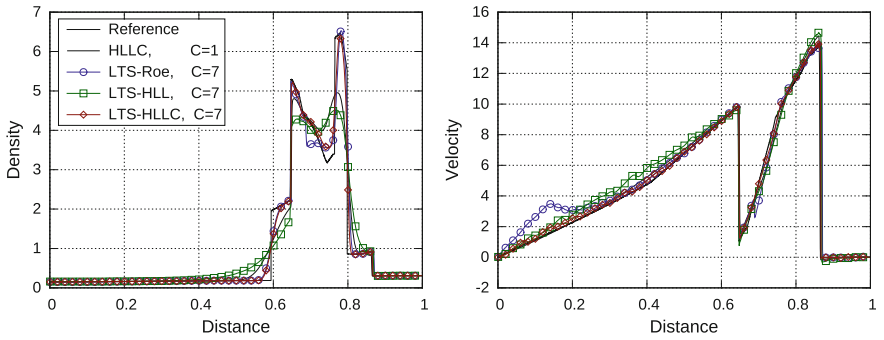
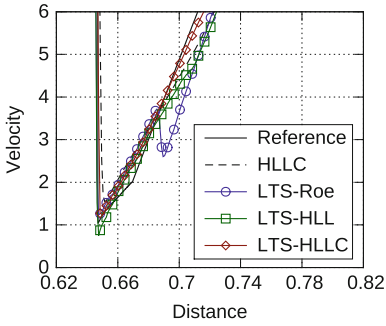


Fig. 3 Comparison between the standard HLLC and different LTS methods for problem (39)

Fig. 4 Entropy violation with the LTS-Roe scheme for problem (39)



5 Conclusions

We used the modified equation analysis to demonstrate that the LTS-HLL scheme proposed by Prebeg et al. [16] with the choice of the wave velocity estimates according to Einfeldt [18] yields entropy-satisfying solutions. We applied the scheme to the one-dimensional Euler equations and numerically demonstrated that the LTS-HLL(C) schemes with the same wave velocity choice also yield entropy-satisfying solutions. In addition, we applied both schemes to the double rarefaction test case and showed that both schemes successfully handle near-vacuum conditions.

Acknowledgements The author was supported by the Research Council of Norway (234126/30) through the SIMCOFLOW project. I am grateful to my supervisors Tore Flåtten, Bernhard Müller, and Marica Pelanti for fruitful discussions. We would like to thank the anonymous reviewer for his helpful and constructive comments, which led to an improvement of the paper.

References

1. R. LeVeque, SIAM J. Numer. Anal. **19**(6), 1091 (1982). https://doi.org/10.1137/0719080
2. R. LeVeque, Commun. Pure Appl. Math. **37**(4), 463 (1984). https://doi.org/10.1002/cpa.3160370405
3. R. LeVeque, SIAM J. Numer. Anal. **22**(6), 1051 (1985). https://doi.org/10.1137/0722063
4. J. Murillo, P. García-Navarro, P. Brufau, J. Burguete, Int. J. Numer. Meth. Fluids **50**(1), 63 (2006). https://doi.org/10.1002/fld.1036
5. M. Morales-Hernández, P. García-Navarro, J. Murillo, J. Comput. Phys. **231**(19), 6532 (2012). https://doi.org/10.1016/j.jcp.2012.06.017
6. M. Morales-Hernández, J. Murillo, P. García-Navarro, J. Burguete, in *Numerical Methods for Hyperbolic Equations*, ed. by E.V. Cendón, A. Hidalgo, P. García-Navarro, L. Cea (CRC Press, 2012), pp. 141–148. https://doi.org/10.1201/b14172-17
7. M. Morales-Hernández, M. Hubbard, P. García-Navarro, A 2D extension of a Large Time Step explicit scheme (CFL > 1) for unsteady problems with wet/dry boundaries. J. Comput. Phys. **263**, 303–327 (2014). https://doi.org/10.1016/j.jcp.2014.01.019
8. M. Morales-Hernández, A. Lacasta, J. Murillo, P. García-Navarro, Appl. Math. Model. **47**, 294 (2017). https://doi.org/10.1016/j.apm.2017.02.043
9. R. Xu, D. Zhong, B. Wu, X. Fu, R. Miao, Chinese Sci. Bull. **59**(21), 2534 (2014). https://doi.org/10.1007/s11434-014-0374-7
10. Z. Qian, C.-H. Lee, A class of Large Time Step godunov schemes for hyperbolic conservation laws and applications. J. Comput. Phys. **230**(19), 7418–7440 (2011). https://doi.org/10.1016/j.jcp.2011.06.008
11. K. Tang, A. Beccantini, C. Corre, Comput. Fluids **93**, 74 (2014). https://doi.org/10.1051/m2an:2004016
12. N.N. Makwana, A. Chatterjee, in *2015 IEEE International Conference on Computational Electromagnetics (ICCEM 2015)* (Institute of Electrical and Electronics Engineers (IEEE), 2015), pp. 330–332. https://doi.org/10.1109/COMPEM.2015.7052651
13. S. Lindqvist, H. Lund, in *VII European Congress on Computational Methods in Applied Sciences and Engineering*, ed. by M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Pleveris, Crete Island, Greece, 5–10 June 2016
14. M. Prebeg, T. Flåtten, B. Müller, Appl. Math. Model. **44**, 124 (2017). https://doi.org/10.1016/j.apm.2016.12.010
15. S. Lindqvist, P. Aursand, T. Flåtten, A.A. Solberg, SIAM J. Numer. Anal. **54**(5), 2775 (2016). https://doi.org/10.1137/15M104935X
16. M. Prebeg, T. Flåtten, B. Müller, Large Time Step HLL and HLLC schemes. ESAIM: M2AN. (2017 In press). https://doi.org/10.1051/m2an/2017051
17. A. Harten, P.D. Lax, B. van Leer, SIAM Rev. **25**(1), 35 (1983). https://doi.org/10.1137/1025002
18. B. Einfeldt, SIAM J. Numer. Anal. **25**(2), 294 (1988). https://doi.org/10.1137/0725021
19. S. Davis, SIAM J. Sci. Stat. Comput. **9**(3), 445 (1988). https://doi.org/10.1137/0909030
20. E.F. Toro, M. Spruce, W. Speares, Shock Waves **4**(1), 25 (1994). https://doi.org/10.1007/BF01414629
21. P. Roe, J. Comput. Phys. **43**(2), 357 (1981). https://doi.org/10.1016/j.jcp.2011.06.008
22. E.F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 3rd edn. (Springer, Berlin, Heidelberg, 2009). https://doi.org/10.1007/b79761
23. B. Einfeldt, C. Munz, P. Roe, B. Sjögreen, J. Comput. Phys. **92**(2), 273 (1991). https://doi.org/10.1016/0021-9991(91)90211-3
24. P. Batten, N. Clarke, C. Lambert, D. Causon, SIAM J. Sci. Comput. **18**(6), 1553 (1997). https://doi.org/10.1137/S1064827593260140
25. R. LeVeque, *Finite Volume Methods for Hyperbolic Problems* (Cambridge University Press, 2002). https://doi.org/10.1017/CBO9780511791253
26. G.A. Sod, J. Comput. Phys. **27**(1), 1 (1978). https://doi.org/10.1016/0021-9991(78)90023-2
27. P. Woodward, P. Colella, J. Comput. Phys. **54**(1), 115 (1984). https://doi.org/10.1016/0021-9991(84)90142-6

Correction Procedure via Reconstruction Using Summation-by-Parts Operators

Philipp Öffner, Hendrik Ranocha and Thomas Sonar

Abstract The correction procedure via reconstruction (CPR, also known as flux reconstruction), is a high-order numerical scheme for conservation laws introduced by Huynh (2007), unifying some discontinuous Galerkin, spectral difference and spectral volume methods. A general framework of summation-by-parts (SBP) operators with simultaneous approximation terms (SATs) is presented, allowing semidiscrete stability for Burgers' equation using nodal bases without boundary nodes or modal bases. The linearly stable schemes of Vincent et al. (2011, 2015) are embedded within this general kind of semidiscretisation. The contributed talk *Artificial Viscosity for Correction Procedure via Reconstruction Using Summation-by-Parts Operators* given by Philipp Öffner extends these results.

Keywords High-order methods · Summation-by-parts
Correction procedure via reconstruction
Flux reconstruction · Skew-symmetric form

Mathematics Subject Classification (2010) 65M70 · 65M60 · 65M06 · 65M12

1 Introduction

This contribution is concerned with numerical methods for scalar conservation laws in one space dimension

$$\partial_t u + \partial_x f(u) = 0, \tag{1}$$

P. Öffner · H. Ranocha (✉) · T. Sonar
Institute Computational Mathematics, TU Braunschweig, Pockelsstraße 14,
38106 Braunschweig, Germany
e-mail: p.oeffner@tu-bs.de

H. Ranocha
e-mail: h.ranocha@tu-bs.de

T. Sonar
e-mail: t.sonar@tu-bs.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_37

equipped with appropriate initial and boundary conditions. For simplicity, periodic boundaries or compactly supported initial data are assumed.

In 2007, Huynh [13] introduced the *flux reconstruction* (FR) method, also known as *correction procedure via reconstruction* (CPR) [14]. This framework unifies some high-order semidiscretisations such as *discontinuous Galerkin* (DG), *spectral difference* (SD) and *spectral volume* (SV) with appropriate choice of parameters, at least for linear equations. Several results about linear stability in a semidiscrete setting are available [15, 30–32], and the method has been implemented in the open-source codes PyFR [34] and Nektar++ [1]. However, much less is known about nonlinear stability [16].

On the other hand, *summation-by-parts* (SBP) operators have their origins in another class of numerical schemes, namely *finite difference* (FD) methods, as described inter alia in the review articles [3, 17, 26] and references cited therein. Mimicking integration by parts, they have been used classically to get L_2 stability for linear equations in bounded domains [9]. Recently, the idea of SBP operators has been applied to nodal DG methods [6] and general nodal bases with appropriate quadrature strength [2].

In this contribution, the concept of SBP operators is applied to CPR methods. Using a certain reformulation of these nodal polynomial collocation schemes, semidiscrete stability results are obtained for norms adapted to the correction procedure, recovering the energy stable schemes of Vincent et al. [31, 32]. Using a skew-symmetric formulation, nonlinear stability for Burgers' equation is obtained [24]. Moreover, a generalised concept of nodal SBP bases not including boundary nodes and modal bases is presented [21].

2 Correction Procedure via Reconstruction

Traditionally, *finite element* (FE) methods approximate the solution of a conservation law in a given finite-dimensional space. Using the method of lines, a semidiscretisation is obtained by projecting the Eq. (1) onto this Hilbert space. For DG methods, these finite-dimensional approximations are commonly piecewise polynomials. Borrowing ideas from *finite volume* (FV) methods, numerical fluxes are used to couple neighbouring elements.

Approximating the integrals appearing in the projection of (1) by discrete quadrature rules with $p+1$ nodes — if polynomials of degree $\leq p$ are considered — results in a polynomial collocation framework. If Gauß nodes are used, the integrals are evaluated exactly, if the flux is linear with constant coefficients, i.e. $f(u) = u$. Otherwise, the idea of L_2 projection is only approximated.

As a polynomial collocation framework, the correction procedure via reconstruction resembles strong form nodal DG methods. For a scalar conservation law (1), it can be described as follows.

At first, the computational domain is partitioned into non-overlapping intervals. On each of these elements, the numerical solution u is given as a polynomial of

degree $\leq p \in \mathbb{N}_0$, represented in a Lagrangian basis using the values $\underline{u}_0, \ldots, \underline{u}_p$ at certain nodes. As usual in collocation frameworks, the flux is computed pointwise at these nodes, resulting in a polynomial representation with coefficients $\underline{f}_i = f(\underline{u}_i)$. For all computations, each cell is mapped to the standard element $[-1, 1]$.

Since the flux is approximated as a polynomial on each element, its derivative can be computed as the exact derivative $\underline{\underline{D}}\,\underline{f}$ of this polynomial. However, similarly to DG methods, the information of neighbouring elements has to be used as well. Therefore, the numerical solution $\underline{u}$ and its flux $\underline{f}$ are interpolated to the boundaries of the interval, yielding the point values u_L, u_R and f_L, f_R, respectively. At each boundary, a common numerical flux $f^{\text{num}}(u_-, u_+)$ is computed using the solution values from the cells to the left and right, respectively. Enforcing the point values of this numerical flux at the boundaries is done using left and right correction functions g_L, g_R, where $g_L(-1) = 1$, $g_L(1) = 1$, $g_L(x) = g_R(-x)$, and g_L, g_R approximate zero in the standard interval $[-1, 1]$ in some sense. These correction functions are polynomials of degree $\leq p + 1$, and the semidiscrete approximation can finally be written as

$$\partial_t \underline{u} + \underline{\underline{D}}\,\underline{f} + (f_L^{\text{num}} - f_L)\underline{g}'_L + (f_R^{\text{num}} - f_R)\underline{g}'_R = 0, \tag{2}$$

where $g'_{L/R}$ is the derivative of the correction function $g_{L/R}$. These correction functions are parameters enabling the recovery of some well-known schemes as described in Sect. 1.

3 Summation-by-Parts Operators

Contrary to finite element methods, where the solution is approximated in some finite-dimensional Hilbert space, *finite difference* (FD) methods are based on the idea to approximate the derivative operator. Classically, a finite set of point values $\underline{f}_i$ is used, and the linear differential operator can be represented by some matrix $\underline{\underline{D}}$.

However, SBP operators have many ideas in common with finite element matrices. In order to mimic integration by parts, a discrete scalar product with associated norm is introduced, represented by a matrix $\underline{\underline{M}}$, corresponding to the mass matrix of FE methods and being linked to some quadrature rule [11]. Approximating the L_2 scalar product, integration by parts

$$\int_\Omega u\,(\partial_x v) + \int_\Omega (\partial_x u)\,v = u\,v\big|_{\partial\Omega} \tag{3}$$

is mimicked on a discrete level as

$$\underline{u}^T\,\underline{\underline{M}}\left(\underline{\underline{D}}\,\underline{v}\right) + \left(\underline{\underline{D}}\,\underline{u}\right)^T\,\underline{\underline{M}}\,\underline{v} = \underline{u}_p\underline{v}_p - \underline{u}_0\underline{v}_0 = \underline{u}^T\,\mathrm{diag}\,(-1, 0, \ldots, 0, 1)\,\underline{v}, \tag{4}$$

if the endpoints of the interval Ω are included in the finite-dimensional representation of functions u, v.

Merging ideas of FE and FD methods, the following analytical setting of SBP operators given in [21] will be used. A finite-dimensional (real) Hilbert space X_V of functions on the volume (interval) is equipped with a basis $\mathscr{B}_V$. With respect to this basis, the mass matrix $\underline{\underline{M}}$ (symmetric, positive definite) represents the scalar product, approximating the L_2 scalar product

$$\underline{u}^T \underline{\underline{M}}\, \underline{v} = \langle \underline{u}, \underline{v} \rangle_M \approx \int_\Omega u\, v = \langle u, v \rangle_{L^2}\,. \tag{5}$$

Additionally, the derivative (divergence) operator is represented as $\underline{\underline{D}}$.

Besides, there is a finite-dimensional (real) Hilbert space X_B of functions on the boundary. In one space dimension, it is two-dimensional and the basis $\mathscr{B}_B$ is given by the point values at the boundary nodes $-1, 1$ of the reference element. On X_B, there is a bilinear form represented by $\underline{\underline{B}}$, mimicking integration with respect to the outer normal as in the divergence theorem. In one space dimension, it reads

$$\underline{u}_B^T \underline{\underline{B}}\, \underline{v}_B = u_B\, v_B \Big|_{-1}^{1}\,. \tag{6}$$

Furthermore, a restriction operator $\underline{\underline{R}}$ couples both Hilbert spaces by performing restriction / interpolation of functions on the volume to the boundary. Finally, the SBP property

$$\underline{\underline{M}}\, \underline{\underline{D}} + \underline{\underline{D}}^T \underline{\underline{M}} = \underline{\underline{R}}^T \underline{\underline{B}}\, \underline{\underline{R}}, \tag{7}$$

mimics integration by parts (3). If a nodal basis including boundary points is used, the familiar boundary operator $\underline{\underline{R}}^T \underline{\underline{B}}\, \underline{\underline{R}} = \mathrm{diag}\,(-1, 0, \ldots, 0, 1)$ is recovered.

This framework can also be extended to multiple dimensions, not relying on, but including tensor product formulations [19], similarly to the numerical setting of [12].

4 Correction Procedure via Reconstruction Using Summation-by-Parts Operators

Reformulating the semidiscrete CPR method (2) as

$$\partial_t \underline{u} + \underline{\underline{D}}\, \underline{f} + \underline{\underline{C}} \left(\underline{f}^{\mathrm{num}} - \underline{\underline{R}}\, \underline{f} \right) = 0, \tag{8}$$

the framework of SBP operators can be introduced. Here, the correction matrix $\underline{\underline{C}} = \left(g_L', g_R' \right)$ contains the derivatives of the correction functions as columns, and

$\underline{f}^{\text{num}} = \left(f_L^{\text{num}}, f_R^{\text{num}}\right)^T$, $\underline{\underline{R}}\,\underline{f} = \left(f_L, f_R\right)^T$. Then, due to the SBP property (7), one gets

Lemma 1 (Lemma 1 in [24]). *If* $\underline{1}^T \underline{\underline{M}}\,\underline{\underline{C}} = \underline{1}^T \underline{\underline{R}}^T \underline{\underline{B}}$, *then the semidiscretisation* (8) *is conservative across elements.*

Proof. Using the representation $\underline{1}$ of the constant function $x \mapsto 1$, in each element

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \int u =& \underline{1}^T \underline{\underline{M}}\,\partial_t \underline{u} = -\underline{1}^T \underline{\underline{M}}\,\underline{\underline{D}}\,\underline{f} - \underline{1}^T \underline{\underline{M}}\,\underline{\underline{C}} \left(\underline{f}^{\text{num}} - \underline{\underline{R}}\,\underline{f}\right) \\ =& -\underline{1}^T \underline{\underline{R}}^T \underline{\underline{B}}\,\underline{\underline{R}}\,\underline{f} + \underline{1}^T \underline{\underline{D}}^T \underline{\underline{M}}\,\underline{f} - \underline{1}^T \underline{\underline{R}}^T \underline{\underline{B}} \left(\underline{f}^{\text{num}} - \underline{\underline{R}}\,\underline{f}\right) = -\underline{1}^T \underline{\underline{R}}^T \underline{\underline{B}}\,\underline{f}^{\text{num}},\end{aligned} \tag{9}$$

where the SBP property (7), the assumption and exact differentiation of constant functions $\underline{\underline{D}}\,\underline{1} = 0$ have been used. Hence, summing the contributions of all elements and using periodic boundary conditions, all terms sum up to zero and the semidiscretisation is conservative. □

Since explicit Runge–Kutta methods preserve linear invariants [10, Theorem IV.1.5], using these in a fully discrete scheme results in a conservative method.

5 Linear Stability

Studying the linear advection equation with constant velocity

$$\partial_t u + \partial_x u = 0, \tag{10}$$

L_2 stability can be translated to a discrete setting using the norm induced by the mass matrix $\underline{\underline{M}}$. By the SBP property (7), proofs relying on integration by parts can be transferred to the semidiscretisation.

Jameson [15] proposed to obtain stability in some norm not necessarily approximating the L_2 norm, since all norms are equivalent in finite-dimensional spaces. Exploiting this, Vincent et al. discovered a whole family of linearly stable CPR schemes [31, 32]. Transferring their results to the reformulation of CPR methods results in

Lemma 2 (Lemma 2 of [24], see also Theorem 1 of [32]). *If the semidiscretisation*

$$\partial_t \underline{u} + \underline{\underline{D}}\,\underline{u} + \underline{\underline{C}} \left(\underline{f}^{\text{num}} - \underline{\underline{R}}\,\underline{u}\right) = 0 \tag{11}$$

of (10) *is used with* $\underline{\underline{C}} = \left(\underline{\underline{M}} + \underline{\underline{K}}\right)^{-1} \underline{\underline{R}}^T \underline{\underline{B}}$, *where* $\underline{\underline{M}} + \underline{\underline{K}}$ *is positive definite and* $\underline{\underline{M}}\,\underline{\underline{K}}$ *is antisymmetric, then the SBP CPR method is linearly stably in the discrete norm* $\|\cdot\|_{M+K}$ *induced by* $\underline{\underline{M}} + \underline{\underline{K}}$, *if an adequate numerical flux* f^{num} *is chosen.*

As for Lemma 1, the proof relies on the SBP property (7). The one-parameter family of [31] can be directly translated to this setting and the multi-parameter family of [32] can also be obtained as described in [24].

The weak coupling of adjacent elements (or boundary conditions) via surface terms $\underline{\underline{C}}\left(\underline{f}^{\text{num}} - \underline{\underline{R}}\,\underline{u}\right)$ consisting of a numerical flux and interpolated flux values resembles *simultaneous approximation terms* (SATs) used in FD methods with SBP operators.

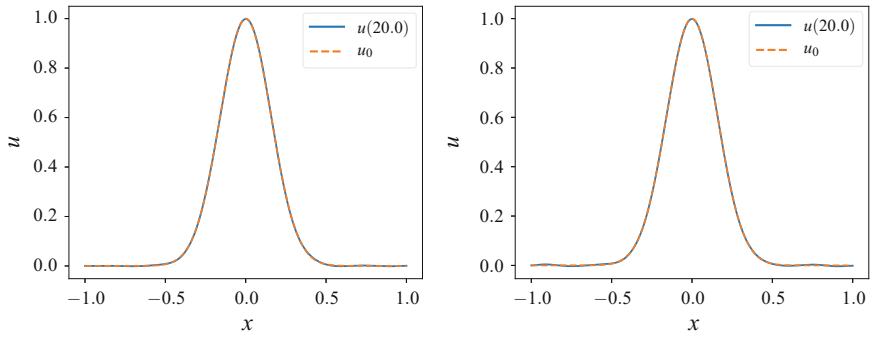
The discrete norm $\|\cdot\|_{M+K}$ in Lemma 2 approximates some kind of Sobolev norm. Thus, this equivalence of norms has to be used very carefully. First, stability and convergence results under mesh refinement have to handled warily, since the dimension of the approximation space increases and the constants for the equivalence of norms may blow up. Secondly, discrete stability results should mimic well-posedness results of the continuous PDE. For linear advection, the initial data are simply transported without change of shape. Thus, if the initial data are regular enough, corresponding Sobolev norms remain constant and this kind of stability may be acceptable. However, if the initial data are rough, the norm $\|\cdot\|_{M+K}$ approximates the Sobolev norm and can blow up. For nonlinear conservation laws, the matter is even worse, since discontinuities can develop even if smooth data are given. Hence, it may be recommended to use the canonical correction matrix $\underline{\underline{C}} = \underline{\underline{M}}^{-1}\underline{\underline{R}}^T\underline{\underline{B}}$ with norm $\|\cdot\|_M$.

Numerical experiments for the constant coefficient linear advection Eq. (10) with $N = 3$ elements using polynomials of degree $\leq p = 9$ have been conducted to evolve the initial condition $u_0(x) = \exp\left(-20x^2\right)$ in the domain $[-1, 1]$ from $t = 0$ to $t = 20$. The classical fourth-order Runge–Kutta method using four stages has been applied with 5, 000 steps.

The numerical solutions at $t = 20$ for Gauß and Lobatto nodes with corresponding diagonal norm matrices given by their quadrature rules and associated canonical correction matrices $\underline{\underline{C}} = \underline{\underline{M}}^{-1}\underline{\underline{R}}^T\underline{\underline{B}}$ (corresponding to parameters $c = c_0 = 0$ for Gauß nodes and $c = c_{\text{Hu}}$ for Lobatto nodes [24, 31]) are shown in Fig. 1. As can be seen there, the choice of Gauß nodes may be slightly better, but there is not much difference at this resolution. Here, a central numerical flux $f^{\text{num}}(u_-, u_+) = \frac{u_- + u_+}{2}$ has been used, resulting in the semidiscrete estimate $\frac{\mathrm{d}}{\mathrm{d}t}\, t\, \|u\|_M^2 = 0$.

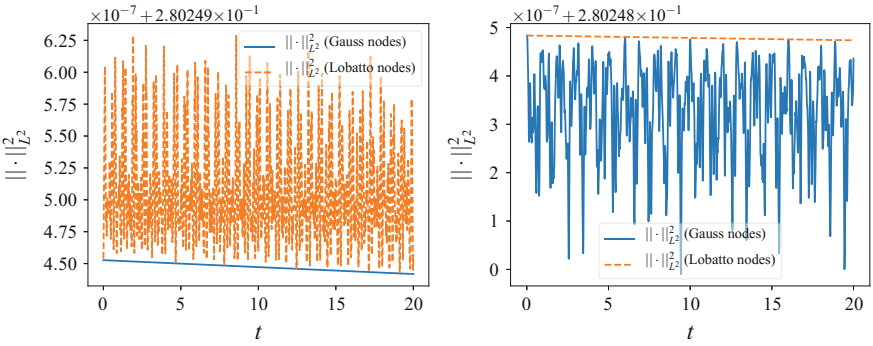
The corresponding energy is plotted in Fig. 2. For Gauß nodes, the energy computed via Gauß quadrature remains nearly constant, whereas the energy computed via Lobatto quadrature is bounded (due to equivalence of norms) but oscillatory, and vice versa for Lobatto nodes. The same phenomenon can be observed if an upwind numerical flux $f(u_-, u_+) = u_-$ is applied, as can be seen in Fig. 3. Here, the time scale is reduced, since the additional dissipation of the upwind flux reduces the oscillations considerably.

The slightly visible loss of energy for the central flux with semidiscrete estimate $\frac{\mathrm{d}}{\mathrm{d}t}\, t\, \|u\|_M^2 = 0$ in Fig. 2 can be explained by the dissipative nature of the explicit Runge–Kutta method, at least if two consecutive steps are considered [23, 25].



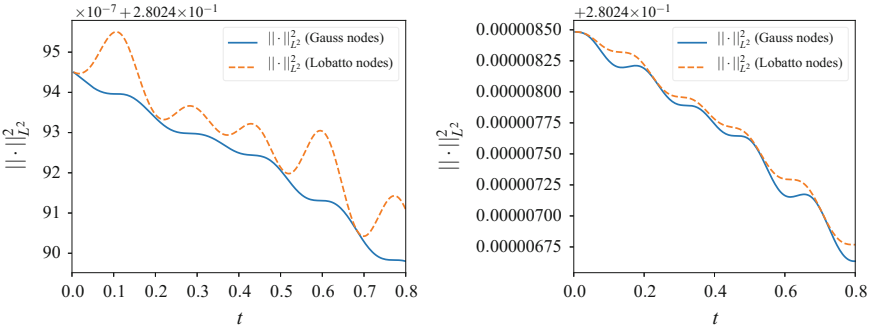
(a) $c = c_0$, Gauß nodes. (b) $c = c_{\mathrm{Hu}}$, Lobatto nodes.

Fig. 1 The numerical solutions of constant velocity linear advection at $t = 20$



(a) $c = c_0$, Gauß nodes. (b) $c = c_{\mathrm{Hu}}$, Lobatto nodes.

Fig. 2 Energies of the numerical solutions with central flux



(a) $c = c_0$, Gauß nodes. (b) $c = c_{\mathrm{Hu}}$, Lobatto nodes.

Fig. 3 Energies of the numerical solutions with upwind flux

6 Nonlinear Stability for Burgers' Equation

Extending the L_2 stability theory of linear equations to nonlinear ones, general convex entropies $U(u)$ can be considered. Using $U(u) = \frac{1}{2}u^2$, the familiar L_2 stability is recovered. Even for the general case, skew-symmetric forms have long been known to be linked to stability [18, 27]. Recently, Fisher et al. [5] provided a justification to use split operator forms in combination with diagonal norm nodal SBP bases regarding the Lax–Wendroff theorem. This theory leads further to flux-differencing forms [4], transferring results from FV methods to SBP schemes.

As classical example of a nonlinear conservation law (1), Burgers' equation

$$\partial_t u + \partial_x \frac{u^2}{2} = 0 \tag{12}$$

will be considered. The nonlinear flux $f(u) = \frac{u^2}{2}$ does not allow the same cancellation of boundary terms used in the proof of linear stability in Lemma 2. However, the split operator form

$$\partial_t u + \frac{1}{3}\partial_x u^2 + \frac{1}{3}u\partial_x u = 0 \tag{13}$$

allows to gain L_2 stability $\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 \leq 0$ using integration by parts. Applying SBP operators with nodal bases including boundary points, the corresponding semidiscretisation

$$\partial_t \underline{u} + \frac{1}{3}\underline{\underline{D}}\,\underline{u^2} + \frac{1}{3}\underline{u}\,\underline{\underline{D}}\,\underline{u} + \underline{\underline{M}}^{-1}\underline{\underline{R}}^T\underline{\underline{B}}\left(\underline{f}^{\mathrm{num}} - \frac{1}{2}\underline{\underline{R}}\,\underline{u^2}\right) \tag{14}$$

results in a conservative (across elements) and stable (in the discrete norm $\|\cdot\|_M$) method, if an adequate numerical flux f^{num} and appropriate boundary conditions are chosen, see inter alia [5, 6, 24].

This kind of split form has often been described as a correction of the product rule $\partial_x(uv) = (\partial_x u)v + u(\partial_x v)$, that is invalid both for weak solutions and in the discrete setting. However, it should be emphasised that it is *multiplication* which is invalid in the discrete setting, not only the product rule. Using this idea, the split form can be extended by introducing new boundary terms and by a more general splitting of the volume terms, resulting in

Theorem 3 (Theorem 2 of [21]). *For a general SBP operator as in Sect. 3, the semidiscretisation*

$$\partial_t \underline{u} + \frac{1}{3}\underline{\underline{D}}\,\underline{u^2} + \frac{1}{3}\underline{u}^*\underline{\underline{D}}\,\underline{u} + \underline{\underline{M}}^{-1}\underline{\underline{R}}^T\underline{\underline{B}}\left(\underline{f}^{\mathrm{num}} - \frac{1}{3}\underline{\underline{R}}\,\underline{u}\,\underline{u} - \frac{1}{6}\left(\underline{\underline{R}}\,\underline{u}\right)^2\right) = 0 \tag{15}$$

is conservative. Additionally, it is stable in the discrete norm $\|\cdot\|_M$ induced by $\underline{\underline{M}}$, if an appropriate numerical flux fulfilling the entropy stability condition of Tadmor [28, 29]

$$(u_+ - u_-) f^{\text{num}}(u_-, u_+) - \frac{1}{6}\left(u_+^3 - u_-^3\right) \leq 0 \tag{16}$$

is chosen, e.g. an entropy conservative flux, a local Lax–Friedrichs flux or Osher's flux.

Here, the $\underline{\underline{M}}$-adjoint $\underline{\underline{u}}^* = \underline{\underline{M}}^{-1}\underline{\underline{u}}^T\underline{\underline{M}}$ has been introduced, motivated by the fact that multiplication operators should be self-adjoint — at least, if an appropriate domain in infinite dimensions is chosen. In the finite-dimensional case, this condition can simply be omitted.

Using the generalisation provided by Theorem 3, non-diagonal nodal bases using, e.g., Chebyshev points as well as modal bases can be applied. For a diagonal norm modal Legendre basis using exact multiplication followed by exact L_2 projection, the multiplication operators are indeed self-adjoint, since the Legendre polynomials are orthogonal, and for polynomials u, v, w of degree $\leq p$,

$$\int \text{proj}(uv)w = \int (uv)w = \int v(uw) = \int v\,\text{proj}(uw). \tag{17}$$

Another hint not to use equivalence of norms and non-canonical correction matrices $\underline{\underline{C}} \neq \underline{\underline{M}}^{-1}\underline{\underline{R}}^T\underline{\underline{B}}$ (c.f. Sect. 5) may be the inability of the authors to prove nonlinear stability results for Burgers' equation (12) using norms different from $\|\cdot\|_M$.

As can be seen here, the idea of numerical fluxes borrowed from finite volume schemes greatly simplifies the generation of stable coupling between elements. It would be much more tedious to get an idea of the boundary terms using only inspiration from SATs in the FD community.

7 Further Research

Since only semidiscrete schemes have been investigated, fully discrete stability appears to be a natural next question. Artificial dissipation [22] and modal filtering [8] have been investigated in the setting presented here and are described in the contribution *Artificial Viscosity for Correction Procedure via Reconstruction Using Summation-by-Parts Operators* in this volume. These results have been extended in [23], where also new stability results for linear equations have been obtained.

Additionally, nonlinear stability results for systems of conservation (or balance) laws are available, e.g. for the shallow water equations, using techniques comparable to those presented here [7, 20, 33].

References

1. C. Cantwell, D. Moxey, A. Comerford, A. Bolis, G. Rocco, G. Mengaldo, D. De Grazia, S. Yakovlev, J.E. Lombard, D. Ekelschot et al., Nektar++: an open-source spectral/hp element framework. Comput. Phys. Commun. **192**, 205–219 (2015)
2. D.C.D.R. Fernández, P.D. Boom, D.W. Zingg, A generalized framework for nodal first derivative summation-by-parts operators. J. Comput. Phys. **266**, 214–239 (2014)
3. D.C.D.R. Fernández, J.E. Hicken, D.W. Zingg, Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations. Comput. Fluids **95**, 171–196 (2014)
4. T.C. Fisher, M.H. Carpenter, High-order entropy stable finite difference schemes for nonlinear conservation laws: finite domains. Technical report NASA/TM-2013-217971, NASA, NASA Langley Research Center, Hampton VA 23681-2199, United States (2013)
5. T.C. Fisher, M.H. Carpenter, J. Nordström, N.K. Yamaleev, C. Swanson, Discretely conservative finite-difference formulations for nonlinear conservation laws in split form: theory and boundary conditions. J. Comput. Phys. **234**, 353–375 (2013)
6. G.J. Gassner, A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods. SIAM J. Sci. Comput. **35**(3), A1233–A1253 (2013)
7. G.J. Gassner, A.R. Winters, D.A. Kopriva, A well balanced and entropy conservative discontinuous Galerkin spectral element method for the shallow water equations. Appl. Math. Comput. **272**, 291–308 (2016)
8. J. Glaubitz, H. Ranocha, P. Öffner, T. Sonar, Enhancing stability of correction procedure via reconstruction using summation-by-parts operators II: modal filtering (2016), arXiv:1606.01056 [math.NA]. Submitted
9. B. Gustafsson, H.O. Kreiss, J. Oliger, *Time-Dependent Problems and Difference Methods*, vol. 123 (Wiley, New York, 2013)
10. E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, vol. 31 (Springer Science & Business Media, Berlin, 2006)
11. J.E. Hicken, D.W. Zingg, Summation-by-parts operators and high-order quadrature. J. Comput. Appl. Math. **237**(1), 111–125 (2013)
12. J.E. Hicken, D.C.D.R. Fernández, D.W. Zingg, Multidimensional summation-by-parts operators: general theory and application to simplex elements. SIAM J. Sci. Comput. **38**(4), A1935–A1958 (2016)
13. H. Huynh, A flux reconstruction approach to high-order schemes including discontinuous Galerkin methods, in *AIAA Paper 2007*, vol. 4079 (2007)
14. H. Huynh, Z.J. Wang, P.E. Vincent, High-order methods for computational fluid dynamics: a brief review of compact differential formulations on unstructured grids. Comput. Fluids **98**, 209–220 (2014)
15. A. Jameson, A proof of the stability of the spectral difference method for all orders of accuracy. J. Sci. Comput. **45**(1–3), 348–358 (2010)
16. A. Jameson, P.E. Vincent, P. Castonguay, On the non-linear stability of flux reconstruction schemes. J. Sci. Comput. **50**(2), 434–445 (2012)
17. J. Nordström, P. Eliasson, New developments for increased performance of the SBP-SAT finite difference technique, *IDIHOM: Industrialization of High-Order Methods-A Top-Down Approach* (Springer, Berlin, 2015), pp. 467–488
18. P. Olsson, J. Oliger, Energy and maximum norm estimates for nonlinear conservation laws. Technical report NASA-CR-195091, NASA, Research Institute for Advanced Computer Science, Moffett Field, CA, United States (1994)
19. H. Ranocha, SBP operators for CPR methods. Master's thesis, TU Braunschweig (2016)
20. H. Ranocha, Shallow water equations: split-form, entropy stable, well-balanced, and positivity preserving numerical methods. GEM Int. J. Geomath. **8**(1), 85–133 (2017). https://doi.org/10.1007/s13137-016-0089-9, arXiv:1609.08029 [math.NA]

21. H. Ranocha, P. Öffner, T. Sonar, Extended skew-symmetric form for summation-by-parts operators (2015), arXiv:1511.08408 [math.NA]. Submitted
22. H. Ranocha, J. Glaubitz, P. Öffner, T. Sonar, Enhancing stability of correction procedure via reconstruction using summation-by-parts operators I: artificial dissipation (2016), arXiv:1606.00995 [math.NA]. Submitted
23. H. Ranocha, J. Glaubitz, P. Öffner, T. Sonar, Time discretisation and L_2 stability of polynomial summation-by-parts schemes with Runge–Kutta methods (2016), arXiv:1609.02393 [math.NA]. Submitted
24. H. Ranocha, P. Öffner, T. Sonar, Summation-by-parts operators for correction procedure via reconstruction. J. Comput. Phys. **311**, 299–328 (2016). https://doi.org/10.1016/j.jcp.2016.02.009, arXiv:1511.02052 [math.NA]
25. Z. Sun, C.W. Shu, Stability of the fourth order Runge–Kutta method for time-dependent partial differential equations (2016), https://www.brown.edu/research/projects/scientific-computing/sites/brown.edu.research.projects.scientific-computing/files/uploads/Stability%20of%20the%20fourth%20order%20Runge-Kutta%20method%20for%20time-dependent%20partial.pdf. Submitted to Annals of Mathematical Sciences and Applications
26. M. Svärd, J. Nordström, Review of summation-by-parts schemes for initial-boundary-value problems. J. Comput. Phys. **268**, 17–38 (2014)
27. E. Tadmor, Skew-selfadjoint form for systems of conservation laws. J. Math. Anal. Appl. **103**(2), 428–442 (1984)
28. E. Tadmor, The numerical viscosity of entropy stable schemes for systems of conservation laws. I. Math. Comput. **49**(179), 91–103 (1987)
29. E. Tadmor, Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. Acta Numer. **12**, 451–512 (2003)
30. P.E. Vincent, P. Castonguay, A. Jameson, Insights from von Neumann analysis of high-order flux reconstruction schemes. J. Comput. Phys. **230**(22), 8134–8154 (2011)
31. P.E. Vincent, P. Castonguay, A. Jameson, A new class of high-order energy stable flux reconstruction schemes. J. Sci. Comput. **47**(1), 50–72 (2011)
32. P.E. Vincent, A.M. Farrington, F.D. Witherden, A. Jameson, An extended range of stable-symmetric-conservative flux reconstruction correction functions. Comput. Methods Appl. Mech. Eng. **296**, 248–272 (2015)
33. N. Wintermeyer, A.R. Winters, G.J. Gassner, D.A. Kopriva, An entropy stable nodal discontinuous Galerkin method for the two dimensional shallow water equations on unstructured curvilinear meshes with discontinuous bathymetry (2016), arXiv:1509.07096v2 [math.NA]
34. F.D. Witherden, A.M. Farrington, P.E. Vincent, PyFR: an open source framework for solving advection-diffusion type problems on streaming architectures using the flux reconstruction approach. Comput. Phys. Commun. **185**(11), 3028–3040 (2014)

A Third-Order Entropy Stable Scheme for the Compressible Euler Equations

Deep Ray

Abstract A third-order WENO reconstruction has been recently proposed (Fjordholm and Ray, J Sci Comput, 68(1):42–63, 2016, [5]) in the context of finite difference schemes for conservation laws and tested for scalar conservation laws. The method, which is called SP-WENO, satisfies the *sign property* required for constructing high-order finite difference schemes for conservation laws that are provably entropy stable. In the present work, we extend the reconstruction procedure to systems of conservation laws in multiple space dimensions, with a focus on the compressible Euler equations. SP-WENO in its original form can lead to large overshoots near discontinuities when tested with the Euler equations. We show that SP-WENO can be modified to control oscillations near discontinuities, without compromising on the accuracy for smooth solutions.

Keywords Euler equations · Finite difference · Entropy stability · WENO
Sign property

MSC [2010] 35L65 · 65M06 · 76N15

1 Introduction

Conservation is one of the most basic and important principles of physics, forming the basis for several scientific models. Consider the following Cauchy problem for a generic two-dimensional hyperbolic system of conservation laws,

$$\begin{aligned} \partial_t \mathbf{U} + \partial_x \mathbf{f}(\mathbf{U}) + \partial_y \mathbf{g}(\mathbf{U}) = 0 \qquad & \forall \quad (\mathbf{x}, t) \in \mathbb{R}^2 \times \mathbb{R}^+, \\ \mathbf{U}(x, y, 0) = \mathbf{U}_0(x, y) \qquad & \forall \quad \mathbf{x} \in \mathbb{R}^2, \end{aligned} \tag{1}$$

D. Ray (✉)
École Polytechnique Fédérale de Laussane,
Lausanne CH-1015, Switzerland
e-mail: deep.ray@epfl.ch

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_38

where $\mathbf{x} = (x, y)$, $\mathbf{U} : \mathbb{R}^2 \times \mathbb{R}^+ \mapsto \mathbb{R}^m$ is the vector of conserved variables, $\mathbf{f}, \mathbf{g}$ are the Cartesian components of the flux vector, and $\mathbf{U}_0$ is the initial condition. In particular, for the two-dimensional Euler equations

$$\mathbf{U} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix}, \quad \mathbf{f}(\mathbf{U}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (E + p)u \end{pmatrix}, \quad \mathbf{g}(\mathbf{U}) = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (E + p)v \end{pmatrix},$$

where ρ, $\mathbf{u} = (u, v)^\top$ and p denote the fluid density, velocity, and pressure, respectively. The quantity E is the total energy per unit volume given by $E = \rho e + \rho|\mathbf{u}|^2/2$, where e is the specific internal energy given by a caloric equation of state, $e = e(\rho, p)$. For an ideal gas, $e = p/(\gamma - 1)\rho$ with $\gamma = c_p/c_v$ denoting the ratio of specific heats.

Solutions to conservation laws can develop discontinuities in finite time even for a smooth initial data [4]. Thus, the solutions are interpreted in a weak (distributional) sense. However, these weak solutions are not necessarily unique and must be supplemented with additional conditions, known as the *entropy conditions*, in order to single out a physically relevant solution. Assume that (1) is equipped with a convex *entropy function* $\eta(\mathbf{U})$ and *entropy flux* $\mathbf{q}(\mathbf{U}) = (q^x(\mathbf{U}), q^y(\mathbf{U}))$ satisfying the compatibility conditions $\partial_{\mathbf{U}} q^x(\mathbf{U}) = \mathbf{V}^\top \partial_{\mathbf{U}} \mathbf{f}(\mathbf{U})$ and $\partial_{\mathbf{U}} q^y(\mathbf{U}) = \mathbf{V}^\top \partial_{\mathbf{U}} \mathbf{g}(\mathbf{U})$. Here, $\mathbf{V} = \partial_{\mathbf{U}} \eta(\mathbf{U})$ is the vector of *entropy variables*. Taking the scalar product of (1) with $\mathbf{V}$ leads to an auxiliary conservation law

$$\partial_t \eta(\mathbf{U}) + \partial_x q^x(\mathbf{U}) + \partial_y q^y(\mathbf{U}) = 0, \tag{2}$$

which is valid when the solution of (1) is smooth. However, for discontinuous solutions, entropy should be dissipated at shocks, and hence, one imposes the entropy condition

$$\partial_t \eta(\mathbf{U}) + \partial_x q^x(\mathbf{U}) + \partial_y q^y(\mathbf{U}) \leqslant 0, \tag{3}$$

which is understood in the sense of distributions. A weak solution of (1) is called an *entropy solution* if (3) holds.

Harten [9] has shown that the Euler equations are equipped with a family of entropy–entropy flux pairs. We choose the following specific pair

$$\eta(\mathbf{U}) = -\frac{\rho s}{\gamma - 1}, \quad q^x(\mathbf{U}) = -\frac{\rho u s}{\gamma - 1}, \quad q^y(\mathbf{U}) = -\frac{\rho v s}{\gamma - 1}, \tag{4}$$

where $s = \ln(p) - \gamma \ln(\rho)$. The corresponding vector of entropy variables is given by

$$\mathbf{V} = \left(\tfrac{\gamma - s}{\gamma - 1} - \beta|\mathbf{u}|^2,\ 2\beta\mathbf{u}^\top,\ -2\beta \right)^\top, \quad \beta = \rho/(2p).$$

Although no global existence and uniqueness results are available for entropy solutions of the generic multi-dimensional system of conservation laws, the entropy conditions play an important role in providing global stability estimates. Formally integrating (3) in space and ignoring the boundary terms by assuming periodic or no-inflow boundary conditions, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^2} \eta(\mathbf{U})\mathrm{d}\mathbf{x} \leqslant 0 \quad \Longrightarrow \quad \int_{\mathbb{R}^2} \eta(\mathbf{U}(\mathbf{x},t))\mathrm{d}\mathbf{x} \leqslant \int_{\mathbb{R}^2} \eta(\mathbf{U}_0(\mathbf{x}))\mathrm{d}\mathbf{x} \qquad \forall\, t > 0.$$

As η is convex, the above entropy bound gives rise to an a priori estimate on the solution of (1) in suitable L^p spaces [4].

Unlike scalar conservation laws, rigorous convergence results for schemes approximating multi-dimensional systems of conservation laws are currently unavailable. Thus, the construction of schemes satisfying a discrete version of the entropy inequality (3) is a reasonable goal. Such schemes are termed as *entropy stable* schemes. Tadmor [20] proposed a novel approach for constructing entropy stable schemes for hyperbolic systems, which consists of two steps: (i) constructing an *entropy conservative* scheme satisfying a discrete version of (2), (ii) adding artificial dissipation to satisfy a discrete entropy inequality. This idea has been used to construct entropy stable schemes for several conservative systems [1, 2, 10, 15, 21, 22]. High-order entropy stable finite difference schemes on Cartesian grids can be constructed using a combination of high-order entropy conservative finite difference fluxes [12] and high-order numerical dissipation. Fjordholm et al. [6] proposed a sufficient condition to construct such high-order numerical dissipation operators leading to entropy stability, which involved the reconstruction of (scaled) entropy variables such that a sign property is satisfied at each interface. This means that the jump in the reconstructed values at every cell face must have the same sign as the jump in the corresponding cell values. These high-order entropy stable schemes have been termed as *TeCNO schemes*.

Only a small class of reconstructions is known to satisfy the sign property. A second-order limited reconstruction with the minmod limiter satisfies the sign property [6]. A third-order sign-preserving reconstruction based on appropriate limiting of quadratic polynomials was proposed [3]. It was shown in [7] that essentially non-oscillatory (ENO) interpolation satisfies the sign property and has been tested numerically in [6] to give accurate results. However, ENO schemes can show deterioration in accuracy due to the selection of unstable stencils [16]. Weighted ENO (WENO) schemes [11, 14], which take a weighted combination of lower-order ENO polynomials to give a higher-order approximation, do not suffer from accuracy deterioration faced by ENO schemes [17]. Recently, a third-order sign-preserving WENO reconstruction, called SP-WENO [5], was proposed and tested with scalar conservation laws.

The primary aim of the present work is to test the performance of SP-WENO in the TeCNO setup for the compressible Euler equations. To control oscillations observed near discontinuities in the solution, a suitable modification to SP-WENO is proposed while ensuring the salient features of the original SP-WENO are retained.

2 Mesh and Finite Difference Scheme

Consider a uniform Cartesian mesh in $\mathbb{R}^2$ with mesh point $\mathbf{x}_{i,j} = (x_i, x_j) = (i\Delta x, j\Delta y)$ forming the cell centers of cells $I_{i,j} = [x_{i-1/2}, x_{i+1/2}) \times [y_{j-1/2}, y_{j+1/2})$ for $(i, j) \in \mathbb{Z}^2$. The cell interfaces are denoted by $\mathbf{x}_{i+1/2,j} = (x_{i+1/2}, y_j)$, $\mathbf{x}_{i,j+1/2} = (x_i, y_{j+1/2})$. A generic semi-discrete finite difference scheme for the system (1) is given by

$$\frac{\mathrm{d}\mathbf{U}_{i,j}}{\mathrm{d}t} + \frac{1}{\Delta x}\left(\mathbf{F}_{i+1/2,j} - \mathbf{F}_{i-1/2,j}\right) + \frac{1}{\Delta y}\left(\mathbf{G}_{i,j+1/2} - \mathbf{G}_{i,j-1/2}\right) = 0. \tag{5}$$

Here, $\mathbf{U}_{i,j}(t) = \mathbf{U}(x_i, y_j, t)$ is the *point value* of the solution at the cell centre $\mathbf{x}_{i,j}$, while $\mathbf{F}_{i+1/2,j}$, $\mathbf{G}_{i,j+1/2}$ are conservative numerical fluxes at the cell interfaces, consistent with $\mathbf{f}$, $\mathbf{g}$ respectively.

We are interested in constructing entropy stable schemes for (1) which satisfy a discrete version of the entropy condition (3). Following the approach of Tadmor [20], we first construct an entropy conservative scheme which satisfies the following discrete relation analogous to (2)

$$\frac{\mathrm{d}\eta(\mathbf{U}_i)}{\mathrm{d}t} + \frac{1}{\Delta x}(\tilde{q}^x_{i+1/2,j} - \tilde{q}^x_{i-1/2,j}) + \frac{1}{\Delta y}(\tilde{q}^y_{i,j+1/2} - \tilde{q}^y_{i,j-1/2}) = 0, \tag{6}$$

where $\tilde{q}^x_{i+1/2,j}$, $\tilde{q}^y_{i,j+1/2}$ are consistent with q^x, q^y respectively. Furthermore, we denote the undivided jump and average across the interface $\mathbf{x}_{i+1/2,j}$ by

$$\Delta\phi_{i+1/2,j} = \phi_{i+1,j} - \phi_{i,j}, \quad \overline{\phi}_{i+1/2,j} = \frac{\phi_{i+1,j} + \phi_{i,j}}{2},$$

with similar expressions for $\Delta\phi_{i,j+1/2}$, $\overline{\phi}_{i,j+1/2}$ across $\mathbf{x}_{i,j+1/2}$. Tadmor [19] has shown that the scheme (5) satisfies (6) if the numerical fluxes $\tilde{\mathbf{F}}_{i+1/2,j} = \tilde{\mathbf{F}}(\mathbf{U}_{i,j}, \mathbf{U}_{i+1,j})$ and $\tilde{\mathbf{G}}_{i,j+1/2} = \tilde{\mathbf{G}}(\mathbf{U}_{i,j}, \mathbf{U}_{i,j+1})$ satisfy the algebraic relations

$$\Delta\mathbf{V}^\top_{i+1/2,j}\tilde{\mathbf{F}}_{i+1/2,j} = \Delta\Psi^x_{i+1/2,j}, \quad \Delta\mathbf{V}^\top_{i,j+1/2}\tilde{\mathbf{G}}_{i,j+1/2} = \Delta\Psi^y_{i,j+1/2}, \tag{7}$$

where $\Psi^x(\mathbf{U}) := \mathbf{V}^\top\mathbf{f}(\mathbf{U}) - q^x(\mathbf{U})$, $\Psi^y(\mathbf{U}) := \mathbf{V}^\top\mathbf{g}(\mathbf{U}) - q^y(\mathbf{U})$ are the *entropy potentials*. For the Euler equations with the entropy–entropy flux pair (4), the entropy potentials are $\Psi^x = \rho u$, $\Psi^y = \rho v$.

The relations in (7) are generally used to construct two-point second-order accurate entropy conservative fluxes [20]. The approach of LeFloch, Mercier, and Rhode [12] can be used to construct higher-order entropy conservative fluxes, using the second-order fluxes as building blocks. The fourth-order entropy conservative fluxes have the expression

$$\begin{aligned}\tilde{\mathbf{F}}^4_{i+1/2,j} &= \frac{4}{3}\tilde{\mathbf{F}}(\mathbf{U}_{i,j},\mathbf{U}_{i+1,j}) - \frac{1}{6}\big(\tilde{\mathbf{F}}(\mathbf{U}_{i-1,j},\mathbf{U}_{i+1,j}) + \tilde{\mathbf{F}}(\mathbf{U}_{i,j},\mathbf{U}_{i+2,j})\big),\\ \tilde{\mathbf{G}}^4_{i,j+1/2} &= \frac{4}{3}\tilde{\mathbf{G}}(\mathbf{U}_{i,j},\mathbf{U}_{i,j+1}) - \frac{1}{6}\big(\tilde{\mathbf{G}}(\mathbf{U}_{i,j-1},\mathbf{U}_{i,j+1}) + \tilde{\mathbf{G}}(\mathbf{U}_{i,j},\mathbf{U}_{i,j+2})\big).\end{aligned} \tag{8}$$

While entropy is conserved for smooth solutions, it must be dissipated near discontinuities in accordance to (3). Thus, the entropy conservative flux is augmented with an entropy variable-based artificial dissipation term as follows

$$\mathbf{F}_{i+1/2,j} = \tilde{\mathbf{F}}^4_{i+1/2,j} - \frac{1}{2}\mathbf{D}_{i+1/2,j}\Delta\mathbf{V}_{i+1/2,j}, \quad \mathbf{G}_{i,j+1/2} = \tilde{\mathbf{G}}^4_{i,j+1/2} - \frac{1}{2}\mathbf{D}_{i,j+1/2}\Delta\mathbf{V}_{i,j+1/2}, \tag{9}$$

where $\mathbf{D}_{i+1/2,j}, \mathbf{D}_{i,j+1/2}$ are symmetric positive semi-definite matrices evaluated at some suitable averaged states. It has been shown in [19] that the scheme with numerical flux given by (9) is entropy stable; i.e., it satisfies

$$\frac{\mathrm{d}\eta(\mathbf{U}_i)}{\mathrm{d}t} + \frac{1}{\Delta x}(q^x_{i+1/2,j} - q^x_{i-1/2,j}) + \frac{1}{\Delta y}(q^y_{i,j+1/2} - q^y_{i,j-1/2}) \leqslant 0,$$

where $q^x_{i+1/2,j}, q^y_{i,j+1/2}$ are consistent with q^x, q^y respectively.

Although any positive semi-definite matrix leads to entropy stability, we choose the diffusion matrix of the form $\mathbf{D} = \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^\top$, where $\mathbf{R}$ is matrix of right eigenvectors of the flux Jacobian, and $\boldsymbol{\Lambda}$ is a nonnegative diagonal matrix that depends on the eigenvalues of the flux Jacobian. In particular, we choose the *Roe-type* diffusion matrix with $\boldsymbol{\Lambda} = \mathrm{diag}(|\lambda^1|, \ldots, |\lambda^m|)$. Other suitable choices for the diffusion matrices are discussed in [6, 20].

Note that the terms $\Delta\mathbf{V}_{i+1/2,j}$, $\Delta\mathbf{V}_{i,j+1/2}$ in (9) are $\mathscr{O}(\Delta x)$, $\mathscr{O}(\Delta y)$, respectively. Thus, the flux (9) leads to a first-order accurate scheme, irrespective of the order of accuracy of the entropy conservative flux used. To obtain a higher-order scheme, we follow the procedure outlined below. For the remainder of this paper, we restrict our discussions to the reconstruction methodology along the x-direction. The reconstruction in y-direction can be done in a similar manner. We omit the subscript j whenever it is clear that it is fixed.

Consider the cell interface at $x_{i+1/2}$ between control volumes I_i and I_{i+1}. Define the vector of scaled entropy variables $\mathbf{Z} = \mathbf{R}^\top_{i+1/2}\mathbf{V}$ corresponding to this particular interface. Thus, the flux in (9) can be rewritten as

$$\mathbf{F}_{i+1/2} = \tilde{\mathbf{F}}^4_{i+1/2} - \frac{1}{2}\mathbf{R}_{i+1/2}\boldsymbol{\Lambda}_{i+1/2}\Delta\mathbf{Z}_{i+1/2}. \tag{10}$$

Let $\mathbf{Z}_i(x)$ and $\mathbf{Z}_{i+1}(x)$ be suitable polynomial reconstructions of $\mathbf{Z}$ in I_i and I_{i+1}, respectively. We denote the reconstructed values at the cell interface, and the difference in the reconstructed states, by

$$\mathbf{Z}^-_{i+1/2} = \mathbf{Z}_i(x_{i+1/2}), \quad \mathbf{Z}^+_{i+1/2} = \mathbf{Z}_{i+1}(x_{i+1/2}), \qquad [\![\mathbf{Z}]\!]_{i+1/2} = \mathbf{Z}^+_{i+1/2} - \mathbf{Z}^-_{i+1/2}.$$

Replacing the original jump $\Delta\mathbf{Z}_{i+1/2}$ in (10) by the reconstructed jump $[\![\mathbf{Z}]\!]_{i+1/2}$ leads to the higher-order accurate flux

$$\mathbf{F}_{i+1/2} = \tilde{\mathbf{F}}^4_{i+1/2} - \frac{1}{2}\mathbf{R}_{i+1/2}\boldsymbol{\Lambda}_{i+1/2}[\![\mathbf{Z}]\!]_{i+1/2}. \tag{11}$$

Fjordholm et al. [6] have shown that the scheme (5) with the numerical flux (11) is entropy stable if the reconstruction at each interface satisfies the following sign property component-wise

$$\text{sign}([\![\mathbf{Z}]\!]_{i+1/2}) = \text{sign}(\Delta\mathbf{Z}_{i+1/2}).$$

These high-order entropy stable schemes are termed as TeCNO schemes. In the next section, we briefly discuss the SP-WENO reconstruction that satisfies the sign property.

3 SP-WENO

The idea of WENO reconstruction is to use a suitable convex combination of all $2k-1$ polynomials used in the kth-order ENO reconstruction at a given interface and obtain a $(2k-1)$th-order accurate reconstruction. For third-order WENO, the left and right states at the interface $x_{i+1/2}$ are evaluated as

$$\begin{aligned} Z^-_{i+1/2} &= \quad w_{0,i+1/2}\left(\frac{Z_i}{2} + \frac{Z_{i+1}}{2}\right) + w_{1,i+1/2}\left(-\frac{Z_{i-1}}{2} + \frac{3Z_i}{2}\right), \\ Z^+_{i+1/2} &= \widetilde{w}_{0,i+1/2}\left(-\frac{Z_{i+2}}{2} + \frac{3Z_{i+1}}{2}\right) + \widetilde{w}_{1,i+1/2}\left(\frac{Z_i}{2} + \frac{Z_{i+1}}{2}\right), \end{aligned} \tag{12}$$

with the weights given as $w_0 = (3/4 - 2C_1)$, $w_1 = (1 - w_0)$, $\widetilde{w}_0 = (1/4 - 2C_2)$ and $\widetilde{w}_1 = (1 - \widetilde{w}_0)$. The functions C_1, C_2 have two important roles to play: (i) ensure third-order accuracy of the reconstructed states $Z^{\pm}_{i+1/2}$ when the solution is smooth and (ii) give least weight to the stencils containing discontinuities.

We now present the choice of C_1 and C_2 corresponding to the SP-WENO reconstruction proposed in [5]. Define the jump ratio at the interface $x_{i+1/2}$ as $\theta^-_i := \Delta v_{i+1/2}/\Delta v_{i-1/2}, \theta^+_i := 1/\theta^-_i$ and the functions $\psi^+_{i+1/2} := (1-\theta^-_{i+1})/(1-\theta^+_i)$, $\psi^-_{i+1/2} := 1/\psi^+_{i+1/2}$. Then, C_1, C_2 are chosen as

$$C_1(\theta_i^+, \theta_{i+1}^-) = \begin{cases} \frac{1}{8}\left(\frac{f^+}{(f^+)^2+(f^-)^2}\right) & \text{if } \theta_i^+ \neq 1,\ \psi^+ < 0,\ \psi^+ \neq -1 \\ 0 & \text{if } \theta_i^+ \neq 1,\ \psi^+ = -1 \\ -\frac{3}{8} & \text{if } \theta_i^+ = 1 \text{ or } \psi^+ \geqslant 0,\ |\theta_i^+| \leqslant 1 \\ \frac{1}{8} & \text{if } \psi^+ \geqslant 0,\ |\theta_i^+| > 1 \end{cases},$$

and $C_2(\theta_i^+, \theta_{i+1}^-) := C_1(\theta_{i+1}^-, \theta_i^+)$ where

$$f^+(\theta_i^+, \theta_{i+1}^-) := \begin{cases} \frac{1}{1+\psi^+} & \text{if } \theta_i^+ \neq 1,\ \psi^+ \neq -1 \\ 1 & \text{otherwise,} \end{cases}, \quad f^-(\theta_i^+, \theta_{i+1}^-) := f^+(\theta_{i+1}^-, \theta_i^+).$$

The SP-WENO reconstruction enjoys the following properties.

1. **Consistency**: The weights must be nonnegative. This is equivalent to the condition $-3/8 \leqslant C_1, C_2 \leqslant 1/8$.
2. **Sign property**: The reconstructed jump using the reconstructed values (12) can be written as

$$[\![Z]\!]_{i+1/2} = \frac{1}{2}\left[\widetilde{w}_0(1-\theta_{i+1}^-) + w_1(1-\theta_i^+)\right]\Delta Z_{i+1/2}. \tag{13}$$

   Thus, the sign property holds if $\left[\widetilde{w}_0(1-\theta_{i+1}^-) + w_1(1-\theta_i^+)\right] \geqslant 0$.
3. **Negation symmetry**: The weights should not be biased toward positive or negative solution values; i.e., they should remain unchanged under the transformation $Z \mapsto -Z$. A sufficient condition to ensure this is that C_1, C_2 are chosen to be functions of quantities that are invariant under this transformation. For instance,

$$C_k := C_k\left(\theta_i^+,\ \theta_{i+1}^-,\ |\Delta Z_{i+1/2}|,\ (|Z_i| + |Z_{i+1}|)\right), \quad k = 1, 2. \tag{14}$$

4. **Mirror property**: If the solution is mirrored about the interface $x_{i+1/2}$, the weights must also get mirrored about $x_{i+1/2}$. Assuming that C_1, C_2 have the form (14) ensuring negation symmetry, the mirror property holds if

$$C_1(a, b, c, d) = C_2(b, a, c, d) \qquad \forall\ a, b, c, d \in \mathbb{R}. \tag{15}$$

5. **Bound on jumps**: It was shown in [5] that the the reconstructed jump with SP-WENO has the bound

$$|[\![Z]\!]_{i+1/2}| \leqslant 2|\Delta Z_{i+1/2}|. \tag{16}$$

Note that for SP-WENO, the functions C_1, C_2 only depend on $\theta_i^+, \theta_{i+1}^-$ with $C_1(\theta_i^+, \theta_{i+1}^-) = C_2(\theta_{i+1}^-, \theta_i^+)$, while the conditions (14) and (15) are much more general. The reconstructed values at the cell interfaces with SP-WENO have been shown to be third-order accurate in [5].

4 Numerical Results with SP-WENO

We test the performance of SP-WENO with the Euler equations. The entropy stable numerical flux is taken to be of the form (11) with a fourth-order entropy conservative flux of the form (8), which is referred to as TeCNO4. We choose the kinetic energy and entropy conservative flux proposed in [1] as the base second-order flux. The matrices in the dissipation term are evaluated using a specific set of averaged states, which ensures that the (first-order) scheme is able to resolve stationary contact discontinuities exactly (see [1] for details). The scaled entropy variables are reconstructed using SP-WENO and ENO-3, both of which have the sign property. The semi-discrete is integrated in time using a Strong Stability Preserving 3-stage Runge–Kutta scheme (SSP-RK3) [8]. For all test cases, we choose $\gamma = 1.4$ except when indicated otherwise.

1D advecting density wave: The domain is chosen as $[0, 2]$ with the initial condition given by $\rho = 1 + 0.5 \sin^4(x), u = 0.5$, and $p = 1$. The solution is simulated till time $T = 0.5$ with CFL $= 0.5$ and periodic boundary conditions. This test case is in the same spirit as that of the test described in [16] for the linear advection equation, where ENO-3 is shown to perform poorly due to selection of linearly unstable stencils. The discrete L_h^1 errors for the density obtained on different meshes are shown in Table 1. The solution with ENO-3 loses its expected order of accuracy which drops well below second order. SP-WENO on the other hand gives more than third-order accuracy.

Shu–Osher test: This test case proposed by Shu and Osher [18] involves the interaction of shocks of different strengths and highly oscillatory smooth waves. The domain is chosen as $[-5, 5]$ with final time $T = 1.8$ and CFL $= 0.4$. The initial condition has a discontinuity at $x = -4$ with $(\rho^L,\ u^L,\ p^L) = (3.857143,\ 2.629369,\ 10.33333)$ and $(\rho^R,\ u^R,\ p^R) = (1.0 + 0.2 \sin(5x),\ 0,\ 1)$ as the left and right states respectively. The solutions with SP-WENO and ENO-3 are shown in Fig. 1 on a mesh with $N = 400$ cells. As the expression for an exact solution is not available, a solution with ENO-3 on a mesh with 2000 cells is used for reference. The TeCNO4 with

Table 1 L_h^1 error of the density for the advecting sine-wave test case. ENO-3 loses accuracy while SP-WENO gives third-order accuracy

N	SP-WENO		ENO-3	
	Error	Rate	Error	Rate
100	2.61e-06	–	3.43e-06	–
200	2.91e-07	3.17	4.46e-07	2.94
400	3.21e-08	3.18	6.66e-08	2.74
600	8.91e-09	3.16	2.88e-08	2.05
800	3.56e-09	3.19	1.79e-08	1.66
1000	1.75e-09	3.18	1.25e-08	1.59

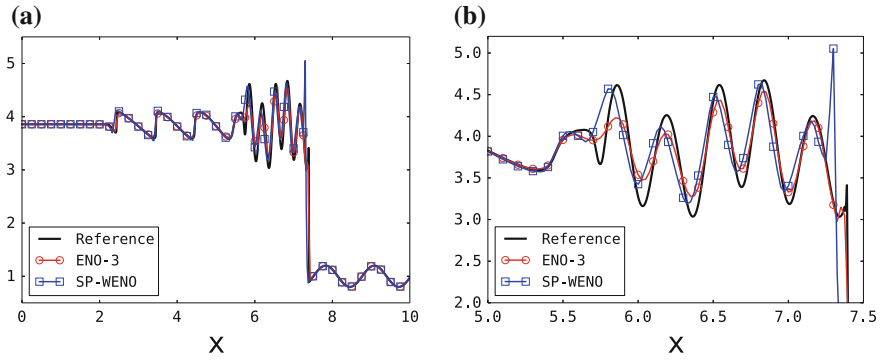


Fig. 1 Shu–Osher test: Solution with TeCNO4 at time $T = 1.8$. SP-WENO gives large overshoots

ENO-3 reconstruction does well in approximating the solution, as compared to the reference solution. However, SP-WENO gives a fairly large overshoot close to the strong shock.

5 SP-WENOc: A Fix for Systems of Conservation Laws

In order to understand why SP-WENO gives unsatisfactory results while approximating discontinuous solutions, we need to take a closer look at the scheme being used. In most scenarios, the reconstructed jump $[\![\mathbf{Z}]\!]_{i+1/2}$ with SP-WENO is zero (see [5] for details). In other words, there is no numerical dissipation in these regions and the scheme is governed only by the fourth-order entropy conservative flux. Among these, two important scenarios are when the solution has a convex or concave profile about the interface $x_{i+1/2}$. In terms of the jump ratios, these are characterized by either $\theta_i^+ < 1, \theta_{i+1}^- > 1$ or $\theta_i^+ > 1, \theta_{i+1}^- < 1$. Let us collectively call these scenarios as the C-region. While such scenarios need not occur at an interface corresponding to a discontinuity in the solution, it may describe an interface in the close proximity of a shock (or a contact). This could lead to large Gibbs oscillations, as was observed in Fig. 1.

A possible fix would be to perturb the reconstruction procedure described by SP-WENO so that the reconstructed jump is nonzero in the C-region. Consider the reconstructed jump written in the form (13). In the C-region, we introduce a small perturbation in terms of a function $\mathscr{G}$ (whose explicit form is defined later) such that

$$[\![Z]\!]_{i+1/2} = \frac{1}{2}\left[\widetilde{w}_0(1-\theta_{i+1}^-) + w_1(1-\theta_i^+) + \mathscr{G}\right]\Delta Z_{i+1/2}.$$

In order to ensure that the perturbation is a consequence of the appropriate choice of the WENO weights, we propose the following modifications

$$\overline{C}_1 = C_1 - \frac{1}{4}\frac{\mathscr{G}}{(1-\theta_i^+)}, \quad \overline{C}_2 = C_2 - \frac{1}{4}\frac{\mathscr{G}}{(1-\theta_{i+1}^-)}, \tag{17}$$

which is well defined in the C-region since $\theta_i^+ \neq 1$ and $\theta_{i+1}^- \neq 1$. We consider the additional modification

$$C_1^\# = \min\left(\max\left(\overline{C}_1, -\frac{3}{8}\right), \frac{1}{8}\right), \qquad C_2^\# = \min\left(\max\left(\overline{C}_2, -\frac{3}{8}\right), \frac{1}{8}\right), \tag{18}$$

to ensure the weights are consistent.

Finally, we chose the perturbation function $\mathscr{G}$ as

$$\mathscr{G} = \left(\min\left(\frac{|\Delta Z_{i+1/2}|}{0.5(|Z_i| + |Z_{i+1}|)}, |\Delta Z_{i+1/2}|\right)\right)^3, \tag{19}$$

Note that with this choice, $C_1^\#, C_2^\#$ have the form given by (14) and (15) and thus satisfy negation symmetry and the mirror property. Furthermore, using the fact that $\mathscr{G} \geqslant 0$ along with the consistency modification (18), the reconstruction can be shown to satisfy the sign property. The bound (16) no longer holds for SP-WENOc. However, a careful analysis leads to an alternate bound of the form

$$\left|[\![Z]\!]_{i+1/2}\right| \leqslant 4\left(\left|\Delta Z_{i-1/2}\right| + \left|\Delta Z_{i+1/2}\right| + \left|\Delta Z_{i+3/2}\right|\right). \tag{20}$$

We refer to the SP-WENO reconstruction with the correction given by (17), (18) and (19) in the C-region as SP-WENOc.

Remark 1. The bounds (16) or (20) are useful in proving Lipschitz continuity of the numerical flux. However, these bounds do not effect the sign property of the reconstructions, thus have no influence on the entropy stability of the scheme.

We make two additional remarks on the choice of $\mathscr{G}$. Firstly, the perturbation to the initial SP-WENO jump is $\mathscr{O}(|\Delta v_{i+1/2}|^4)$ for smooth solutions (assuming $C_1^\# = \overline{C}_1$ and $C_2^\# = \overline{C}_2$). This ensures that the superior order of convergence observed with the TeCNO4 scheme is retained. Secondly, the quantity $|\Delta Z_{i+1/2}|/(|Z_i| + |Z_{i+1}|)$ can have very bad scaling if $|Z_i|, |Z_{i+1}|$ are very small. Thus, by taking a minimum with $|\Delta Z_{i+1/2}|$ we are able to bound the value of $\mathscr{G}$.

Before we proceed to test SP-WENOc with the Euler equations, we first check whether SP-WENOc truly leads to a third-order reconstruction at the cell interfaces. We consider the smooth function $u(x) = \sin(10\pi x) + x$ and reconstruct the interface values using the function value at the cell centers. The discrete L_h^1 norm of interface errors are shown in Table 2. We clearly see that SP-WENOc retains the superior accuracy of SP-WENO despite the perturbation introduced in the C-region.

Table 2 L_h^1 reconstruction errors with SP-WENO and SP-WENOc

N	SP-WENO		SP-WENOc	
	Error	Rate	Error	Rate
40	8.59e-02	–	8.79e-02	–
80	6.73e-03	3.67	7.35e-03	3.58
160	5.01e-04	3.75	5.27e-04	3.80
320	3.64e-05	3.78	3.78e-05	3.80
640	2.59e-06	3.81	2.68e-06	3.82

6 Numerical Results with SP-WENOc

1D advecting density wave: The L_h^1 solution errors with SP-WENOc in the TeCNO4 setup are given in Table 3. Clearly, SP-WENOc performs at par with the original SP-WENO. In fact, the errors are almost identical.

Shu–Osher test: Large overshoots were observed with SP-WENO while solving the discontinuous Shu–Osher test. Although the new SP-WENOc reconstruction does not completely remove the overshoot, it definitely gives much better control over the magnitude of oscillation as can be seen in Fig. 2. This indicates that the numerical dissipation does not vanish in key regions under the proposed modification.

Vortex advection: This is a two-dimensional problem describing the advection of an isentropic vortex. The initial conditions are taken from [13]. The domain is taken as $[-5, 5] \times [-5, 5]$ with the initial vortex centered at the origin. The vortex advects in the horizontal direction with velocity 0.5 till $T = 20$, at the end of which the vortex completes one full cycle. The L_h^1 errors for density, pressure, and x-velocity component are shown in Table 4. Once again, both SP-WENO and SP-WENOc give very similar results, with more than third-order accuracy.

Shock–vortex interaction: This test consists of the interaction of a left-moving shock wave with a right-moving isentropic vortex. The domain is chosen as $[0, 1] \times [0, 1]$ while the initial conditions are identical to those prescribed in [6]. The domain is discretized with 200×200 cells, and the final time is chosen as $T = 0.35$. The

Table 3 L_h^1 error of density with SP-WENO reconstructions, for the advecting sine-wave test case

N	SP-WENO		SP-WENOc	
	Error	Rate	Error	Rate
100	2.61e-06	–	2.60e-06	–
200	2.91e-07	3.17	2.91e-07	3.16
400	3.21e-08	3.18	3.21e-08	3.18
600	8.91e-09	3.16	8.91e-09	3.16
800	3.56e-09	3.19	3.55e-09	3.19
1000	1.75e-09	3.18	1.74e-09	3.18

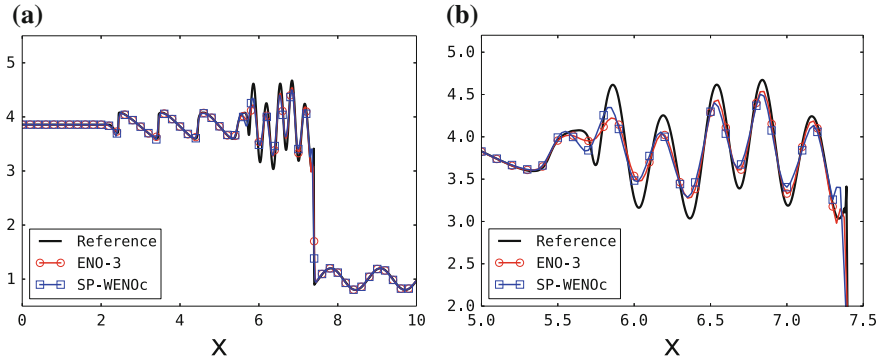
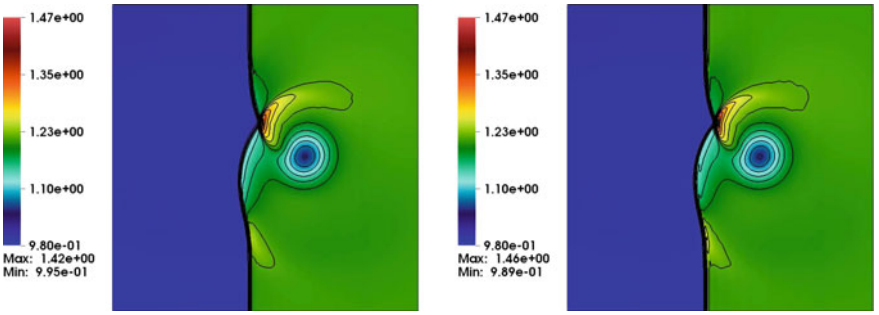


Fig. 2 Shu–Osher test: Solution with TeCNO4 at time $T = 1.8$. Overshoots are controlled by the new SP-WENOc

Table 4 L_h^1 error for advecting isentropic vortex

	Density				Pressure				x-Velocity			
N	SP-WENO		SP-WENOc		SP-WENO		SP-WENOc		SP-WENO		SP-WENOc	
	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
40	1.43e-02	–	1.46e-02	–	2.01e-02	–	2.09e-02	–	3.68e-02	–	3.66e-02	–
80	1.82e-03	2.97	1.79e-03	3.03	2.54e-03	2.98	2.53e-03	3.05	6.89e-03	2.41	6.74e-03	2.44
160	1.33e-04	3.77	1.35e-04	3.72	1.90e-04	3.74	1.95e-04	3.70	5.73e-04	3.58	5.78e-04	3.54
320	1.04e-05	3.67	1.06e-05	3.67	1.46e-05	3.70	1.50e-05	3.70	4.57e-05	3.65	4.60e-05	3.65



(a) TeCNO4 + ENO-3 (b) TeCNO4 + SP-WENOc

Fig. 3 Density profiles for shock–vortex interaction at time T = 0.35

numerical solutions with ENO-3 and SP-WENOc are comparable with well resolved shock lines, as shown in Fig. 3.

7 Conclusions

The sign-preserving SP-WENO reconstruction used in conjunction with TeCNO schemes performs well in approximating solutions of scalar conservation laws [5]. However, it leads to undesirably large oscillations close to discontinuities when tested with the Euler equations, which can be attributed to the absence of numerical dissipation in the proximity of a shock or contact discontinuity. A modification to the reconstruction is proposed to ensure the dissipation does not vanish in key areas, while maintaining high order of accuracy in smooth regions. The new method termed as SP-WENOc preserves most of the crucial properties of the original method and gives better control of overshoots near discontinuities.

Future work would test the performance of SP-WENOc for other important systems of conservation laws such as the shallow water equations and the magnetohydrodynamics equations.

Acknowledgements This research work benefited from the support of the AIRBUS Group Corporate Foundation Chair in Mathematics of Complex Systems, established in TIFR/ICTS, Bangalore.

References

1. P. Chandrashekar, Commun. Comput. Phys. **14**(5), 1252–1286 (2013)
2. P. Chandrashekar, C. Klingenberg, SIAM J. Numer. Anal. **54**(2), 1313–1340 (2016)
3. X. Cheng, Y. Nie, J. Hyperbolic Differ. Equ. **13**(1), 129–145 (2016)
4. C.M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, vol. 325, 3rd edn., Grundlehren der Mathematischen Wissenschaften (Springer, Berlin, 2010)
5. U.S. Fjordholm, D. Ray, J. Sci. Comput. **68**(1), 42–63 (2016)
6. U.S. Fjordholm, S. Mishra, E. Tadmor, SIAM J. Numer. Anal. **50**(2), 544–573 (2012)
7. U.S. Fjordholm, S. Mishra, E. Tadmor, FoCM **13**(2), 139–159 (2013)
8. S. Gottlieb, C.W. Shu, E. Tadmor, SIAM Rev. **43**(1), 89–112 (2001). (electronic)
9. A. Harten, J. Comput. Phys. **49**(1), 151–164 (1983)
10. F. Ismail, P.L. Roe, J. Comput. Phys. **228**(15), 5410–5436 (2009)
11. G.-S. Jiang, C.-W. Shu, J. Comput. Phys. **126**(1), 202–228 (1996). (Academic Press Professional, Inc.)
12. P.G. Lefloch, J.M. Mercier, C. Rohde, SIAM J. Numer. Anal. **40**(5), 1968–1992 (2002)
13. A. Lerat, F. Falissard, J. Sidès, J. Comput. Phys. **225**(1), 635–651 (2007)
14. X.-D. Liu, S. Osher, T. Chan, J. Comput. Phys. **115**(1), 200–212 (1994)
15. S. Mishra, U.S. Fjordholm, E. Tadmor, in *Foundations of Computational Mathematics Proceedings FoCM held Hong Kong 2008*. London Mathematical Society Lecture Note Series, vol. 363 (2009), pp. 93–139
16. A.M. Rogerson, E. Meiburg, J. Sci. Comput. **5**(2), 151–167 (1990)
17. C.-W. Shu, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, vol. 1697, Lecture Notes in Mathematics (Springer, Berlin, 1998), pp. 325–432
18. C.-W. Shu, S. Osher, J. Comput. Phys. **83**(1), 32–78 (1989)
19. E. Tadmor, Math. Comput. **43**(168), 369–381 (1984)
20. E. Tadmor, Acta Numer. **12**, 451–512 (2003)
21. E. Tadmor, W. Zhong, in *Mathematics and Computation, a Contemporary View: The Abel Symposium 2006*, Proceedings of the Third Abel Symposium, Alesund, Norway, 25–27 May 2006 (2006), pp. 67–94
22. A.R. Winters, G.J. Gassner, J. Comput. Phys. **304**, 72–108 (2016)

Did Numerical Methods for Hyperbolic Problems Take a Wrong Turning?

Philip Roe

Abstract Methods for the numerical solution of hyperbolic conservation laws have been dominated for several decades by discontinuous data representations and one-dimensional physical arguments. Although these ideas bestow useful properties, it will be argued that they are unrealistic and ultimately restrictive. A strategy is described that avoids one-dimensional concepts in favour of discriminating between advective (one-dimensional) and acoustic-type (multidimensional) behaviour. It is successfully applied to the Euler equations.

Keywords Conservation laws · Godunov-type methods · Hyperbolic problems Discontinuous reconstruction

1 Introduction

The first time that I presented a CFD paper to an international audience was at Stanford in 1980. On the evening of the conference banquet, Bram van Leer proposed to fill a table with believers in upwinding and Riemann solvers, but I remember that we came up short, and had to invite two believers in flux-corrected transport to make up the number. I cannot imagine having such a difficulty today; upwinding and its associated paraphernalia have become indispensable clichés in numerical treatments of hyperbolic problems.

The benefits of introducing physical reasoning into the computational method are felt in numerous ways, even when the reasoning is based only on one-dimensional analysis. Clean shock profiles are easier to produce, boundary conditions fit naturally into the method, negative densities and pressures can be avoided, fewer tuning

P. Roe (✉)
Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI, USA
e-mail: philroe@umich.edu

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_39

parameters are required, and dissipation is often reduced. These are so valuable that any limitations are usually forgiven. But the limitations are serious enough to preclude some applications. Dissipation is sometimes increased considerably. This is the case especially if shocks or contacts lie obliquely across the mesh, and for this reason it is found that unstructured grids cannot be used within boundary layers. Dissipation is also greatly increased in regions of low-speed flow, so that stagnation points generate substantial entropy layers, both locally and downstream. To reduce these effects, meshes have to be expensively refined in the problematic regions. There have been many papers published that recognise these issues, proposing a variety of "multidimensional upwind schemes". Often, these are qualified as "truly", "genuinely" or "fully" multidimensional, which to me betrays a certain insecurity among these authors as to whether their objectives were achieved. Many seem to assume that a large number of one-dimensional events must somehow constitute a multidimensional event.

One-dimensional physics is almost inevitably introduced when the data is represented discontinuously, as is the case with finite-volume methods and with discontinuous Galerkin methods. If there appears to be a discontinuity along some path in the data, then waves will propagate away from that path and will do so at right angles to it. The solution may then display behaviour that depends more on the computational grid than on any aspect of the data.

I believe that methods can be found that have all of the virtues but none of the vices of traditional upwind schemes. However, the ideas that express the physics need to be quite different from those currently in use, and so do the numerical techniques that express the ideas. In this paper, I want to return to the origin of our current methods, and then to stress a distinction that has been almost ignored. This is the great difference between the advective and acoustic modes of propagating information, in any number of dimensions except one. For ease of presentation, almost all of the discussion will actually be of the two-dimensional case, although the method is first derived in three dimensions. The greatest attention will be paid to third-order schemes, which are probably best for many applications, and for which a number of simplifications are permissible.

2 Godunov's Question

The prototype of all upwind schemes is the method proposed by Sergei Godunov [8]. Initial data is projected into the space of cellwise constant functions, thereby creating in the data one-dimensional discontinuities along the faces of the cells. The flux on every edge is then evaluated by solving the Riemann problem for the states separated by that face. There was some initial resistance to this method from researchers more accustomed to conventional finite-difference or finite-volume methods, to whom this proposal seemed both unnatural and expensive. Their anxiety was somewhat relieved by explaining to them that Godunov's method did not so much seek to approximate an evolution operator as to project the data into a form to which the

exact evolution operator could easily be applied, at least for a small timestep. In fact that had been Godunov's original idea, although he put it rather more exactly. He posed the question:

> **If the initial data were to be replaced by its cellwise constant projection, how would it evolve?**

He has provided a fascinating account [9] of the origins of the scheme that bears his name. After developing in one dimension the method inspired by the question above, and finding it to be brilliantly successful [8], he and his colleagues were eager to try it in two dimensions, and it was then that they hit a snag, described in this excerpt.

> The first question we faced in our attempts to generalise our method for two dimensions was how to find **a solution of the two-dimensional Riemann problem** with arbitrary initial conditions. ..... In order for a two-dimensional method to be absolutely similar to the one-dimensional one we needed **an analytical solution of the gas-dynamic equations with four initial discontinuities coming together in a single point**. Naturally, we did not have these solutions at that time and they are still unknown (for general initial conditions). At this point, **a roguish suggestion** was made which was to use only the solutions of a classical Riemann problem involving only planar waves. Thus, the interaction of the four cells with a common vertex was neglected altogether. This removed **a nice physical interpretation** underlying the construction of the one-dimensional scheme. Quite naturally, there were many arguments during the discussion of this **hardly justifiable suggestion**. L. V. Brushlinkii..... carried out the analytical solution for an acoustic wave propagating in stationary media, which took into account the interaction of the cells sharing a common vertex. His solution was implemented in a scheme completely analogous to the one-dimensional one. **To our surprise and pleasure there were no significant differences.** Afterwards, only the simpler model was used.

Quite clearly, the original intention was that once the cellwise constant representation had been obtained, its exact evolution over a small period of time should be determined, including the effects of four-way interactions at the call corners. There is a precision to this proposal that makes it attractive to a theoretician, who might hope eventually for a proof of convergence. However, it proved to be both difficult and ineffective, so that resort had to made to an easier but less fundamental question. In fact, now that we have a clear idea of what "corner Riemann problems" actually look like [15, 34], it seems rather obvious that incorporating such information is very unlikely to improve the accuracy, because we would be working with a subgrid model of the flow that would look like Fig. 1.[1]

[1]Confusingly, however, several researchers [4, 16, 27] discovered that including the corner flux did improve things for the *scalar* problem $\partial_t u + a\partial_x u + b\partial_y u = 0$. I propose that it is important to make the distinction between situations where information follows the fluid path (a one-dimensional domain of dependence) and information that spreads through the fluid by an acoustic (or similar) mechanism. Purely convective disturbances can never, of course, give rise to patterns such as those in Fig. 1, so that "corner terms" make more sense for scalar problems, since no acoustic disturbances are present.

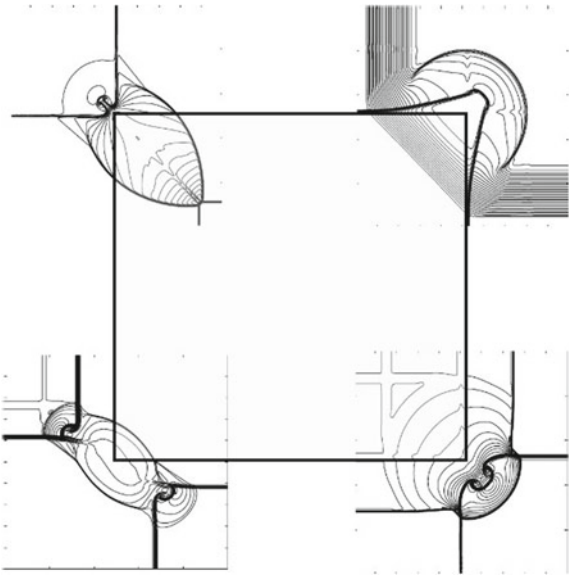


Fig. 1 If Godunovs original concept had been carried through consistently, it would have implied that during each timestep the flow would be behaving like this. Note that, for illustrative purposes, the four corner flows are chosen independently and do not connect with each other

3 Limitations of One-Dimensional Treatments

The part of Godunov's concept that has survived is of course the discontinuous reconstruction. Just as first-order Godunov methods treat the solution as piecewise constant within each cell and MUSCL schemes as piecewise linear. so discontinuous Galerkin methods can use higher-order polynomials. In all of these cases, the solution on the cell boundary is normally multivalued and must be resolved, usually by solving a Riemann problem to some approximation.

Most early applications of the Godunov method [5, 32] attempted to build directly on the satisfactory treatment of one-dimensional problems. They used operator splitting to reduce the multidimensional problems to a sequence of one-dimensional problems. To solve, for example, the Euler equations in the form

$$\partial_t \mathbf{u} + \mathbf{A}\partial_x \mathbf{u} + \mathbf{B}\partial_y \mathbf{u} = 0 \tag{1}$$

the Strang splitting method [28] would be employed, of solving for half a timestep the equation $\partial_t \mathbf{u} + \mathbf{A}\partial_x \mathbf{u} = 0$, then for a full timestep $\partial_t \mathbf{u} + \mathbf{B}\partial_y \mathbf{u} = 0$, and then for another halfstep $\partial_t \mathbf{u} + \mathbf{A}\partial_x \mathbf{u} = 0$. For smooth solutions, this is second-order accurate even in the case where $\mathbf{A}$, $\mathbf{B}$ do not commute, but for discontinuous solutions there is an $\mathcal{O}(1)$ error even if the split equations are solved exactly. This was noted qualitatively by Collela [5], and more quantitatively by Roe [25].

Waves that move in the x-direction are eigenvectors of $\mathbf{A}$ and their speeds are the eigenvalues of $\mathbf{A}$. When the propagation direction changes to y, these waves are projected onto the eigenvectors of $\mathbf{B}$ and now move with different speeds. Figure 2 shows

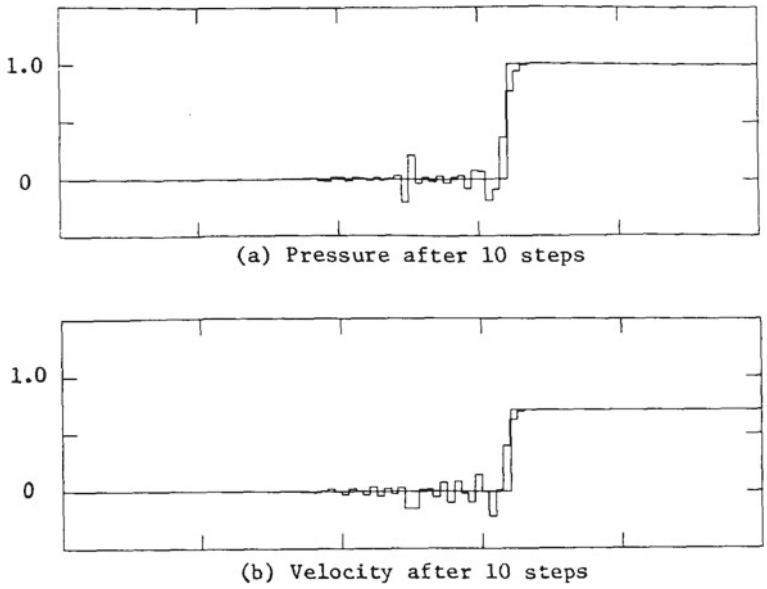


Fig. 2 Exact solution [25] to the operator-split acoustic equations, starting from an oblique shock inclined at 60°

the exact solution to the dimensionally split acoustic equations, ten split timesteps after starting with an oblique acoustic discontinuity. Of course, we do not see anything like this in practical calculations because the small wavelength of the exact solution is too small to be resolved on the mesh. In fact, we are using a subgrid model whose frequency is too short to be resolved on our grid. We are saved from the consequences of our bad physics by the dissipation in our bad numerics.

I think that the use of splitting methods arose from a feeling that multidimensional problems were being reduced to one-dimensional problems that we knew how to do. The analysis above demonstrates that this was in general an illusion and also that the physics could not be represented any better by the use of "unsplit" methods that effectively applied the one-dimensional operators simultaneously [3, 4] rather than sequentially. Technically, these are not operator-splitting methods, but there is a conceptual splitting involved that only sees the physics through a one-dimensional lens. The prolonged popularity of these methods, still the basis of almost all codes in current use, reflects the fact that they are a great improvement over the methods that came before, mostly of the Lax–Wendroff or Jameson–Schmidt–Turkel type, stabilised by artificial viscosity. Apparently, for solving hyperbolic problems, bad physics is an improvement on no physics at all.

However, the exclusive use of one-dimensional physics encourages some misconceptions. To most finite-volume practitioners, a flux is the flow rate of a conserved quantity through a surface, including any contributions from surface forces such as pressure; it is a vector, and it seems appropriate to define it on a face. However, the

flux in that sense is just one component of a "vector of vectors", and if such an object is to be stored anywhere it must be at the location that gives it most meaning. Placing it at a vertex will affect the flow through all of the adjacent faces. It is then possible to give guarantees about multidimensional aspects of the algorithm having to do with the divergence or curl of a vector field. For example, magnetohydrodynamic flows should enforce the condition $\nabla \cdot \mathbf{B} = 0$. If this condition is true in some discrete sense of the initial conditions, it should remain true after any number of timesteps. Because this invariance of $\nabla \cdot \mathbf{B}$ is a direct consequence of the governing equations, it will indeed be observed to some approximation in any consistent numerical solution for short times, but there may be secular errors such that it ceases to hold after longer times. This produces unrealistic solutions or causes the code to fail.

Morton and Roe [19] considered the equivalent problem[2] for vorticity in the acoustic equations on various grids, but most simply on a square grid. They showed that a natural measure of vorticity would be preserved invariant in time if, and only if, the fluxes through each cell face were derived by averaging fluxes from the vertices of that face (see also [20]). For a finite-volume method in two dimensions, this implies that the flux through a face must depend on at least six cells, and on 18 in three dimensions, even for a first- or second-order scheme. When fluxes are obtained from Riemann solvers they depend on only two cells. It seems self-evident that methods designed exclusively around one-dimensional processes cannot be expected to have any special properties with regard to multidimensional behaviour. Lung and Roe [17] showed that methods for solving the acoustic system on a nine-point stencil can have isotropic errors only if vertex fluxes are employed.

Another shortcoming of the dimension-by-dimension approach is an incorrect view of the flows close to stagnation points (which are to be found in many otherwise high-speed flows). For a two-dimensional stagnation point at the origin, the local solution is $u = kr\cos\theta$, $v = kr\sin\theta$, $p = p_0 + \mathcal{O}(k^2r^2)$. Riemann solvers will find acoustic waves whose strength is $\mathcal{O}(r)$ (and believe that they must be stabilised) when in fact this flow is incompressible and contains no acoustic behaviour that needs stabilising. This accounts for the spurious entropy production commonly found in these regions.

3.1 Does Discontinuous Reconstruction Help?

Despite the arguments above, the practical success of Godunov-type methods has given rise to a belief that it is natural and advantageous to look for solutions to hyperbolic problems in a function space that allows discontinuities. These spaces form the foundation of the discontinuous Galerkin method, and this is widely held to be responsible for the remarkable accuracy (superconvergence) that this method displays in some circumstances. I have recently been able to understand some simple

[2] Actually not completely equivalent. In MHD, we must enforce $\nabla \cdot \mathbf{B} = 0$, whereas in acoustics the requirement is that $\partial_t \nabla \times \mathbf{v} = 0$, even if $\nabla \times \mathbf{v}$ is not initially zero.

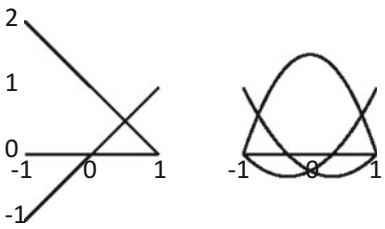


Fig. 3 Trial functions $u(\xi)$ for (left) the DG1 method, and (right) Scheme V

examples of this, in the prototypical example of linear advection $\partial_t u + \partial_x u = 0$. It is instructive to compare two advection schemes that seem quite unlike [26].

Figure 3 shows at left the two linear trial functions $\phi_{1,2}(x)$ employed by the DG1 method to form a reconstruction $u(x,t) = \sum_j \omega_j(t)\phi_j(x)$ in each cell. To derive the method, the time derivatives of the amplitudes $d/dt(\omega_j)$ are found that will minimise each cell residual, given some choice of the flux entering the cell. This does not of course determine what the order of the scheme will be, but it will be the best possible. For sensible choices of the flux, the local truncation error is found, both analytically and experimentally, to be of order h^3, which appears surprising given that we only have a linear reconstruction. Note that this is a semi-discrete method that must be coupled with some timestepping method.

At the right of Fig. 3 are shown the three quadratic trial functions $\phi_{1,2,3}(x)$ used by van Leer in Scheme V of his 1977 paper [29] on "A new approach to linear advection". These enable a quadratic reconstruction of $u(x)$ in each cell, and the reconstructed problem is solved exactly as $u(x, t + \Delta t) = u(x - \Delta t, t)$. The flux through an interface $j + \frac{1}{2}$ can be calculated from the data in the upwind cell as

$$f_{j+\frac{1}{2}} = \int_0^{\Delta t} u_j(x_{j+\frac{1}{2}} - t)\, dt \tag{2}$$

and then the cell averages are updated in the usual finite-volume manner.

$$u_j^{n+1} = u_j^n - \tfrac{\Delta t}{\Delta x}(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}) \tag{3}$$

Note that this is a fully discrete method with no need for separate timestepping. Subsequent rediscoveries of this or closely similar methods can be found in [1, 14, 24, 33]

It is rather surprising, in view of their very different derivations, that these two schemes are intimately related. Their structure is closer than appears. In both schemes, each cell needs to store just two degrees of freedom, but can make use of three. In the reconstruction stage, Scheme V makes use of its own cell average and both of the point values that it shares with neighbours. In the DG1 method, cell j can be thought of as "borrowing" a degree of freedom from the upwind cell $j - 1$ to use for calculating $f_{j-\frac{1}{2}}$. It can be shown that the DG1 scheme is the small Courant number limit of Scheme V, merely expressed in a different basis [26]. The solution

from a DG scheme of any order can be made superconvergent in all norms by defining the solution in each cell to be the polynomial defined by the degrees of freedom in that cell, plus the degrees of freedom borrowed from the upwind cell. This gives rise to a continuous discrete solution, identical to that of Scheme V. Paradoxically, the **discontinuous** representation of the solution in DG methods for linear advection produces superconvergent solutions only because there is also a **continuous** interpretation.

However, Scheme V has a number of advantages if its fully discrete form is used, as shown in Fig. 4. It is stable up to a Courant number of 1.0 rather than 0.409. It needs no iterations to advance in time, rather than at least three in DG1. It can obtain similar accuracy with half the mesh points and half the timesteps. These multiply out to give Scheme V an overall advantage of a factor of about 30, and this would become 60 or 120 in two or three dimensions. Although each step of Scheme V is more expensive than each substep of DG1, this is not nearly enough to counterbalance the other considerations. The upwinding in Scheme V does not derive from the solution to any Riemann problem. It derives instead from the use of exclusively "upwind" data to evaluate the flux in (2). The upwinding occurs within cells rather than at interfaces. In combination with the symmetrical update (3), this produces a stencil that is optimally upwinded in the sense of Iserles and Strang [13].

4 A Multidimensional Method

The remainder of this paper will describe a method for solving the Euler equations that is free from discontinuous reconstructions or one-dimensional mechanisms. Hence, it is not subject to the criticisms put forward above. No claim is made that this method is unique or that it is the best, but it is displayed as a demonstration that a coherent set of ideas can be put together, and that some remarkable improvements may ensue. To begin with, a list of properties is given that are required if the method is to be a

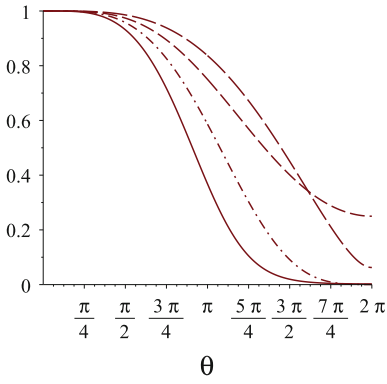


Fig. 4 Amplification factors for Scheme V at various Courant numbers. Solid line $\nu = 0$ (DG1), dash-dot $\nu = 0.25$, short dash $\nu = 0.5$, long dash $\nu = 0.75$

credible improvement on current methods. This is followed by a set of observations that are useful in achieving the requirements.

4.1 Objectives and Tools

4.1.1 Required Properties

1. Conservative in the usual sense.
2. Third-order accuracy (better than second,[3] but still fairly inexpensive [30])
3. Fully discrete, allowing for much larger timesteps.
4. Low storage and computationally intensive (for GPUs and exascale computers)
5. Works on poor-quality unstructured grids.
6. Not based on one-dimensional thinking but should recover regular upwinding in regions of one-dimensional flow.
7. Does not employ discontinuous representation, but captures narrow transitions of any kind.
8. Does not use Riemann solvers, but distinguishes correctly between advective and non-advective behaviour.
9. Applicable over the full range of Mach number.
10. Extendable to the Navier–Stokes equations.
11. Need not be applicable to all conservation laws, but may exploit specific properties of Euler and Navier–Stokes equations (at least to begin with).

Although several details still need to be ironed out, this list of objectives now seems to be well within reach.

4.1.2 Some Useful Observations

The following observations have proved useful in constructing the method

1. **fluxes** The evaluation of interface fluxes need not have any conservation property of its own. As in Scheme V, the flux through the boundaries of the control volumes should be evaluated in accordance with the local information flow, but we may use whichever variables are simplest.
2. **order** The flux evaluation can be done to one order of accuracy less than the accuracy of the conserved variables. (Because the leading error in the fluxes will integrate to zero around a closed control volume.).

[3] A personal preference, but strongly held. Schemes with odd order of accuracy have dispersion and dissipation error of the same order, bur for even error dispersion dominates [12], leading to very oscillatory representations of discontinuities, and hence to a need for strong limiting.

3. **splitting** For the linearised Euler equations, the convective and acoustic operators commute, so that operator splitting is a sound strategy. There is a corresponding splitting in the nonlinear case.
4. **no upwinding** If splitting is used, upwinding is only needed for the convective part of the problem.
5. **Poisson** The linear acoustic problem has an exact solution (Poisson's) that is simple enough to replicate van Leer's "new approach to advection" [29].

4.1.3 Representing the Solution

The representation of the data combines finite-difference, finite-element and finite-volume aspects. The initial data at $t = n\Delta t$ is represented by point values at the vertices and the mid-points of element sides, as in quadratic Lagrange elements (Fig. 5). Because the three edge values are shared with another element and the three vertex values are typically shared with six other elements, this calls for two degrees of freedom per element. Because these independent degrees of freedom are used to compute the interface fluxes, the scheme is called the active flux scheme. The cell averages of the conserved variables are also held.

The unknown point values are represented in terms of primitive variables $(\rho, \mathbf{v}, p)$ because they describe the physics very simply. It is explained below how to update these point values, and they will be updated to $(n + \frac{1}{2})\Delta t$ as well as to $(n + 1)\Delta t$. (The two updates, which are only required to second order (4.1.2, 2) could be combined into one because they use the same data.) The point values can be converted to flux variables and then the flux integral round each element can be evaluated by Simpson's Rule and used to update the cell averages of the conserved variables. Since these averages would not in general coincide with those obtained by integrating the interpolant of the boundary points, the Lagrange elements are enriched with a cubic "bubble function" that vanishes on the element boundary but makes a finite contribution to the integral.

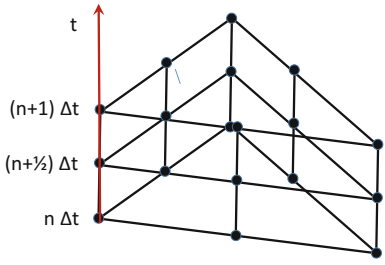


Fig. 5 The solution is represented by point values of the primitive variables at the vertices and edge mid-points of a simplicial element. The mean values of the conserved variables are also stored for each element

4.2 Techniques

4.2.1 Operator Splitting

Whenever a system of partial differential equations involves the sum of two or more operators, so that $\partial_t \mathbf{u} + \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{u} \ldots = 0$, we can consider applying the operators successively

$$\mathbf{u}^{n+1} = (\mathbf{I} - \mathbf{A}\Delta t)(\mathbf{I} - \mathbf{B}\Delta t) \ldots \mathbf{u}^n$$

This is attractive if the individual operators **A** and **B** are simple, and it introduces no error at the p.d.e. level if they commute. This requires that **A** and **B** share the same eigenvalues, implying that information propagates along the same paths in the split and unsplit systems. Dimensional splitting, where each operator involves only the derivatives in one direction, was discussed in Sect. 3 and found not to meet this condition.

Intuitively, a nice splitting would be between advective and acoustic behaviour, but unfortunately this is not possible in conservation form. The problem lies in the energy equation

$$\partial_t E = -\nabla \cdot ((E + p)\mathbf{v}) = -(\nabla \cdot (E\mathbf{v}) + (\mathbf{v} \cdot \nabla)p + p\nabla \cdot \mathbf{v}) \tag{4}$$

The conservative term $\nabla \cdot (p\mathbf{v})$ splits into two terms, $(\mathbf{v} \cdot \nabla)p$ that relates to the advection of molecular kinetic energy, and $p\nabla \cdot \mathbf{v}$ that relates to the increase of internal energy by acoustic waves doing pressure work. To achieve a physically correct splitting, this term must be split into components that are not in conservation form. A correct advective/acoustic splitting cannot be a flux splitting. Nevertheless, the suggestion has often been made to include the term $p\mathbf{v}$ in the "acoustic part of the flux" [11, 34], sometimes with apparent success, but see the results in [18].

However, the fluxes do not have to be in any sense conservative, and in order to calculate them a very simple nonconservative splitting can be applied to the Euler equations written as $\partial_t \mathbf{u} + \mathbf{C}\mathbf{u} + \mathbf{D}\mathbf{u} = 0$ with $\mathbf{u} = (\rho, u, v, p)^T$ and with

$$\mathbf{C} = (u\partial_x + v\partial_y)\mathbf{I}, \qquad \mathbf{D} = \begin{pmatrix} 0 & \rho\partial_x & \rho\partial_y & 0 \\ 0 & 0 & 0 & \rho^{-1}\partial_x \\ 0 & 0 & 0 & \rho^{-1}\partial_y \\ 0 & \gamma p\partial_x & \gamma p\partial_y & 0 \end{pmatrix} \tag{5}$$

It is clear that these matrices do commute because **C** is a multiple of the identity. We can therefore write the second-order expansion

$$\mathbf{u}^{n+1} = (\mathbf{I} - \Delta t(\mathbf{C} + \mathbf{D}) + \tfrac{1}{2}\Delta t^2(\mathbf{C} + \mathbf{D})^2)\mathbf{u}^n \tag{6}$$

$$= (\mathbf{I} + \Delta t\mathbf{C} + \tfrac{1}{2}\Delta t^2\mathbf{C}^2)(\mathbf{I} + \Delta t\mathbf{D} + \tfrac{1}{2}\Delta t^2\mathbf{D}^2)\mathbf{u}^n + \mathcal{O}(\Delta t^3) \tag{7}$$

or reverse the order of the operators. These solutions correspond to the data evolving acoustically and then translating or *vice versa*. The split operators **C** and **D** have precisely the correct eigenstructure (4.1.2, 3) to describe the two kinds of evolution.[4] A benefit of splitting in this way is that the acoustic operator **D** does not need to be upwinded because it has no directional bias (4.1.2, 4). The advection operator **C**, of course, should be upwinded along the flow direction. After the fluxes have been updated by this nonconservative splitting, they can be integrated to give an update of the cell-average quantities that is conservative in the usual finite-volume sense.

The update of the points on the boundary of the element has two almost independent stages, advective and acoustic. In the nonlinear case, these do not commute, but operator splitting turns out to be still valid if carried out in the alternating manner discovered by Strang [28]. The advective stage is fairly straightforward. It is a form of semi-Lagrangian interpolation that incorporates streamline curvature, but the acoustic stage is more novel and will now be described.

4.2.2 The Acoustic Update

The initial-value problem for the scalar wave equation in three dimensions

$$\partial_{tt}\phi - c^2\nabla^2\phi = 0 \tag{8}$$

has a well-known solution [31] in terms of spherical means, published by Poisson in 1818 [23], and put to numerical use in [2, 6, 10]. This is not, however, the form in which acoustic disturbances present themselves in the Euler equations. Instead we see the first-order system form of (8):

$$\begin{aligned} \partial_t p + c\nabla\cdot\mathbf{u} &= 0 \\ \partial_t \mathbf{u} + c\nabla p &= 0 \end{aligned} \tag{9}$$

The distinction is essential, in that (8) only applies to flows without vorticity. Nevertheless, the Poisson formula can be modified so that it applies to (9), thus

$$p(\mathbf{x}, s) = p(\mathbf{x}, 0) - ct\,M_{ct}\nabla\cdot\mathbf{u} + \int_0^{ct} s\,M_{cs}\nabla^2 p(\mathbf{x}, 0)\,ds \tag{10}$$

$$\mathbf{u}(\mathbf{x}, s) = \mathbf{u}(\mathbf{x}, 0) - ct\,M_{ct}\nabla p + \int_0^{ct} s\,M_{cs}\nabla(\nabla\cdot\mathbf{u}(\mathbf{x}, 0))\,ds \tag{11}$$

[4] For a splitting to be physically meaningful, it should not introduce any behaviour that does not exactly match something found in the unsplit problem. This means that there should be no new eigenvectors unless their eigenvalues vanish. The matrix **D** has rank two and has two left nullvectors, $(a^2, 0, 0, -1)$ and $(0, -\partial_y, \partial_x, 0)$. These signify that the linearised acoustic operator correctly makes no response to changes of entropy or to vorticity.

where $M_r\phi(\mathbf{x}, t)$ denotes the mean value of the function $\phi(\mathbf{x}, t)$ over a sphere of radius r with centre $\mathbf{x}$. The formula (10) is in fact equivalent to the classical Poisson solution of the scalar wave equation, but (11) is a little different, since $\nabla(\nabla \cdot \mathbf{u}) = \nabla^2\mathbf{u} + \nabla \times (\nabla \times \mathbf{u})$. These formulas can be applied to two-dimensional flows by the method of descent. The integrals reduce to very simple expressions for low-order polynomial data [7]. If descent is taken further to one dimension, then regular characteristic solutions are recovered, as they must be. There can be no argument about this being a genuinely/truly/fully multidimensional approach.

A nice property of (10), (11) is that they contain only those precise quantities, the pressure gradient and the velocity divergence, that are responsible for change in either linear or nonlinear problems. These two quantities are, as they should be, rotationally invariant. There is some resemblance to a Lax–Wendroff expansion. The second term contains all of the solution depending on odd powers of t and the third contains all of the even powers. The spherical means can be expanded in powers of the Laplacian operator, and then truncated to produce Lax–Wendroff schemes of any order. However, if the true spherical means are used, as is easily done for polynomial reconstructions, then the resulting numerical method is exact in the linear case for globally polynomial data (of the appropriate order, here quadratic) on any grid.

The acoustic update is carried out [6, 7] by integrating (10), (11) over the disc shown in Fig. 6. This is efficiently coded as a loop over elements, calculating the integrals for all of the six segments that fall within each element. These integrals have very simple closed forms. When the loop is finished the complete discs will have been taken into account. To deal with nonlinear problems it has proved sufficient to use a local sound speed for each segment of the disc. The physical limit of stability would be if an arc of some disc, for example, the one centred on a point P, were to leave one of the elements to which P belongs. The solution at P would then depend on data to which it has no access. In practice, we have found that this condition defines very precisely the maximum timestep that can be taken, and we use this to define a Courant number of 1.0.

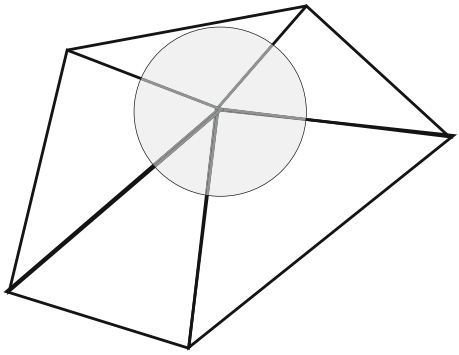


Fig. 6 Each point value receives its acoustic update by integrating (10), (11) over the Mach disc that surrounds it

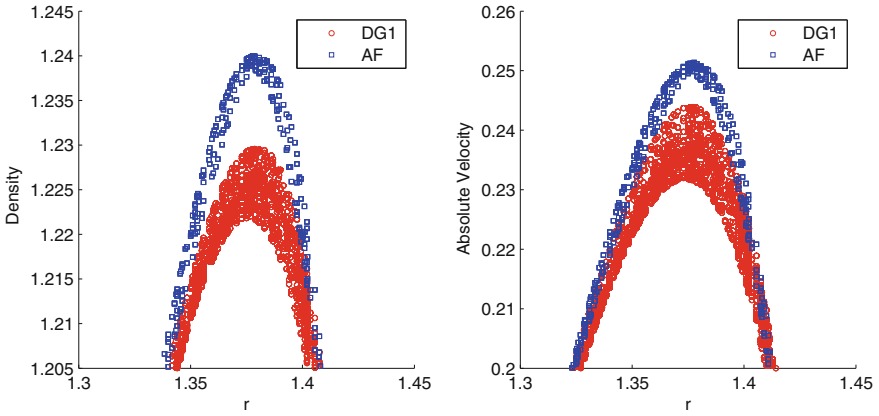


Fig. 7 A nonlinear expanding wave computed on an unstructured triangular grid. Solutions near the peak of the wave (left) density, and (right) velocity magnitude, from the present method in blue, and from DG1 in red

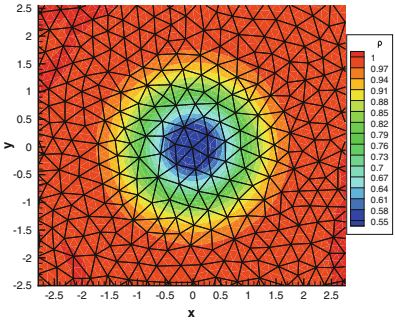
In Fig. 7 some results are shown for a nonlinear wave system,

$$\partial_t \rho + c\nabla \cdot \mathbf{v} = 0, \qquad \partial_t \mathbf{v} + c\nabla(\rho^\gamma) = 0, \tag{12}$$

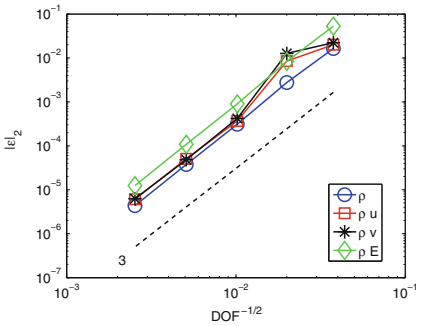
with the initial data consisting of a narrow Gaussian hill. The solution is, of course, a circular wave that expands and decays. The figure shows plots, of density and velocity magnitude against radius, in close-up around the expanding wave. Results are shown for this method, compared with DG1. For this problem, both methods are third-order accurate, but the amplitude of the wave is much better resolved by the present method, which moreover displays only about one quarter of the scatter. Just as for Scheme V in one dimension, the maximum Courant number is 1.0 rather than 0.4, and moreover only 25% of the elements are needed. Other things being assumed equal, the active flux method is 60 times more effective. These results are on unstructured grids and are very slightly better than results on triangular grids obtained by drawing diagonals through square cells. Those display slight directional bias.

5 The Full Euler Equations

Results are shown for the well-known "travelling vortex" test problem. The initial data in this case is that $\rho = 1 - 0.09046\ \exp(1 - r^2)$, $p = \rho^\gamma$, $u = 1 - 0.79577\ y\ \exp\frac{1}{2}(1 - r^2)$, $v = 0.79577\ x\ \exp\frac{1}{2}(1 - r^2)$ with $r^2 = x^2 + y^2$. The vortex travels to the right, at a Mach number of 0.845 inside a square domain $[-10, 10] \times [-10, 10]$



(a) Density contours superposed on the mesh



(b) Convergence of conserved variables

Fig. 8 Results for an inviscid travelling vortex

under periodic boundary conditions until it returns to its initial position, a distance equal to 20 times its core radius. The density contours plotted in Fig. 8a show that excellent radial symmetry has been maintained on an unstructured mesh having about 8 elements across the core. This grid is the middle one of five plotted in Fig. 8b, which demonstrates consistent third-order convergence. On this problem, the DG1 scheme displayed only second-order accuracy. The results from the active flux method were more similar to those from DG2. A more detailed account of this test is provided in Table 1.

Finally, we remark that the extension to the Navier–Stokes equations and other dissipative systems should be possible provided the equations are expressed in hyperbolic form [21, 22]. No new ideas will be required.

6 Summary

The almost universal reliance on upwind methods and Riemann problems comes from trying to insist that discrete information should propagate, so far as possible, along the same paths and in the same combinations as information in the continuum. In one dimension, the Riemann problem separates left-going information from right-going information in a satisfactory way. In higher dimensions, there is another important distinction to make, that between information that travels with the medium (advection) and that which travels though the medium (wavelike, in this context acoustic). The two modes have different domains of dependence, and stencils appropriate to one are not suitable for the other. One method has been sketched out that does make this distinction, with considerable gains in efficiency.

Table 1 Summary of convergence rates for the translating vortex

Level	$h = \sqrt{(DOF)}$	$\|\varepsilon\|_1$	Order	$\|\varepsilon\|_2$	Order
ρ					
1	3.766218e-02	6.130045e-03		1.640785e-02	
2	1.997206e-02	1.115998e-03	2.6855	2.760080e-03	2.8101
3	1.017604e-02	8.542078e-05	3.8112	3.097935e-04	3.2435
4	5.098392e-03	1.013181e-05	3.0848	3.769987e-05	3.0476
5	2.531159e-03	1.266074e-06	2.9700	4.348971e-06	3.0842
ρu					
1	3.766218e-02	1.111388e-02		2.012856e-02	
2	1.997206e-02	3.150319e-03	1.9875	8.509234e-03	1.3573
3	1.017604e-02	1.291610e-04	4.7371	3.671614e-04	4.6613
4	5.098392e-03	1.774744e-05	2.8719	5.051998e-05	2.8699
5	2.531159e-03	2.050113e-06	3.0823	6.069465e-06	3.0262
ρv					
1	3.766218e-02	1.236316e-02		2.233989e-02	
2	1.997206e-02	3.743640e-03	1.8834	1.272513e-02	0.8872
3	1.017604e-02	1.433839e-04	4.8381	4.185226e-04	5.0639
4	5.098392e-03	1.744371e-05	3.0481	4.832258e-05	3.1237
5	2.531159e-03	2.124504e-06	3.0067	6.244666e-06	2.9221
ρE					
1	3.766218e-02	1.898995e-02		5.308827e-02	
2	1.997206e-02	3.608992e-03	2.6177	8.351326e-03	2.9158
3	1.017604e-02	2.811710e-04	3.7850	9.040065e-04	3.2973
4	5.098392e-03	3.526191e-05	3.0041	1.085951e-04	3.0664
5	2.531159e-03	4.335015e-06	2.9933	1.240334e-05	3.0984
s					
1	4.559608e-02	3.458372e-03		1.112317e-02	
2	2.430365e-02	7.044419e-04	2.5289	2.844254e-03	2.1674
3	1.242164e-02	7.203506e-05	3.3973	3.233087e-04	3.2397
4	6.233828e-03	8.489049e-06	3.1016	3.939339e-05	3.0532
5	3.097460e-03	1.092373e-06	2.9317	4.992295e-06	2.9535

Acknowledgements To the students who have been brave enough to share my venture into the unknown, especially Timothy Eymann, Doreen Fan and Brad Maeng. Thanks also to Ju Wang, Kyle Ding, Praveen Chakravarthy and Professor Krysztof Fidkowski. Support under NASA Cooperative Agreement NNX12AJ70A is gratefully acknowledged.

References

1. R. Akoh, S. Ii, F. Xiao, A multi-moment finite volume formulation for shallow water equations on unstructured mesh. J. Comput. Phys. **229**(12) (2010)
2. B. Alpert, L. Greengard, T. Hagstrom, An integral evolution formula for the wave equation. J. Comput. Phys. **162**(2) (2000)
3. J.B. Bell, C.N. Dawson, G.R. Shubin, An unsplit, higher order Godunov method for scalar conservation laws in multiple dimensions. J. Comput. Phys. **74**(1), 1–24 (1988)
4. P. Colella, Multidimensional upwind methods for hyperbolic conservation laws. J. Comput. Phys. **87**(1), 171–200 (1990)
5. P. Colella, Glimms method for gas dynamics. SIAM J. Sci. Stat. Comput. **3**(1), 76–110 (1982)
6. T.A. Eymann, P.L. Roe, Multidimensional active flux schemes, in *21st AIAA Computational Fluid Dynamics Conference* (2013)
7. D. Fan, P.L. Roe, Investigations of a new scheme for wave propagation, in *22nd AIAA Computational Fluid Dynamics Conference* (2015)
8. S.K. Godunov, A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. Matematicheskii Sbornik **89**(3) (1959)
9. S.K. Godunov, Reminiscences about difference schemes. J. Comput. Phys. **153**(1) (1999)
10. T. Hagstrom, High-resolution difference methods with exact evolution for multidimensional waves. Appl. Numer. Math. **93** (2015)
11. D.W. Halt, R.K. Agarwal, A novel algorithm for the solution of compressible Euler equations in wave/particle split (WPS) form, in *11th AIAA CFD Conference, Orlando, FL* (1993)
12. G.W. Hedstrom, Models of difference schemes for $u_t + u_x = 0$ by partial differential equations. Math. Comput. **29** (1975)
13. A. Iserles, G. Strang, The optimal accuracy of difference schemes. Trans. Am. Math. Soc. **277**(2), 779–803 (1983)
14. S.A. Karabasov, V.M. Goloviznin, Compact accurately boundary-adjusting high-resolution technique for fluid dynamics. J. Comput. Phys. **228**(19) (2009)
15. D. Lax, X.-D. Liu, Solution of two-dimensional Riemann problems of gas dynamics by positive schemes. SIAM J. Sci. Comput. **19**(2) (1998)
16. R.J. LeVeque, High-resolution conservative algorithms for advection in incompressible flow. SIAM J. Numer. Anal. **33**(2), 627–665 (1996)
17. T.B. Lung, P.L. Roe, Toward a reduction of mesh imprinting. Int. J. Numer. Methods Fluids **76**(7), 450–470 (2014)
18. P.R.M. Lyra, K. Morgan, A review and comparative study of upwind biased schemes for compressible flow computation. Part I: 1D first order schemes. Arch. Comput. Methods Eng. **7**(1) (2000)
19. K.W. Morton, P.L. Roe, Vorticity-preserving Lax–Wendroff-type schemes for the system wave equation. SIAM J. Sci. Comput. **23**(1), 170–192 (2001)
20. S. Mishra, E. Tadmor, Constraint preserving schemes using potential-based fluxes. II. Genuinely multidimensional systems of conservation laws. SIAM J. Numer. Anal. **49**(3), 1023–104 (2011)
21. H. Nishikawa, New-generation hyperbolic Navier–Stokes schemes: O (1/h) speed-up and accurate viscous/heat fluxes, in *20th AIAA Computational Fluid Dynamics Conference* (2011)
22. I. Peshkov, E. Romenski, A hyperbolic model for viscous Newtonian flows. Contin. Mech. Thermodyn. **28**(1–2) (2016)
23. S.-D. Poisson, Mémoire sur la théorie des ondes. Mém. Acad. R. Sci. Inst. Fr. **2**, 70–186 (1818)
24. M.V. Popov, S.D. Ustyugov, Piecewise parabolic method on local stencil for gasdynamic simulations. Comput. Math. Math. Phys. **47**(12) (2007)
25. P.L. Roe, Discontinuous solutions to hyperbolic systems under operator splitting, *Upwind and High-Resolution Schemes* (Springer, Berlin, 1997), pp. 470–490
26. P.L. Roe, A simple explanation of superconvergence for discontinuous Galerkin solutions to $u_t + u_x = 0$. Commun. Comput. Phys. **21**(4) (2017)

27. P.L. Roe, D. Sidilkover, Optimum positive linear schemes for advection in two and three dimensions. SIAM J. Numer. Anal. **29**(6), 1542–1568 (1992)
28. G. Strang, On the construction and comparison of difference schemes. SIAM J. Numer. Anal. **5**(3), 506–517 (1968)
29. B. Van Leer, Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. J. Comput. Phys. **23**(3) (1977)
30. Z.J. Wang, et al., High order CFD methods: current status and perspective. Int. J. Numer. Methods Fluids **72**(8) (2013)
31. G.B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974)
32. P. Woodward, P. Colella, The numerical simulation of two-dimensional fluid flow with strong shocks. J. Comput. Phys. **54**(1) (1984)
33. X. Zeng, A high-order hybrid finite difference-finite volume approach with application to inviscid compressible flow problems: a preliminary study. Comput. Fluids **98** (2014)
34. Y. Zheng, *Systems of Conservation Laws, Two-dimensional Riemann Problems* (Springer, Berlin, 2001)

Astrophysical Fluid Dynamics and Applications to Stellar Modeling

Friedrich K. Röpke

Abstract The modeling of astrophysical objects poses a challenging multiscale multiphysics problem. Because of their large spatial extent, the description of physical processes dominating the formation, structure, and evolution of such objects is typically based on effective theories such as fluid dynamics or thermodynamics. The modeling ansatz resulting from this approach is the Euler equations in combination with appropriate source terms. In contrast to terrestrial systems, the astrophysical equations of state are usually more complex and the ranges of relevant scales in space, time, density, velocity etc., in the considered objects are orders of magnitude wider. Simulations therefore require an efficient description of physical effects, elaborate numerical techniques, and models of unresolved phenomena. We exemplify this by focusing on processes in stars. This multiphysics problem is characterized by coupling the compressible Euler equations to the simultaneous effects of gravity, nuclear reactions, hydrodynamic instabilities, and mixing processes in the stellar fluid. It implies a multis because the processes act on scales in space and time that can easily be separated by ten orders of magnitude. The traditional astrophysical approach to this challenge—one-dimensional models parametrizing the description of unresolved effects—lacks predictive power. The dramatic increase in computational power, however, enables multidimensional dynamical simulations. They pave the way to the next generation of stellar models and promise new insights into the physical processes in stars. We discuss to which degree the currently applied techniques are able to cope with the scale problems. Among other techniques, we point out the importance of finding algorithms that allow for efficient parallelization and

F. K. Röpke (✉)
Heidelberger Institut für Theoretische Studien, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
e-mail: friedrich.roepke@h-its.org

F. K. Röpke
Institut für Theoretische Astrophysik, Zentrum für Astronomie der Universität Heidelberg, Philosophenweg 12, 69120 Heidelberg, Germany

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_40

the use of problem-adapted geometries of the discretization grids. Further progress critically depends on continuous improvement of the methods, and input from applied mathematics will play a key role in this development.

Keywords Stellar fluid dynamics · Supernovae · Low-Mach number methods Numerical simulations

1 Introduction

The theory of fluid dynamics is suitable to describe many processes in astrophysics. In this contribution, we discuss its use in astrophysical simulations starting out from an overview of the challenges in astrophysical modeling in Sect. 2. In Sect. 3, we focus on the application of fluid dynamical concepts to the modeling of *stars*. While conventional approaches to describing stellar structure and evolution deal with the tremendous scale problems by reducing the dimensionality of hydro*static* setups, predictive models require a hydro*dynamical* description in three spatial dimensions. First steps toward such a new generation of stellar models are discussed, and examples are given in Sect. 4. Conclusions are drawn in Sect. 5.

2 The Challenges of Astrophysical Fluid Dynamics

From the largest structures in the universe, over galaxy clusters, galaxies, globular clusters, stars, planets down to protoplanetary dust, the spatial scales of astrophysical objects span more than 25 orders of magnitude. Inside these objects, physical processes on "microscopic" scales may dominate their evolution and have to be accounted for in astrophysical modeling. Densities encountered in the objects of interest span an even wider range of values. The intergalactic medium is very dilute (typical densities are $\rho \sim 10^{-28}\,\mathrm{g\,cm^{-3}}$). Sun-like stars have $\rho \sim 1\,\mathrm{g\,cm^{-3}}$ and neutron stars reach densities above $10^{14}\,\mathrm{g\,cm^{-3}}$, so that the astrophysically relevant range covers more than 40 orders of magnitude. Temporal scales in stellar processes range from fractions of a millisecond to gigayears. Of course, this overstates the problem somewhat as not all these extreme scales are relevant in a single setup. Nonetheless, even for individual astrophysical objects the range of physically relevant scales is huge.

A variety of physical processes and phenomena governs the structure and evolution of astrophysical objects. These include nuclear reactions, quantum mechanical effects, particle physics phenomena, magnetic fields, gravity, energy transport by various mechanisms, and plasma physics effects.

It is thus clear that the description of astrophysical phenomena poses a great multiscale multiphysics challenge to theoretical modeling. Various strategies are taken to meet this challenge. One of them is the divide-and-conquer approach. In

this spirit, only certain small regions in astrophysical objects or isolated astrophysical processes are simulated, or the dimensionality of the problem at hand is reduced by assuming symmetries. This approach has limited applicability, because

- often several (counteracting) effects act at the same order of magnitude, and it is impossible to separate a single physical process that dominates the astrophysical evolution,
- small-scale processes often depend on large-scale physics (e.g., in the cases of star and galaxy formation),
- an artificial reduction of dimensionality by assumed, but in reality not precisely existing symmetries introduces free tunable parameters that deteriorate the predictive power of models (although these are often used to fit the models to observations),
- the complex interplay of physical processes has to be simulated in full detail because in astrophysics (in contrast to other fields of physics) there are virtually no experiments and validation of theoretical models requires comparison to astronomical observations that result from a variety of physical effects.

Despite these shortcomings, divide-and-conquer approaches are widely used in astrophysical research. Often they are the only realistic way to tackle the complex multiscale multiphysics problems. It is, however, important to be aware of their limitations, and the current work aims at removing free parameters by multidimensional simulations that take into account as many of the relevant physical processes as possible (examples are given in Sect. 4 below).

Another useful approach is to base the physical description of processes on large spatial scales on effective theories. In particular, the concepts of thermodynamics are suitable to deal with the scale problem in densities by specifying equations of state for astrophysical matter and fluid dynamics helps to deal with the spatial scale problem.

3 Application to Stars

In the following, the theoretical description of stars will be discussed with a particular emphasis on fluid dynamical models. As will be detailed below, conventional approaches to stellar structure and evolution theory are based on simplifying assumptions that avoid the solution of multidimensional fluid dynamical problems. These have been instrumental for the understanding of stars over the past decades. New observations and the importance of stellar theory for other branches of astrophysics, however, call for improved modeling approaches.

The fundamental questions in stellar astrophysics include

1. What happens in stellar interiors (hidden to direct astronomical observations)?
2. How do stars evolve in time?
3. How do stars end their lives?

A fourth question, namely that of how stars form, can be addressed with similar approaches, but has its particular challenges. It is beyond the scope of this presentation and will not be discussed here further.

3.1 Why Stars?

Stellar objects are the fundamental building blocks of the visible universe. They form the smallest entities which are dynamically relevant to the evolution of larger cosmological structures. At the same time, stars make the universe accessible to astronomical observations. The largest part of knowledge we have about astrophysical and cosmological processes results from classical optical astronomy, to which stars are the primary targets.

Another important consequence of stellar evolution is the formation of heavy elements and the enrichment of the universe with them. In Big Bang nucleosynthesis, hydrogen and helium were generated almost exclusively. With very few exceptions, all heavier chemical species our world is made of were produced in stars, stellar explosions, or other processes involving stellar objects.

Stars form from the gravitational collapse of overdense regions in the interstellar medium and process stellar material under conditions of high densities and temperatures in nuclear reactions. Explosive nucleosynthesis in supernovae further contributes and expels the newly formed chemical elements back into the interstellar medium out of which the next generation of stars forms. This closes the "cosmic cycle of matter" that continuously enriches the chemical composition of material in galaxies.

Last, but not least, stars are by themselves fascinating physical objects. They feature matter under various thermodynamic conditions. In neutron stars, densities above that of nuclear matter are reached making them laboratories for fundamental physics. Physical processes in them have been and are a primary subject of astrophysical research.

3.2 Stellar Fluid Dynamics

A way to describe the structure and the evolution of stars is to augment the equations of fluid dynamics with suitable source terms:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{1}$$

$$\frac{\partial (\rho X_i)}{\partial t} + \nabla \cdot (\rho X_i \mathbf{v}) = -\rho \omega_{X_i}, \qquad i = 1, \ldots, N, \tag{2}$$

$$\frac{\partial (\rho \mathbf{v})}{\partial t} + \nabla \cdot \rho \mathbf{v} \otimes \mathbf{v} + \nabla p = -\rho \nabla \Phi, \tag{3}$$

$$\frac{\partial (\rho E)}{\partial t} + \nabla \cdot (\rho E \mathbf{v}) + \nabla (p \mathbf{v}) = -\rho \mathbf{v} \cdot \nabla \Phi + \rho S. \tag{4}$$

Here, the symbols for the fluid dynamical quantities bear their usual meanings. The large separation between the astrophysical scales of interest and the viscosity scale justifies the use of the Euler equations instead of modeling viscosity effects explicitly.

The system accounts for the multiphysics nature of the problem. Equation (2) describes the change of the mass fraction of one of the N species i due to reactions. In the case of stellar processes, these are usually *nuclear* reactions. The source term on the r.h.s. of Eq. (2) and the second term on the r.h.s. of Eq. (4) correspond to the implied change in composition and energy release/consumption, respectively. To determine them, an extended nuclear reaction network has to be calculated concurrently with the fluid dynamics. The source term on the r.h.s. of Eq. (3) and the first term on the r.h.s. of Eq. (4) are due to gravity. For these, the Poisson equation has to be solved. In principle, a variety of additional terms could be introduced to account for diffusion, conduction, magnetic fields, etc.

The set of Eqs. (1)–(4) is closed by an appropriate equation of state. For stellar matter under the moderate conditions as found in stellar interiors, it should include contributions from an ideal gas of nuclei, arbitrarily degenerate and relativistic electrons, radiation, Coulomb interactions, electron–positron pairs, and ionization. Usually, the resulting expressions are complicated and in practice not available in closed analytic form but rather as tabled values. The equation of state for neutron stars is not completely understood and an active field of research. Ionization effects further complicate the equation of state. They play a role in the least densest regions close to the surface of normal stars.

The set of Eqs. (1)–(4) gives rise to several distinct timescales. One of them is associated with hydrostatic equilibrium. Writing the momentum equation (3) in Lagrangian form,

$$\frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} = -\frac{1}{\rho} \nabla p - \nabla \Phi; \qquad \frac{\mathrm{D}}{\mathrm{D}t} \equiv \frac{\partial}{\partial t} + \mathbf{v} \nabla, \tag{5}$$

shows that stellar material is not accelerated if the force due to gravity is balanced by the force due to the pressure gradient,

$$\frac{1}{\rho} \nabla p = -\nabla \Phi, \tag{6}$$

and hence an equilibrium configuration is assumed. The timescale on which the system reacts to perturbations of this *hydrostatic equilibrium* can be derived from assuming pressure support to vanish instantaneously. This gives

$$\tau_{\mathrm{h}} = \sqrt{\frac{R^3}{GM}}. \tag{7}$$

The *hydrostatic timescale* τ_{h} for the Sun is about 20 minutes. Another timescale is introduced into the system by the nuclear reaction source terms in Eqs. (2) and (4). The *nuclear burning timescale* of species i, which is set by the corresponding reaction rates, is given as the rate of change of its mass fraction,

$$\tau_i^{\mathrm{nuc}} = \frac{X_i}{\dot{X}_i}. \tag{8}$$

In many situations, one finds $\tau_{\mathrm{nuc}} \gg \tau_{\mathrm{h}}$; i.e., the star reacts to any changes in its structure induced by nuclear burning on the very short hydrostatic timescale and restores equilibrium. Burning is then said to proceed in hydrostatic equilibrium. An example is hydrogen burning in main sequence stars, where

$$\tau_{\mathrm{H-burning}} \sim \frac{1}{10}\frac{0.007Mc^2}{L} \approx 10^{10}\left(\frac{M}{M_\odot}\right)\left(\frac{L_\odot}{L}\right)\mathrm{yr}. \tag{9}$$

Thus, for the Sun τ_{nuc} is on the order of 10 Gyr. In explosive events such as supernovae, in contrast, $\tau_{\mathrm{nuc}} \lesssim \tau_{\mathrm{h}}$ and burning proceeds as a dynamical process.

The velocities encountered in stellar objects are no less diverse. For meridional circulation, for instance, Mach numbers $M = c/|\mathbf{v}|$ of about 10^{-11} are typical (c denotes the speed of sound). Convective motions in stars feature $M \sim 10^{-4}$, and stellar explosions proceed in the transonic regime.

4 Toward a New Generation of Stellar Models

Vastly different spatial, temporal, and velocity scales are sometimes equally relevant for the evolution of a single stellar object. Approaches to solve the full multidimensional problem therefore challenge modeling concepts, numerical techniques, and supercomputational resources.

A basic approach, however, is to use a finite-volume discretization of Eqs. (1)–(4) on various grid geometries. The usual problems of appropriate boundary conditions and suitable initial conditions are particularly severe in the astrophysical context, because in open space no physical boundaries exist outside the stellar objects and initial conditions cannot be directly determined but have to be inferred (usually indirectly) from astronomical observations. For solving the fluid dynamics equations, standard Riemann solvers are employed and the source terms are treated in an operator

splitting approach. For the choice of the Riemann solver, it has to be accounted for the fact that astrophysical equation of state is usually not available in closed analytic from but given as tabulated values.

In the following, examples from stellar astrophysical research are given. Each of them illustrates an approach to deal with one (or more) of the challenges discussed above. Of course, many other applications and approaches exist, but a comprehensive review is beyond the scope of this contribution.

4.1 Example: Type Ia Supernova Simulations

Stellar explosions occurring as supernovae are a prominent example of a severe spatial scale problem (in combination with a pronounced multiphysics challenge). This is caused by the fact that the physical processes driving the explosions are microphysical and the largest scales of interest are given by the exploding stars themselves. Moreover, supernova explosion modeling is hampered by the fact that the exact physical conditions at the onset of the explosions are poorly known and therefore the initial conditions of the hydrodynamical explosion models are uncertain. Because stellar explosions take place in the transonic regime, the temporal and velocity-scale problems are modest compared to other astrophysical problems (see Sect. 4.2).

Core collapse supernovae mark the end stages of the evolution of massive stars ($M \gtrsim 8 \ldots 10\, M_\odot$). For these astrophysical explosions, the multiphysics problem is particularly challenging because of the large number of physical processes that are only partially understood and act simultaneously (see [1] for a recent review).

Here, we focus however on another class of supernovae, the so-called thermonuclear supernovae. These are associated with the astronomical class of Type Ia supernovae. These events are in the focus of astrophysical interest because of their application as distance indicators in the universe which led to the conclusion of its accelerated expansion [2, 3]. This is interpreted as being caused by an unidentified form of "dark" energy that dominates the energy content of today's universe. Moreover, they significantly contribute to the cosmic cycle of matter and the chemical enrichment of galaxies being responsible for producing the majority of iron in the universe.

The main problem with simulating the explosions of Type Ia supernovae is that there is no direct observation of the progenitor system available. These events are only detected while brightening in the explosion. It is therefore clear that the progenitor must be a relatively faint astronomical source. A number of arguments, however, indicate that Type Ia supernovae are caused by thermonuclear explosions of white dwarf stars consisting primarily of carbon and oxygen (see [4] for a review). Such white dwarf stars, however, are eternally stabilized by a quantum mechanical effect—the degeneracy of electrons in their matter—and it must be explained how they reach a critical state for the explosion. It is suspected that interaction with a binary companion triggers the explosion. Unfortunately, the state of the white dwarf at the onset of the

supernova and the nature of the companion and the interaction remain unclear and no astronomical effort could thus far identify them beyond doubt.

Therefore, hydrodynamical simulations of thermonuclear supernova explosions have to start out from guesses of the initial conditions. Over the recent years, pipelines of models were established (see, e.g., [5]) that aim at removing free parameters from the description of the involved physical processes by performing simulations in multiple spatial dimensions. Earlier one-dimensional approaches, in contrast, assumed spherical symmetry and introduced free tunable parameters in describing physical processes. The consistent multidimensional treatment allows the faithful prediction of observables from the models that can be compared directly to astronomical data. This way it is possible to assess the validity of modeling assumptions and in particular the guess of the initial conditions.

The modeling of the thermonuclear species conversion and energy release (usually called "thermonuclear burning") is challenged by the wide range of physically relevant length scales. The initializing nuclear reaction, the fusion of ^{12}C nuclei, is extremely sensitive to temperature. Under the relevant conditions, its rate scales with $\sim T^{20}$. This implies that burning is only significant in the thin layer where temperature peaks. Therefore, it advances as a thin combustion wave. The width of this wave is very small (millimeters to centimeters) compared to the overall scale of interest for the explosion simulations (the radius of a white dwarf is several thousand kilometers).

This justifies to model combustion waves as discontinuities separating the fuel material from the nuclear "ashes." The jump conditions corresponding to these weak solutions imply that there are two distinct modes for flame propagation: *subsonic deflagrations* and *supersonic detonations*. These are also distinct in the microphysical mechanisms of flame advancement. While deflagrations are mediated by the thermal conduction of the electrons that heat up material in front of the reaction zone and give rise to a subsonic flame speed, detonations are driven by shock waves (see [6] for details of combustion in thermonuclear supernovae). Both processes cannot be resolved in numerical simulations covering the entire exploding star.

One way to model the propagation of deflagrations [7] is based on the level-set technique [8]. It allows to propagate the discontinuity representing the unresolved deflagration front with a given speed. This speed on the smallest scales is set by microphysical transport. On larger, but still unresolved scales, however, the mechanism that determines the flame velocity is turbulent acceleration. This is because the flame propagates from the center of the white dwarf star outwards and leaves behind light and hot ashes creating an inverse density stratification in the gravitational field of the star. This setup is buoyancy-unstable, and Rayleigh–Taylor instability with secondary Kelvin–Helmholtz instabilities at the flame front generates turbulent eddies. These decay in an energy cascade down to unresolved scales. On a wide range of scales, the flame is dragged around by turbulent eddies of various sizes ("flamelet regime of turbulent combustion," see [6, 9]). This increases the flame surface and leads to strong acceleration. The efficiency of the nuclear burning is set by this effect, and a valid model must take it into account correctly. The resulting scale problem can be addressed in a large-eddy simulation (LES) approach. A turbulent subgrid-scale

model is employed to determine the effective flame speed near the grid scale, and this serves as an input for the level-set-based deflagration flame model (e.g., [10–13]).

While in many astrophysical simulations the spatial scale problem is tackled in adaptive mesh refinement (AMR) approaches, for the case of thermonuclear supernovae a combination of level-set and LES techniques has proven particularly efficient.

4.2 Example: Multidimensional Models of Stellar Interiors

The main challenge of multidimensional models for astrophysical processes in stellar interiors is the wide range of temporal scales involved. Additionally, the associated motions of stellar material often proceed at low velocities. Therefore, a standard approach in stellar evolution theory is to simplify matters by assuming spherical symmetry (see [14] for a standard treatment). It is usually assumed that stellar evolution can be approximated by a sequence of hydrostatic equilibrium solutions. These change slowly due to contraction and nuclear reactions releasing gravitational binding energy and nuclear energy that replenish the loss from the stellar surface by radiation.

The resulting one-dimensional hydrostatic model, however, leads to a coarse description and parametrization of inherently multidimensional and dynamical effects such as energy transport by convection and mixing by instabilities. It lacks predictive power as parameters can be tuned to fit the observations. Nonetheless, the limits of parametrization seem to be reached with modern observations of stars (in particular the detailed data on stellar abundances and asteroseismology) that are in tension with theoretical expectations.

Full multidimensional simulations of the evolution of stars from their formation to the end of their life cycle (for some stars explosions as supernovae) are out of reach for current theoretical modeling and computational resources. Well-chosen evolutionary phases, or isolated processes inside stars, however, become accessible to multidimensional simulations owing to progress in modeling techniques and the ever-increasing power of supercomputers. Such simulations can help to make one-dimensional stellar evolution calculations more reliable by fixing free parameters.

Multidimensional hydrodynamical modeling of stellar processes is based on the set of Eqs. (1)–(4) with suitable source terms. A timescale problem arises from the discrepancy between the dynamical timescale and the much longer nuclear timescales. Moreover, flows in stellar interiors are characterized by extremely low Mach numbers and they act over long time spans. This requires special numerical approaches that are able to deal with low-Mach number fluid dynamics and can cover long periods of time with reasonable computational effort. Formal requirements for suitable schemes are stated and discussed in detail in [15].

For an approach to the low-Mach number problem, consider the homogeneous Euler equations for simplicity. Non-dimensionalization introduces a reference Mach number M_r (see [15–17]). In the zero Mach number limit, two decoupled solution

spaces are found [15, 18–20]: For incompressible solutions, the velocity fluctuations scale with M_r^2, whereas for sound waves they scale linearly with M_r.

At low Mach numbers, standard Riemann solvers show excessive dissipation. This can be interpreted as a result of unresolved artificial sound waves caused by the Riemann problems constructed at the cell interfaces. An example for this is shown in Fig. 4 of [16] where the Gresho vortex [21] is followed with a standard Roe flux discretization. For a maximum Mach number in the setup of 10^{-1}, the vortex is preserved; for 10^{-2}, it becomes blurred; and below 10^{-3}, it is not recognizable after one revolution of the vortex. The kinetic energy of the Gresho vortex is dissipated at a higher rate for lower Mach numbers (see Fig. 6 of [16]).

A way to avoid excessive dissipation at low Mach numbers is *flux preconditioning* [22], where the upwinding (numerical dissipation) term is modified with a suitable preconditioning matrix (an alternative approach is to modify the underlying equations as implemented in the astrophysical simulation code MAESTRO [23]). Several such matrices are possible, but not all correct the low-Mach number scaling of the dissipation term fully. These are not suitable for astrophysical flows where source terms due to gravity are important (see the discussion in [15–17]). Therefore, [16] suggested a new low-Mach number preconditioning matrix that ensures correct scaling:

$$P_\mathbf{V} = \begin{pmatrix} 1 & n_x \frac{\rho\delta M_\mathrm{r}}{c} & n_y \frac{\rho\delta M_\mathrm{r}}{c} & n_z \frac{\rho\delta M_\mathrm{r}}{c} & 0 \\ 0 & 1 & 0 & 0 & -n_x \frac{\delta}{\rho c M_\mathrm{r}} \\ 0 & 0 & 1 & 0 & -n_y \frac{\delta}{\rho c M_\mathrm{r}} \\ 0 & 0 & 0 & 1 & -n_z \frac{\delta}{\rho c M_\mathrm{r}} \\ 0 & n_x \rho c \delta M_\mathrm{r} & n_y \rho c \delta M_\mathrm{r} & n_z \rho c \delta M_\mathrm{r} & 1 \end{pmatrix} \tag{10}$$

where $\delta = \frac{1}{\mu} - 1$ and $\mu = \min[1, \max(M_\mathrm{local}, M_\mathrm{cut})]$. Here M_local is the local Mach number at the cell interface, which is limited to a lower value of M_cut to avoid singularity of $P_\mathbf{V}$. Tests down to Mach numbers of 10^{-10} [15] demonstrate that with this preconditioning the numerical dissipation is low and independent of the Mach number.

The application to stellar models requires to consider *reactive* fluid dynamics *with gravity* as in Eqs. (1)–(4). This is implemented in the Seven-League Hydro (SLH) code that solves the compressible Euler equations with source terms in one, two, and three spatial dimensions. In addition to flux preconditioning, it features other techniques to deal with low-Mach number flows [24]. It allows for explicit and implicit time discretization (for an alternative approach to multidimensional modeling of stellar interiors with time-implicit methods see [25, 26]). The physics modules of the SLH code include radiation in the diffusion limit, a general equation of state for stellar matter, and a general nuclear reaction network solver. The code features flexible grid geometries based on arbitrary curvilinear meshes [27]. These allow for computational grids adapted to overall spherical stellar objects avoiding grid singularities at the center, as would be the case for spherical coordinates.

Implicit time discretization is essential for practical simulations of low-Mach number flows, because otherwise the CFL stability criterion would restrict the time step to prohibitively short intervals. For accuracy reasons, however, it should be set according to the fluid flow timescale, which still results in a gain in time step duration of $1/M$. Tests imply that implicit time stepping becomes more efficient already at moderately low Mach numbers of M 0.1 . . . 0.2. The implicit time stepping in SLH uses "Explicit first stage, Singly Diagonally Implicit Runge–Kutta" (ESDIRK) schemes [28] of second to fifth order. The nonlinear solver is based on Newton–Raphson iteration with the last Runge–Kutta stage as initial guess. The computation of the Jacobian uses automatic differentiation: Each quantity carries its derivatives with respect to every independent variable. The linear solver is based on iterative techniques (multigrid and/or Krylov methods). The implementation has been successfully tested for scaling to large numbers of processors (up to half a million cores [29]), which is a prerequisite for its applicability to astrophysical problems.

Particular care is required when setting up stellar profiles resulting from one-dimensional stellar evolution calculations as initial conditions for multidimensional hydrodynamical simulations. The mapping from the one-dimensional model to a three-dimensional setup is non-trivial, because multidimensional and dynamical effects have been parametrized in an inconsistent way. Moreover, the models are usually very close to hydrostatic equilibrium. Hydrodynamical models do not automatically guarantee this delicate balance between gravity and pressure gradient, and maintaining it over many dynamical timescales is challenging. An approach to deal with this problem is to use well-balancing techniques (e.g., [30, 31]).

An example application of the SLH code to stellar modeling is the simulation of convective mixing in Population-III stars. These are supposed to be the first stars formed in the universe. Therefore, they are made of pristine Big Bang nucleosynthesis (BBN) material, i.e., a mixture consisting almost exclusively of hydrogen and helium. At some stage of their evolution, the fusion of hydrogen to helium runs out of fuel in the stellar core and the next nuclear burning phase—the fusion of helium forming carbon—commences, while in a shell surrounding the core hydrogen is burned. Both burning regions produce energy at a high rate. This causes convection in them.

For stellar hydrogen burning, two modes exist: a direct fusion between hydrogen nuclei (protons)—called "pp-chain"—and a cyclic reaction sequence involving carbon, nitrogen, and oxygen nuclei as catalysts—called "CNO-cycle." At high enough temperatures, as found in the hydrogen-burning shell of the considered Population-III star model, burning in the CNO-cycle is much more efficient in energy production than the pp-chain. For it to become active, the catalyst carbon nuclei must be in place in sufficient abundances. These are not contained in the pristine BBN material the star is made of. Carbon is produced, however, in convective helium burning in the core, and the question is whether this material can reach the hydrogen-burning shell. In-between this shell and the core, there is a non-convective layer. One-dimensional stellar evolution models are unable to predict whether convective motions in the core and the hydrogen-burning shell are able to mix carbon-rich material through this non-convective layer, because convection and convective overshooting is treated in a

highly parametrized way. Three-dimensional hydrodynamical simulations that treat this process consistently are work in progress with the SLH code.

In such simulations, the velocity-scale problem is addressed with low-Mach number techniques, for which flux preconditioning is one possible approach, and the timescale problem is accounted for by implicit time discretization.

4.3 Example: Common Envelope Phases in Binary Stellar Evolution

As a third example of fluid dynamical simulations in stellar astrophysics, a particular phase in binary stellar evolution is discussed. More than half of the stars in the universe are found in multiples, and some are close enough that binary interaction takes place. Usually, the more massive star evolves faster than its less massive companion and enters the giant phase earlier. In this phase, a very dense stellar core is formed inside a dilute and very extended envelope. This envelope may reach and even swallow the companion star, resulting in an object, in which two stellar cores revolve inside a *common envelope.* The outcome of this interaction is transfer of energy from the two orbiting cores to the envelope that is eventually ejected leaving behind a close binary system of compact stars. The common envelope phase has important consequences for progenitor systems of supernova explosions and for compact binaries that are sources of gravitational waves.

Obviously, there is no inherent symmetry in this problem and conventional one-dimensional binary stellar evolution codes use parametrized descriptions that severely reduce their predictive power. Past attempts to simulate this common envelope phase were based on smooth particle hydrodynamics (SPH) or employed static grids (some approaches used adaptive mesh refinement)—see [32] for a recent review.

Generally, the common envelope phase challenges numerical simulations by the wide ranges of involved time and space scales, but in addition it is a problem that requires Galilean invariance of the employed scheme.

Static mesh approaches do not comply with the "Lagrangian nature" of the problem. They are not Galilean invariant. Following several stable orbits of the stellar cores is difficult with such schemes. SPH has problems resolving fluid dynamical instabilities. Moreover, due to its mass-adaptive nature, it does not allow for high resolution in the dilute envelope, which, however, is of primary interest to solve the problem.

A technique that avoids the disadvantages of both approaches while it nearly retains the Lagrangian nature of SPH and allows for the high resolution of grid-based methods is that of moving meshes. In astrophysics, this technique was pioneered by the cosmological code AREPO [33]. This code was modified and applied to the problem of common envelope evolution [34]. A version of the AREPO code that follows the evolution of magnetic fields [35] was recently also employed to simulations of the common envelope phase in binary stellar evolution [36].

5 Conclusions

Astrophysical simulations are often based on hydrodynamics and complicated by the various source terms that represent the numerous physical processes at work in complex astrophysical objects. Apart from this multiphysics problem, the wide ranges of relevant scales in time, space, and velocity render astrophysical simulations a challenge (and at the same time an ideal test bed) for numerical schemes. A combination of new numerical techniques and efficient parallelization allowed to use some of the world's fastest supercomputers to study questions of stellar astrophysics. Examples were given in this chapter. They illustrate the challenges of fluid dynamical modeling in astrophysics, the design of new methods, and the application to the simulation of processes in stellar astrophysics.

Progress of theoretical (stellar) astrophysics cannot solely rely on the ever-increasing power of supercomputers. The challenging nature of these problems will always require to go to the limits of numerical methods and computational resources. It is essential to develop new mathematical approaches and numerical schemes that account for the specific needs of astrophysical simulations. General-purpose hydrodynamics codes have only restricted applicability to routine simulations, while breakthroughs require innovative approaches. Close collaboration between applied mathematics and computational astrophysics has proven fruitful and paves the way to future developments.

Acknowledgements FKR thanks the organizers of the "Hyp2016" conference for the kind invitation to this stimulating event. Collaboration and discussions with the fluid mechanics group at the Mathematics Department of the University of Würzburg—in particular with Christian Klingenberg, Markus Zenk, Wasilij Barsukow, and Jonas Berberich—are gratefully acknowledged. The examples discussed in this chapter reflect in large parts results of the work of Fabian Miczek, Philipp Edelmann, Sebastian Ohlmann, and many others in the supernova/stellar astrophysics group at the Max Planck Institute for Astrophysics, Garching, the Astrophysics Group at the University of Würzburg, and the "Physics of Stellar Objects" group at the Heidelberg Institute for Theoretical Studies. The simulations of common envelope phases were carried out in close collaboration with Rüdiger Pakmor and Volker Springel. The work of FKR is supported by the Klaus Tschira Foundation.

References

1. H.T. Janka, Ann. Rev. Nucl. Part. Sci. **62**, 407 (2012)
2. A.G. Riess, A.V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P.M. Garnavich, R.L. Gilliland, C.J. Hogan, S. Jha, R.P. Kirshner, B. Leibundgut, M.M. Phillips, D. Reiss, B.P. Schmidt, R.A. Schommer, R.C. Smith, J. Spyromilio, C. Stubbs, N.B. Suntzeff, J. Tonry, AJ **116**, 1009 (1998)
3. S. Perlmutter, G. Aldering, G. Goldhaber, R.A. Knop, P. Nugent, P.G. Castro, S. Deustua, S. Fabbro, A. Goobar, D.E. Groom, I.M. Hook, A.G. Kim, M.Y. Kim, J.C. Lee, N.J. Nunes, R. Pain, C.R. Pennypacker, R. Quimby, C. Lidman, R.S. Ellis, M. Irwin, R.G. McMahon, P. Ruiz-Lapuente, N. Walton, B. Schaefer, B.J. Boyle, A.V. Filippenko, T. Matheson, A.S.

Fruchter, N. Panagia, H.J.M. Newberg, W.J. Couch, The Supernova Cosmology Project, ApJ **517**, 565 (1999)
4. W. Hillebrandt, J.C. Niemeyer, ARA&A **38**, 191 (2000)
5. W. Hillebrandt, M. Kromer, F.K. Röpke, A.J. Ruiter, Front. Phys. **8**, 116 (2013)
6. F. K. Röpke, W. Schmidt, in *Interdisciplinary Aspects of Turbulence*, Lecture Notes in Physics, ed. by W. Hillebrandt, F. Kupka (Springer, Berlin, 2009), pp. 255–289
7. M. Reinecke, W. Hillebrandt, J.C. Niemeyer, R. Klein, A. Gröbl, A&A **347**, 724 (1999)
8. S. Osher, J.A. Sethian, J. Comput. Phys. **79**, 12 (1988)
9. N. Peters, *Turbulent Combustion* (Cambridge University Press, Cambridge, 2000)
10. M. Reinecke, W. Hillebrandt, J.C. Niemeyer, F. Röpke, W. Schmidt, D. Sauer, in *Proceedings of the 11th Workshop on "Nuclear Astrophysics", Ringberg Castle*, ed. by W. Hillebrandt, E. Müller (Max-Planck-Institut für Astrophysik, Garching, 2002), MPA/P13, pp. 54–56
11. F.K. Röpke, W. Hillebrandt, A&A **431**, 635 (2005)
12. F.K. Röpke, W. Hillebrandt, W. Schmidt, J.C. Niemeyer, S.I. Blinnikov, P.A. Mazzali, ApJ **668**, 1132 (2007)
13. M. Fink, M. Kromer, I.R. Seitenzahl, F. Ciaraldi-Schoolmann, F.K. Röpke, S.A. Sim, R. Pakmor, A.J. Ruiter, W. Hillebrandt, MNRAS **438**, 1762 (2014)
14. R. Kippenhahn, A. Weigert, A. Weiss, *Stellar Structure and Evolution* (Springer, Berlin, 2012)
15. W. Barsukow, P.V.F. Edelmann, C. Klingenberg, F. Miczek, F.K. Röpke, J. Sci. Comput. 1–24 (2017)
16. F. Miczek, F.K. Röpke, P.V.F. Edelmann, A&A **576**, A50 (2015)
17. W. Barsukow, P.V.F. Edelmann, C. Klingenberg, F.K. Röpke, in *Workshop on Low Velocity Flows, Paris, 5–6 November 2015, ESAIM: Proceedings and Surveys*, ed. by S. Dellacherie, et al., vol. 56 (2017). In print
18. S. Dellacherie, J. Comput. Phys. **229**(4), 978 (2010)
19. S. Schochet, J. Differ. Equ. **114**(2), 476 (1994)
20. S. Klainerman, A. Majda, Commun. Pure Appl. Math. **34**(4), 481 (1981)
21. P.M. Gresho, S.T. Chan, Int. J. Numer. Methods Fluids **11**(5), 621 (1990)
22. E. Turkel, Ann. Rev. Fluid Mech. **31**, 385 (1999)
23. A.S. Almgren, J.B. Bell, M. Zingale, J. Phys. Conf. Ser. **78**(1), 012085 (2007)
24. M.S. Liou, J. Comput. Phys. **214**(1), 137 (2006)
25. M. Viallet, I. Baraffe, R. Walder, A&A **531**, A86 (2011)
26. M. Viallet, T. Goffrey, I. Baraffe, D. Folini, C. Geroux, M.V. Popov, J. Pratt, R. Walder, A&A **586**, A153 (2016)
27. K. Kifonidis, E. Müller, A&A **544**, A47 (2012)
28. C.A. Kennedy, M.H. Carpenter, Additive Runge–Kutta schemes for convection-diffusion-reaction equations. Technical report, NASA Technical Memorandum (2001)
29. N. Hammer, F. Jamitzky, H. Satzger, M. Allalen, A. Block, A. Karmakar, M. Brehm, R. Bader, L. Iapichino, A. Ragagnin, V. Karakasis, D. Kranzlmüller, A. Bode, H. Huber, M. Kühn, R. Machado, D. Grünewald, P.V.F. Edelmann, F.K. Röpke, M. Wittmann, T. Zeiser, G. Wellein, G. Mathias, M. Schwörer, K. Lorenzen, C. Federrath, R. Klessen, K. Bamberg, H. Ruhl, F. Schornbaum, M. Bauer, A. Nikhil, J. Qi, H. Klimach, H. Stüben, A. Deshmukh, T. Falkenstein, K. Dolag, M. Petkova in *Parallel Computing: On the Road to Exascale, Proceedings of the International Conference on Parallel Computing, ParCo 2015, 1–4 September 2015, Edinburgh, Scotland, UK*, ed. by G.R. Joubert, H. Leather, M. Parsons, F.J. Peters, M. Sawyer. Advances in Parallel Computing, vol. 27 (IOS Press, 2016), pp. 827–836
30. P. Chandrashekar, C. Klingenberg, SIAM J. Sci. Comput. **37**(3), B382 (2015)
31. V. Desveaux, M. Zenk, C. Berthon, C. Klingenberg, Int. J. Numer. Methods Fluids **81**(2), 104 (2016)
32. N. Ivanova, S. Justham, X. Chen, O. De Marco, C.L. Fryer, E. Gaburov, H. Ge, E. Glebbeek, Z. Han, X.D. Li, G. Lu, T. Marsh, P. Podsiadlowski, A. Potter, N. Soker, R. Taam, T.M. Tauris, E.P.J. van den Heuvel, R.F. Webbink, A&A Rev. **21**, 59 (2013)
33. V. Springel, MNRAS **401**, 791 (2010)
34. S.T. Ohlmann, F.K. Röpke, R. Pakmor, V. Springel, ApJ **816**(1), L9 (2016)
35. R. Pakmor, A. Bauer, V. Springel, MNRAS **418**, 1392 (2011)
36. S.T. Ohlmann, F.K. Röpke, R. Pakmor, V. Springel, E. Müller, MNRAS **462**(1), L121 (2016)

Nonlinear Stability of Localized and Non-localized Vortices in Rotating Compressible Media

Olga S. Rozanova and Marko K. Turzynsky

Abstract We study nonlinear stability of steady isolated vortices in two-dimensional compressible media in an uniformly rotating reference frame. First, we consider a vortex with linear profile of velocity. Its behavior can be completely described by a quadratically nonlinear system of ODEs. We find that the stability property depends only on one parameter, the ratio of relative vorticity of vortex to the Coriolis constant. We find the domain of this parameter ensuring nonlinear stability. Further, we consider more general class of isolated steady vortices, containing decaying at infinity and compactly supported vortices as particular cases. At every point of the plane, this isolated steady vortex can be approximated by a solution with linear profile of velocity. Thus, at every point of the plane, there arises a nonlinear system of ODEs with initial data generated by derivatives of the steady vortex state. It is hypothesized that if at every point the solution to this ODEs system falls in the domain of attraction of an equilibrium, then the steady vortex is nonlinearly stable. We compare this nonlinear stability hypothesis with Rayleigh criterium of linearized stability with respect to radial perturbation. In particular, we find that the rotation has a stabilizing effect.

Keywords Compressible media · Rotation · Vortex motion · Stability · Blow up

1 Introduction

The vortex motion in rotating fluid is important due to application in geophysical models, and a review of the state of art can be found in [8, 11]. In the middle scale approximation, the geophysical motion can be considered on a l-plane (i.e.,

O. S. Rozanova (✉) · M. K. Turzynsky
Department of Mechanics and Mathematics, Moscow State University,
Moscow 119991, Russia
e-mail: rozanova@mech.math.msu.su

M. K. Turzynsky
e-mail: M13041@yandex.ru

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_41

on a plane tangent to the Earth surface at a fixed point). In this approximation, the Coriolis parameter l is constant and the coordinate frame is uniformly rotating. Most important example of middle scale atmospherical vortex motion is a hurricane. Since the vortex dynamics can be complicated, it is natural to begin with the study of certain elementary processes. One example is the isolated circular free vortex. Its stability/instability properties are of fundamental interest due to the presence of strain and shear in the ambient flow. A huge literature, both theoretical and experimental, is devoted to stability of vortices in incompressible media [15]. Despite their importance, the compressible vortices are studied to a lesser extent [1, 21]. There exist numerical works concerning stability of isolated compressible vortices [10, 16, 17]; however from the theoretical point of view, only linear analysis of stability is performed [6, 7].

This paper is a step to analysis of nonlinear stability of compressible 2D vortices in a uniformly rotating frame. A very particular case of this vortex is the motion with a linear profile of velocity. This class of vortices can be considered in dynamical setting, since the problem is reduced to studying of a nonlinear system of ODEs. The equilibrium (stable or unstable) of this system corresponds to steady vortex. Further, there exists a large class of steady compressible vortices [20]. Any other steady vortex, localized or not, can be approximated at every point of plane by a vortex with linear profile of velocity. We study the solution to the Cauchy problem for the respective ODEs systems and notice that if all their solutions are periodic (fall in the basin of attraction of some stable equilibrium), then the steady state satisfies known necessary stability conditions. This allows us to make a guess that the above property of solutions of approximating ODEs systems can be a criterion of nonlinear stability for the steady-state vortex.

2 Bidimensional Models of Rotating Compressible Medium

The two-dimensional system of motion of inviscous compressible barothropic medium on a rotating plane consists of three equations for density $\rho(t, x)$, velocity $\mathbf{U}(t, x) = (u_1, u_2)$ and pressure $p(t, x)$:

$$\rho(\partial_t \mathbf{U} + (\mathbf{U} \cdot \nabla)\mathbf{U} + \mathscr{L}\mathbf{U}) + \nabla p = 0, \tag{1}$$

$$\partial_t \rho + \operatorname{div}(\rho \mathbf{U}) = 0, \tag{2}$$

where $p = \mathscr{C}\rho^\gamma$, $\mathscr{C} = \text{const}$. Here, $\mathscr{L} = lL$, $L = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, $\gamma \in (1, 2]$ is the heat ratio, $l > 0$ is the Coriolis parameter. In geophysical applications, the centrifugal force is included in geopotential and disappears under averaging over the height due to hydrostatic balance [9, 18].

Let us introduce a new variable $\Pi = p^{\frac{\gamma-1}{\gamma}}$ and reduce (1), (2) to

$$\partial_t \mathbf{U} + (\mathbf{U}\cdot\nabla)\mathbf{U} + \mathscr{L}\,\mathbf{U} + c_0\,\nabla\Pi = 0, \tag{3}$$

$$\partial_t \Pi + (\nabla\Pi\cdot\mathbf{U}) + (\gamma-1)\,\Pi\,\mathrm{div}\,\mathbf{U} = 0, \tag{4}$$

$c_0 = \frac{\gamma}{\gamma-1}\mathscr{C}^{\frac{1}{\gamma}}$. This system was used in [18, 19] in geophysical context.

3 Non-localized Vortex

We look for the solution of (3), (4) in the form

$$\mathbf{U}(t,\mathbf{x}) = Q\mathbf{x}, \qquad Q = \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix}, \tag{5}$$

$$\Pi(t,\mathbf{x}) = A(t)x_1^2 + B(t)x_1x_2 + C(t)x_2^2 + K(t). \tag{6}$$

In fact, the solution is the first term of the Taylor series expansion at the point of minimum or maximum of pressure. In this way, we can keep a maximum possible members in this expansion to obtain an exact solution of (3) and (4).

Thus, we get a closed ODE system for the components of the matrices Q and $R = \begin{pmatrix} A(t) & \frac{1}{2}B(t) \\ \frac{1}{2}B(t) & C(t), \end{pmatrix}$:

$$\dot{R} + RQ + Q^TR + (\gamma-1)\,R\,\mathbf{tr}\,Q = 0, \tag{7}$$

$$\dot{Q} + Q^2 + lLQ + 2c_0R = 0, \tag{8}$$

$$\dot{K} + 2(\gamma-1)\,K\,\mathbf{tr}\,Q = 0.$$

The last equation is linear with respect to K, whereas the system of matrix equations (7), (8) consists of seven nonlinear ODEs and has a very complicated behavior.

As one can check, the point

$$a = d = 0,\ b = -c = b^*,\ A = C = A* = \frac{b^*(b^*-l)}{2c_0},\ B = 0 \tag{9}$$

is the only equilibrium of system (7), (8). The degenerate case $A^* = 0$ of constant pressure corresponds to the case $b^* = 0$ or $b^* = l$.

Direct computations show that the following properties hold.

Theorem 1. *System (7) has three first integrals:*

$$(b-c-l)\mathscr{D}^{-\frac{1}{2\gamma}} = I_1, \tag{10}$$

$$((d-a)B + 2bA - 2cC - l(A+C))\mathscr{D}^{-\frac{\gamma+1}{2\gamma}} = I_2, \tag{11}$$

$$((a^2+c^2)C + (b^2+d^2)A + (ac+bd)B - \frac{4c_0}{\gamma-1}\mathscr{D})\mathscr{D}^{-\frac{\gamma+1}{2\gamma}} = I_3, \tag{12}$$

where $\mathscr{D} = AC - B^2/4$.

3.1 Axisymmetric case

System (7), (8) has a closed submanifold of solutions with additional properties $a = d$, $c = -b$, $A = C$, $B = 0$. These solutions correspond to the axisymmetric motion. Here, we get a system of three ODEs:

$$\dot{A} + 2\gamma a A = 0, \quad \dot{a} + a^2 - b^2 + lb + 2c_0 A = 0, \quad \dot{b} + 2ab - la = 0. \tag{13}$$

If $A \neq 0$, then (10) implies $b(t) = \frac{l}{2} + \mathscr{C}|A(t)|^{\frac{1}{\gamma}}$, $\mathscr{C} = \text{const}$, and (13) can be reduced to the following system:

$$\dot{A}(t) = -2\gamma a A, \qquad \dot{a}(t) = -a^2 - \frac{l^2}{4} + \mathscr{C}^2 A^{\frac{2}{\gamma}} - 2c_0 A.$$

On the phase plane (A, a), there always exists a unique equilibrium $(A^*, a^*) = (A_0, 0)$, stable in the Lyapunov sense (a center), where A_0 is a root of equation $\frac{l^2}{4} + 2c_0 A = \mathscr{C}^2 A^{\frac{2}{\gamma}}$. This case is considered in [18].

If $\mathscr{C} = 0$, we have a particular case of motion with a constant vorticity $2b = l$.

If $A = 0$, then (13) can be reduced to one Riccati equation $\dot{z} = -z^2 + ilz$ for the function $z(t) = a(t) + ib(t) : \mathbb{C} \mapsto \mathbb{C}$. It has an explicit solution $z = \frac{lz(0)}{(l+iz(0))e^{-ilt} - Iz(0)}$, $a = \Re z$, $b = \Im z$. In the last case, the pressure is constant; nevertheless, the motion is vortical.

3.2 Range of Instability

Theorem 2. *If* $\frac{b^*}{l} < \frac{1-\sqrt{2}}{2}$ *or* $\frac{b^*}{l} > \frac{1+\sqrt{2}}{2}$, *then the equilibrium of system (7), (8) is unstable.*

Proof. Integral (10) is sufficiently simple and can be used to reduce (7), (8) to the system of six equations, and the equation for $b(t)$ will be excluded. The eigenvalues of matrix corresponding to the linearization at the equilibrium point of this system are the following:

$$\lambda_{1,2} = \pm\sqrt{-(2(2-\gamma)b^*(b^*-l) + l^2)},$$

$$\lambda_{3,4,5,6} = \pm\sqrt{2}\sqrt{-l\left(b^* + \frac{l}{4}\right) \pm \sqrt{\left(b^* + \frac{l}{2}\right)^2\left(\frac{l^2}{4} + b^*l - (b^*)^2\right)}}.$$

Since $(2-\gamma)b^*(b^*-l) + l^2 > 0$ for $\gamma \in (1,2]$, then $\Re(\lambda_{1,2}) = 0$. Eigenvalues λ_i, $i = 3,4,5,6$, have zero real part if and only if b^* satisfies the following inequalities simultaneously: $l(b^* + \frac{l}{4}) \geq 0$, $\frac{l^2}{4} + b^*l - (b^*)^2 > 0$, $l^2\left(b^* + \frac{l}{4}\right)^2 > \left(b^* + \frac{l}{2}\right)^2\left(\frac{l^2}{4} + b^*l - (b^*)^2\right)$, that is $b^* \in \left[\frac{1-\sqrt{2}}{2}\,l,\ \frac{1+\sqrt{2}}{2}\,l\right]$. For other b^*, the eigenvalues $\lambda_{3,4,5,6} = \pm\alpha \pm i\beta$, $\alpha \neq 0$, $\beta \neq 0$; therefore, there exists an eigenvalue with a positive real part. Thus, the Lyapunov theorem implies instability of the equilibrium for $b^* < \frac{1-\sqrt{2}}{2}\,l$ and $b^* > \frac{1+\sqrt{2}}{2}\,l$.

Remark 1. If the coordinate system is not rotating ($l = 0$), then the vortex is always cyclonic (the anticyclonic domain shrinks as $l \to 0$). The equilibrium point is unstable in the Lyapunov sense both in axisymmetric and general case. Nevertheless, in the axisymmetric case the equilibrium has a type of stable/unstable node and is quasi-asymptotically stable, whereas in general case the matrix of linearization has eigenvalues with nonzero real parts [5]. Thus, we can see that the rotation has a stabilizing effect.

3.3 Range of Stability

Theorem 3. *For $\frac{b^*}{l} \in (0,1)$, the equilibrium (9) is stable in the Lyapunov sense.*

Proof. Let us consider the function $\Lambda(a,b,c,d,A,B,C) = b^*I_2 - I_3 - \Lambda_0$, where I_2 and I_3 are given by equalities (11) and (12), and the constant Λ_0 is the value of Λ at the equilibrium point (9). Straightforward computation shows that Λ is the Lyapunov function. Indeed, at the equilibrium point $\Lambda = 0$ the first derivatives of Λ vanish, the matrix of the second derivatives is positive definite. Namely, the eigenvalues if this matrix (divided by $\frac{\partial^2\Lambda}{\partial a^2}$) are the following: $\{1,1,1,1,-\frac{c_0}{2A^*},-\frac{c_0}{A^*},-\frac{c_0}{2\gamma A^*}\}$. This implies that there exists a neighborhood $\mathcal{V}$ of the equilibrium point such that $\Lambda > 0$ in $\mathcal{V}$ everywhere except of the equilibrium point. Moreover, the total derivative of Λ by virtue of system (7), (8) is zero. According to the Lyapunov theorem, this implies stability of the equilibrium point (9).

3.4 Range of Possible Stability

If $\frac{b^*}{l} \in \Sigma = \left(\frac{1-\sqrt{2}}{2}, 0\right] \cup \left[1, \frac{1+\sqrt{2}}{2}\right)$, then we have only necessary condition for the stability of equilibrium, since the matrix, corresponding to the system, linearized at

the equilibrium, has three pairs of purely imaginary complex conjugate roots (see proof of Theorem 2). They can be written as $\pm i\,\omega_j$, $j = 1, 2, 3$, $\omega_j \in \mathbb{R}$.

We are going to prove that in the case of rationally independent frequencies almost all trajectories in ε-neighborhood of the equilibrium are quasi-periodic. This means that the equilibrium is "practically" stable in the Lyapunov sense. We apply a semi-analytical method based on the Bibikov theorem [3] (Theorem 15.5).

3.4.1 Semi-analytical Method to Prove Stability: Non-resonant Frequencies

Let us consider system

$$\dot{X} = \mathscr{A}X + \mathscr{P}(X), \quad X \in \mathbb{R}^n, \tag{14}$$

where a constant matrix A has purely imaginary eigenvalues $\pm i\omega_k$, $k = 1, \ldots, m$, $n = 2m$, the frequencies ω_j are rationally independent, the vector-valued function $\mathscr{P}(X)$ does not contain free and linear terms. The system can be written in diagonalized form

$$\dot{y}_k = i\omega_k y_k + Y_k(y, \bar{y}), \qquad \dot{\bar{y}}_k = -i\omega_k \bar{y}_k + \bar{Y}_k(y, \bar{y}). \tag{15}$$

Further, (15) is formally equivalent to its normal form

$$\dot{y}_k = y_k(i\omega_k + P_k(y, \bar{y})), \qquad \dot{\bar{y}}_k = \bar{y}_k(-i\omega_k + \bar{P}_k(y, \bar{y})), \tag{16}$$

where $P_k(y\bar{y})$ denotes a (formal) series in powers of products $y_1\bar{y}_1, \ldots, y_m\bar{y}_m$ without constant terms [4]. The normalizing transform has the form

$$x_k = y_k + h_k(y_k, \bar{y}_k), \quad \bar{x}_k = \bar{y}_k + \bar{h}_k(y_k, \bar{y}_k), \tag{17}$$

where the series $h_k(y_k, \bar{y}_k)$ are also formal.

Further, let us assume that the neutrality condition holds:

$$P_k(y_1\bar{y}_1, \ldots, y_m\bar{y}_m) = i\,H_k(y_1\bar{y}_1, \ldots, y_m\bar{y}_m), \tag{18}$$

where H_k are series with real coefficients, moreover,

$$\det|\partial H/\partial\rho|_{\rho=0} \neq 0, \quad \rho = (\rho_1, \ldots, \rho_m), \quad \rho_k = y_k\bar{y}_k,\ k = 1, \ldots, m. \tag{19}$$

Then system (16) has as integral surfaces invariant m-dimensional tori $y_k\bar{y}_k = c_k > 0$, $k = 1, ..., m$, and possesses quasi-periodic solutions. If the equivalence of systems (14) and (16) was not only formal, but analytical, then the invariant tori to system (16) would correspond to invariant tori to system (14).

Theorem 15.1 [3] implies that despite divergence of normalizing transform (17), in some sense "most" of invariant tori to the system (16) correspond to invariant tori to system (14). Namely, there exists $\varepsilon > 0$ and series $h_k(y, \bar{y}, \rho)$, $\rho = y\bar{y}$, $k = l, ..., n$, convergent in an ε-neighborhood of the origin $\mathscr{V}_\varepsilon$ for every ρ belonging to a measurable set $\mathscr{M}_\varepsilon \in \mathscr{V}_\varepsilon$, where $\lim\limits_{\varepsilon \to 0} \frac{\text{mes}\, \mathscr{M}_\varepsilon}{\text{mes}\, \mathscr{V}_\varepsilon} = 1$, such that change of variables (17) reduces system (14)–(16), where H_k are convergent for $y \in \text{mes}\, \mathscr{V}_\varepsilon$, for every $\rho \in \text{mes}\, \mathscr{M}_\varepsilon$ and have real coefficients.

To apply the Bibikov theorem, we have to reduce the system (7), (8), (10) to normal form.

The algorithm is the following.

- The system (7), (8), (10) has to be written in the form (15) at the equilibrium point $a = d = 0, b = -c = b^*, A = C = A^* = \frac{b^*(b^*-l)}{2c_0}$, $B = 0$:

$$\frac{dx_\nu}{dt} = \lambda_\nu x_\nu + \sum a^\nu_{jh} x_j x_h + \sum b^\nu_{jhk} x_j x_h x_k + ..., \tag{20}$$

  where indices ν take values $\pm 1, \pm 2, \pm 3$; $\bar{\lambda}_\nu = \lambda_{-\nu}$ and $\bar{x}_\nu = x_{-\nu}$. Here, x_ν are new variables, λ_ν and $\lambda_{-\nu}$ correspond to the complex conjugate eigenvalues, coefficients a^ν_{jh} and b^ν_{jhk} are complex-valued and symmetrized, $j, h, k, \nu = \mp 1, \mp 2, \mp 3$.
- According to the Bruno theorem [3, 4, 22], there exists a formal change of variables $x_j = y_j + \sum \alpha^j_{lm} y_l y_m + \sum \beta^j_{lmn} y_l y_m y_n + ...$, where $\alpha^j_{lm} = \alpha^j_{ml}$, $\beta^j_{lmn} = \text{id}$, $j, l, m, n = \mp 1, \mp 2, \mp 3$, reducing the system (20) to normal form:

$$\frac{dy_\nu}{dt} = \lambda_\nu y_\nu + y_\nu S(y, \bar{y}) = \lambda_\nu y_\nu + y_\nu \sum_{(\Lambda, Q)=0} g_{\nu Q} y_1^{q_1} y_{-1}^{q_{-1}} y_2^{q_2} y_{-2}^{q_{-2}} y_3^{q_3} y_{-3}^{q_{-3}}, \tag{21}$$

  where $\nu = \mp 1, \mp 2, \mp 3, q_j \in \mathbb{Z}, q_\nu \geq -1, q_j > 0,\ j \neq \nu, \sum q_n \geq 1$. For our case, condition $(\Lambda, Q) = 0$ means

$$\omega_1(q_1 - q_{-1}) + \omega_2(q_2 - q_{-2}) + \omega_3(q_3 - q_{-3}) = 0, \quad \omega_\nu = \Im \lambda_\nu.$$

  If we restrict ourselves by the non-resonant case, where ω_ν are rationally independent, we obtain $q_\nu = q_{-\nu}$, $\nu = 1, 2, 3$. Thus, the series $S(y, \bar{y})$ contains infinitely many terms.
- To prove that almost all trajectories in a ε-neighborhood of the equilibrium $a = d = 0, b = -c = b^*, A = C = A^* = \frac{b^*(b^*-l)}{2c_0}$, $B = 0$ are quasi-periodic, we have to show that $S(\mathbf{y}) = iH(y_1 y_{-1}, y_2 y_{-2}, y_3 y_{-3})$, where H is a real-valued vector-function. Thus, we have to check that the coefficients $g_{\nu \mathbf{Q}}$ are purely imaginary.

3.4.2 Method of Computing $g_{\nu Q}$, Truncated Case

In [22], Chap. VIII, Sect. 4, the author analyzes resonances and normal forms of analytic autonomous (not necessarily conservative) sixth-order systems with three pairs of distinct pure imaginary eigenvalues of the matrix of the linear part, as in our case. Provided the normal form (21) is truncated up to terms of power not higher than three, there exist explicit formulae for calculation of coefficients of normalizing transformation and normal forms.

Namely, the truncated normal form (21) is

$$\frac{dy_\nu}{dt} = \lambda_\nu y_\nu + y_\nu S_3(y, \bar{y}) = \lambda_\nu y_\nu + y_\nu (g_1^\nu y_1 y_{-1} + g_2^\nu y_2 y_{-2} + g_3^\nu y_3 y_{-3}). \quad (22)$$

Method of computing g_i^j is the following [22].

1. We denote the vector of solutions $(A, B, C, a, c, d)^T$ as Z and rewrite system (7), (8), (10) as
$$Z' = GZ + F(Z), \quad (23)$$
where G is a matrix of linearization at the equlibrium, and $F(Z)$ is the nonlinear part.
2. We reduce G to its diagonalized form D by means of the non-degenerate matrix C, such that $G = C^{-1}DC$. The change of variables $Y = CZ$ reduces (23) to $Y' = DY + CF(C^{-1}Y)$.
3. We expand the matrix $CF(C^{-1}Y)$ of nonlinear part to the Taylor series in the new variables at the equilibrium and find the coefficients of second order, a_{jh}^ν, and third order, b_{jhk}^ν.
4. We find coefficients α_{lm}^ν, g_h^ν by the following formulae:

$$\alpha_{lm}^\nu = \frac{a_{lm}}{\lambda_l + \lambda_m - \lambda_\nu}; \quad g_{|\nu|}^\nu = 3b_{\nu\nu-\nu}^\nu + 2\sum_j (2a_{\nu j}^\nu \alpha_{\nu-\nu}^j + a_{-\nu j}^\nu \alpha_{\nu\nu}^j); \quad (24)$$

$$g_\nu^h = 6b_{\nu h-h}^\nu + 4\sum_j (a_{\nu j}^\nu \alpha_{h-h}^j + a_{hj}^\nu \alpha_{-h\nu}^j + a_{-hj}^\nu \alpha_{\nu h}^j),$$

where $h \neq |\nu|$; $\nu = \pm 1, \pm 2, \pm 3$.

To check condition (18), we have to prove that the real parts of g_ν^h are zero. We performed computations according to formula (24), taking values of $\gamma \in (1, 2]$ and $b^*/l \in \Sigma$ with the step 0.001. They confirm that (18) holds with a good reliability. The real parts of g_ν^h do not vanish in a very small neighborhood of boundary points $\frac{1-\sqrt{2}}{2} \approx -0.2071$ and $\frac{1+\sqrt{2}}{2} \approx 1.2071$ and points 0 and 1. Thus, we check numerically that except of small neighborhood of these points the neutrality condition (18) holds for the truncated up to third terms normal form (22).

Further, as follows from the Bibikov–Pliss criterium [2], for the case of two pairs of pure imaginary eigenvalues, the terms of the third order can help to prove the stability of equilibrium for non-resonant frequencies. Namely, if

$$\dot{y}_1 = iy_1 + i(\frac{g_1}{i}y_1|y_1|^2 + \frac{h_1}{i}y_1|y_2|^2) + \ldots, \quad \dot{y}_2 = i\lambda y_2 + i(\frac{g_2}{i}y_2|y_1|^2 + \frac{h_2}{i}y_2|y_2|^2) + \ldots,$$

λ is irrational, g_i,h_i, $i = 1, 2$, are real and $h_1h_2 - g_1g_2 \neq 0$, then "almost all" of an ε-neighborhood of equilibrium is filled by almost-periodic motions; i.e., the equilibrium is "practically" stable.

Nevertheless, the first integrals (11) and (12) reduce system (7), (8) to four equations such that the matrix if linearization has two pairs of purely imaginary eigenvalues. The only problem is that the eigenvalues and coefficients of the respective normal form hardly can be found explicitly. Nevertheless, in general situation $h_1h_2 - g_1g_2$ does not vanish, and therefore, one can conclude that the equilibrium is basically stable.

Remark 2. If (18) does not hold in some point $b^*/l \in \Sigma$, this suggests that the system has a resonance in this point, which can be stable or unstable. As mentioned earlier, in our case such points are close to the boundaries of Σ with unstable domain. For case of two pairs of purely imaginary eigenvalues, there exist many known results about resonances (e.g. [12]). In particular, it is known that the resonances 1:2 are basically unstable [14].

4 Localized Vortex

In [20], we constructed a class of steady vortices of the form $\mathbf{U} = \nabla_\perp \Phi = (\Phi_{x_2}, -\Phi_{x_1})$, $\Pi = -\frac{1}{c_0}\left[l\Phi + \int(\Phi_{x_2}\Phi_{x_1x_2} - \Phi_{x_1}\Phi_{x_2x_2})\,dx_1 + \int(-\Phi_{x_2}\Phi_{x_1x_1} + \Phi_{x_1}\Phi_{x_1x_2})\,dx_2\right]$, where $\Phi = \Phi(x_1^2 + x_2^2)$ is a sufficiently smooth function. This is an axisymmetric vortex, and it can be localized or not in dependence on Φ. In particular, if $\Phi = \frac{b^*}{2}(x_1^2 + x_2^2)$, we obtain the solution, corresponding to the equilibrium of ODE system considered in Sect. 3.1. If Φ has a compact support, the vortex is localized within a bounded domain. Let us consider closer the case of decaying at infinity $\Phi = -B_0\, e^{-\frac{\sigma}{2}(x_1{}^2+x_2{}^2)}$. Then

$$u_1 = B_0\,\sigma\, x_2\, e^{-\frac{\sigma}{2}(x_1{}^2+x_2{}^2)}, \qquad u_2 = -B_0\,\sigma\, x_1\, e^{-\frac{\sigma}{2}(x_1{}^2+x_2{}^2)}, \tag{25}$$

$$\Pi = -\frac{1}{2c_0}\left(B_0^2\,\sigma\, e^{-\sigma(x_1^2+x_2^2)} - 2l\,B_0\, e^{-\frac{\sigma}{2}(x_1^2+x_2^2)}\right) + R_0, \quad R_0 = \text{const}. \tag{26}$$

It is easy to see that in a neighborhood of the origin, the structure of solution is similar to the solution with a linear profile of velocity, considered in Sect. 3 where $\Phi = \frac{b^*}{2}(x_1^2 + x_2^2)$, $b^* = B_0\sigma$.

4.1 Linear Analysis, Radial Perturbations

The following extension of the Rayleigh stability criterium (e.g., [11]) to the case of compressible rotating fluid is known.

Theorem 4. *Let $V_\theta(r)$ be the tangential component of velocity, $r = \sqrt{x_1^2 + x_2^2}$. Condition*

$$\mathscr{R}(r) = \frac{1}{r^3}\frac{d}{dr} r^2 \rho(r)(V_\theta + \frac{l}{2}r)^2 \geq 0, \tag{27}$$

is a necessary and sufficient condition of stability of an axisymmertic steady flow to the system (1), (2) with respect to radial perturbations in the linearized setting.

To prove, it is enough to take into account the change of density in the momentum equation in the derivation of the criterium for the incompressible fluid in [13], (see also [21]).

However, it is known that (27) does not ensure stability with respect to asymmetric perturbation even in the incompressible case (see references in [21]).

It is easy to check that for non-localized steady-state solution from Sect. 3, the condition (27) holds. It is natural, since the assumption on the radial symmetry of density means that we are in the frame of Sect. 3.1, where the equilibrium is stable.

Nevertheless, for sufficiently fast-rotating localized vortices (with sufficiently large $|B_0|$) $\mathscr{R}(r)$ becomes negative for some values of r, see Fig. 1.

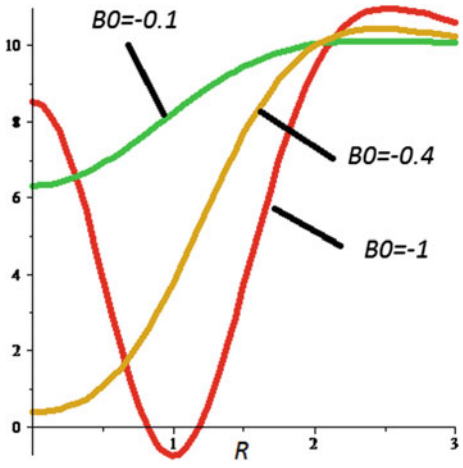


Fig. 1 Typical graphs of $\mathscr{R}(r)$ for the case of the steady state (25), (26) with $c_0 = 1, l = 1, \sigma = 1$, $R_0 = 10, \gamma = \frac{9}{7}$. For relatively fast vortices, the instability zone is close to the maximum of vorticity. As $|B_0|$ further increases, there arises an instability zone at the center

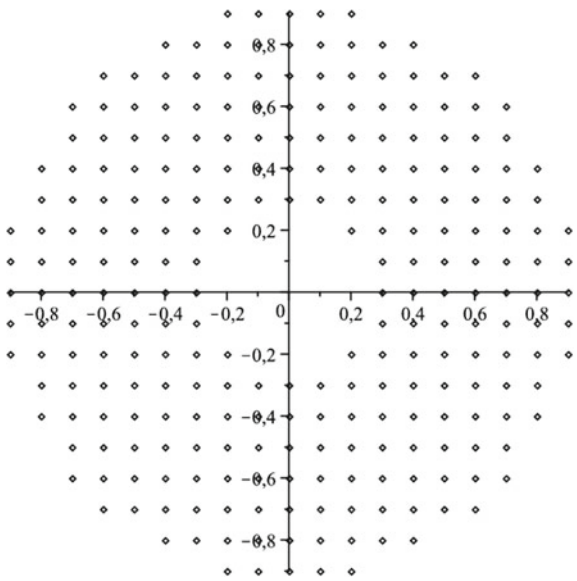


Fig. 2 Points, where solution to system (7), (8) blows up. Initial data correspond to the steady state (25), (26) with $\gamma, c_0, l, \sigma, R_0$ as in Pic.1, $B0 = -0.4$ (this case is stable with respect to radial perturbations)

4.2 "Linear Profile Velocity" Approximation

Let us assume that we know a steady-state solution $\mathbf{U}^0 = (u_1^0, u_2^0)$, Π^0 to the system (3), (4). Then in a neighborhood of every point (x_1, x_2), we can construct an approximate solution taking in the Taylor series the zero- and first-order terms in $\mathbf{U}^0$ and the zero-, first-, and second-order terms in Π^0:

$$u_1 = u_{10}(t) + a(t)x_1 + b(t)x_2, \quad u_2 = u_{20}(t) + c(t)x_1 + d(t)x_2,$$

$$\Pi = K(t) + M(t)x_1 + N(t)x_2 + \frac{1}{2}A(t)x_1^2 + B(t)x_1x_2 + \frac{1}{2}C(t)x_2^2,$$

with the initial conditions computed from $\mathbf{U}^0, \Pi^0$ at the point $(0, 0)$. Namely, $u_{10}(0) = u_1^0$, $u_{20}(0) = u_2^0$, $K(0) = \Pi^0$, $M(0) = \partial_{x_1}\Pi^0$, $N(0) = \partial_{x_2}\Pi^0$,

$$a(0) = \partial_{x_1}u_1^0, \quad b(0) = \partial_{x_2}u_1^0, \quad c(0) = \partial_{x_1}u_2^0, \quad d(0) = \partial_{x_2}u_2^0, \tag{28}$$

$$A(0) = \partial_{x_1^2}\Pi^0, \quad B(0) = \partial_{x_1x_2}\Pi^0, \quad C(0) = \partial_{x_2^2}\Pi^0. \tag{29}$$

Coefficients a, b, c, d, A, B, C are subject to nonlinear system of ODE (7), (8). Coefficients u_{10}, u_{20}, K, M, N can be found from linear equations (see the details in [18, 19]).

We solve (7), (8) numerically using the Fehlberg fourth–fifth-order Runge–Kutta method. The solution either blows up quickly or demonstrates an oscillating character.

In the latter case, we conclude that the solution is in the domain of attraction of an equilibrium. We are going to propose the following hypothesis:

Hypothesis 1 If a steady state $\mathbf{U}^0$, Π^0 is such that at every point (x_1, x_2) the solution to the system (7), (8) with the data (28), (29) falls in the domain of attraction of an equilibrium, then this steady state is nonlinearly stable in the Lyapunov sense.

Numerical experiments with the ODEs system show that the domain of stability based on our hypothesis is much more restricted than the domains of stability based on the Rayleigh criterium, see Fig. 2. For example, if we consider the steady state (25), (26) with $c_0 = 1$, $l = 1$, $\sigma = 1$, $R_0 = 10$, $\gamma = \frac{9}{7}$, then the Rayleigh criterium implies that the solution is unstable for $B_0 \lesssim -0.5$ and $B_0 \gtrsim 3$. Nevertheless, the domain of nonlinear instability based on Hypothesis 1 is $B_0 \lesssim -0.2$ and $B_0 \gtrsim 0.5$. The equilibrium corresponding to the linear approximation of velocity at the origin is unstable for $B_0 < (1 - \sqrt{2})/2 \approx -0.21$ and $B_0 > (1 + \sqrt{2})/2 \approx 1.21$.

Let us notice that from the point of view of Hypothesis 1, any steady localized vortex in irrotational coordinate frame is unstable. In this sense, the rotation has a stabilizing effect.

Computations made directly from the 2D system of compressible media for the case of fixed coordinate frame show that the stability depends both on intensity and steepness of vortex [17] (in our notation, they are B_0 and σ, respectively). Analysis based on Hypothesis 1 confirms this effect as well.

5 Conclusion

We study nonlinear stability of steady vortex in a compressible media. We begin with vortex with linear profile of velocity. In this case, the problem can be reduced to analysis of stability of equilibrium of related ODEs system. This analysis turns out to be very nontrivial. For some range of parameters, it requires to construct a Lyapunov function in the absence of standard algorithms. For other range, the study of stability requires application of theory of normal forms and invariant tori. Any other steady vortex, localized or not, can be approximated at every point of plane by a vortex with linear profile of velocity. We hypothesize that if at every point the solution to this ODEs system falls in the domain of attraction of an equilibrium, then the steady vortex is nonlinearly stable. Then, we give arguments in support of this hypothesis.

References

1. D. Bershader, Compressible vortices, in *Fluid Vortices Fluid Mechanics and Its Applications*, vol. 30, Chap. VII (Springer, Netherlands), pp. 291–316
2. Y.N. Bibikov, V.A. Pliss, On the existence of invariant tori in the neighborhood of the zero solution of a system of ordinary differential equations. Diff. Eq. **3**(11) (1967)

3. Y.N. Bibikov, *Local Theory of Nonlinear Analytic Ordinary Differential Equations*. Lecture Notes in Mathematics, vol. 702 (Springer, Berlin, 1979)
4. A.D. Bruno, *Local Methods in Nonlinear Differential Equations*, Springer Series in Soviet Mathematics (Springer, Berlin, 1989)
5. O.I. Bogoyavlensky, *Methods in the Qualitative Theory of Dynamical Systems in Astrophysics and Gas Dynamics*, Springer Series in Soviet Mathematics (Springer, Berlin, 1985)
6. W.M. Chan, T.K. Shariff, T.H. Pulliam, Instabilities of two-dimensional inviscid compressible vortices. J. Fluid Mech. **253**, 173–209 (1993)
7. J.-M. Chomaz, S. Ortiz, F. Gallaire, P. Billant, Stability of quasi two-dimensional vortices, in *Fronts, Waves and Vortices in Geophysical Flows*, ed. by J.-B. Flór. Lecture Notes in Physics, vol. 805 (Springer, Berlin, 2010)
8. F.V. Dolzhansky, *Fundamentals of Geophysical Hydrodynamics*. Encyclopaedia of Mathematical Sciences, vol. 103 (Springer, Berlin, 2013)
9. A.E. Gill, *Atmosphere-Ocean Dynamics* (Academic Press, USA, 1982)
10. T. Hiejima, Stability of compressible stream-wise vortices. Phys. Fluids **27**, 074107 (2015)
11. E.J. Hopfinger, G.J.F. van Heijst, Vortices in rotating fluids. Ann. Rev. Fluid Mech. **25**, 241–289 (1993)
12. L.G. Khazin, E.E. Shnol, *Stability of Critical Equilibrium States* (Manchester University Press, Manchester, 1991)
13. R.C. Kloosterziel, G.J.F. van Heijst, An experimental study of unstable barotropic vortices in a rotating fluid. J. Fluid Mech. **223**, 1–24 (1991)
14. A.L. Kunitsyn, On the stability in the critical case of pure imaginary roots under conditions of internal resonance. Diff. Eq. **7**, 1704–1706 (1971, in Russian)
15. V.V. Meleshko, H. Aref, A bibliography of vortex dynamics, 1858–1956. Adv. Appl. Mech. **41**, 197–292 (2007)
16. I. Menshov, Y. Nakamura, Instability of isolated compressible entropy-stratified vortices. Phys. Fluids **17**, 034102 (2005)
17. I. Menshov, Tearing instability of isolated compressible vortices. Int. J. Aeroacoust. **13**, 113–140 (2014)
18. O.S. Rozanova, J.-L. Yu, C.-K. Hu, Typhoon eye trajectory based on a mathematical model: comparing with observational data. Nonlinear Anal. Real World Appl. **11**, 1847–1861 (2010)
19. O.S. Rozanova, J.-L. Yu, C.-K. Hu, On the position of vortex in two-dimensional model of atmosphere. Nonlinear Anal. Real World Appl. **13**, 1941–1954 (2012)
20. O.S. Rozanova, Frozen and almost frozen structures in the compressible rotating fluid. Bull. Braz. Math. Soc. New Ser. **47**, 715–726 (2015)
21. D.A. Shalybkov, Hydrodynamic and hydromagnetic stability of the Couette flow. PhysicsUspekhi **52**(9), 915–935 (2009)
22. V.M. Starzhinskii, *Applied Methods in the Theory of Nonlinear Oscillations* (Mir, Moscow, 1980)

Coupled Scheme for Hamilton–Jacobi Equations

Smita Sahu

Abstract In this paper, we will present some coupled numerical schemes for Hamilton–Jacobi equation by using the scheme proposed in Falcone and Sahu (Coupled scheme for linear and Hamilton-Jacobi-Bellman equations, 2016 [11]). The approach is general and in principle can be applied to couple many different schemes, for example one can couple an accurate method well adepted where the solution is smooth with another method designed to treat discontinuities and/or jumps in the gradients. Clearly, one has to decide where to apply the first or the second method, and this is done by means of a switching parameter which must be computed in every cell at every time step. In this paper, we investigate, in particular, the coupling between an anti-dissipative scheme by Bokanowski and Zidani (J Sci Comput 30(1):1–33 2007, [4]) which has been proposed in order to deal with discontinuous solutions and a semi-Lagrangian scheme by Falcone and Ferretti (Semi-Lagrangian approximation schemes for linear and Hamilton-Jacobi equations. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, 2014 [10]) which is more adept to deal with Lipschitz continuous solutions and is more accurate for regular solutions provided a high-order local interpolation operator is used for the space reconstruction. We will show that how the coupling can be done for two schemes which typically use two different grid reconstructions.

Keywords Hamilton-Jacobi-Bellman equations · Semi-Lagrangian schemes · Anti-dessipative schemes · Viscosity solutions

1 Introduction

Our aim is to propose a new method to build schemes for first-order time-dependent Hamilton–Jacobi (HJ) equations coupling two schemes for viscosity solution which have different properties. We will consider the following model problem

S. Sahu (✉)
Department of Mathematical Sciences, Durham university, Durham, UK
e-mail: smita.sahu@durham.ac.uk

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_42

$$\partial_t v + H(x, \nabla v) = 0, \quad (t, x) \in [0, T] \times \mathbb{R} \tag{1}$$

$$v(0, x) = v_0(x), \quad x \in \mathbb{R}. \tag{2}$$

equation, where the Hamiltonian H is convex in the second argument. A typical example comes from optimal control theory, where $H(x, \nabla v) = \max_{a \in A}\{f(x, a)v_x(t, x)\}$ and a represents the control, and it is well known that in this framework, the solution v of the equation (1) corresponds to the value function of the problem [1, 2]. Typically, the solution is Lipschitz continuous if the data are Lipschitz continuous but also discontinuous solutions can be considered and they actually appear in several applications to control problems with state constraints, games and image processing. This is our main motivation to deal here with discontinuous initial conditions, and in general, the coupled scheme will be designed in order to be able to track discontinuous solutions. However, since the typical situation is to have a piecewise regular solution which only has discontinuities or jumps of the derivatives at isolated points, it is natural to try to diversify the method in the subdomains where the solution is regular and in the cells where the solution exhibits this kind of singularities. To this end, we will couple two schemes which have been already proposed in the literature and for which we know a number of properties which will turn to be useful for the construction of the coupled scheme. Let us also mention that hybrid schemes for hyperbolic conservation laws have been proposed in the literature to capture shocks for hyperbolic conservation laws and contact discontinuities for the compressible Euler system (see Chap. 22 in the book by Laney [13] for more information and references). The coupled scheme proposed here follows the same ideas although our goal is to solve HJ equations and the schemes chosen for the coupling are different.

It is well known that, in the one-dimensional case, there is a strong link between HJ equations and hyperbolic conservation laws. Namely, the viscosity solution of the evolutive HJ equation is the primitive of the entropy solution of the corresponding hyperbolic conservation law with the corresponding Hamiltonian. Most of the numerical ideas to solve hyperbolic conservation law can be extended to HJ equations. In the last decades, many numerical schemes have been proposed for HJ equations using different techniques, for example finite differences, Markov chain, semi-Lagrangian (SL) [10], high-order filtered scheme [7, 12, 15, 16]; these schemes have been shown to be stable and convergent under mild regularity assumptions on the solution and to be the first order accurate for the approximation of Lipschitz continuous solutions. However, it can be interesting to deal with discontinuous viscosity solutions so these schemes have to be adepted in order to obtain reasonable approximations which do not diffuse too much around the discontinuities of Dv and/or v and do not oscillate.

For discontinuous solutions, an anti-dissipative (AD) scheme has been proposed [4] and a convergence result has been proved in one dimension [5, 6]. That scheme has been initially proposed for hyperbolic conservation laws [9, 14] and then extended to Hamilton–Jacobi equations in one dimension. Another class of schemes which have been shown to be rather effective is that of SL scheme (see Falcone and Ferretti book [10] for a comprehensive presentation of this approach). SL schemes give good results and are naturally multidimensional, and they can be very accurate

in the regions of regularity for the solution provided a high-order local reconstruction in space is used. Despite these interesting features, SL schemes are not efficient for discontinuous initial data since they use a local interpolation operator for the computation at the foot of the characteristics.

In this paper, we present a new coupling of semi-Lagrangian scheme and Ultra-Bee scheme (a particular anti-dissipative scheme) for (1). We intend to take the advantage of the properties of the two methods introducing an indicator parameter.

Organization of the paper. In Sect. 2, we will recall the basic results about the semi-Lagrangian (SL) method [10] and the Ultra-Bee (UB) scheme [4] which we will use in the coupling. In Sect. 3, we present the general form of the coupled scheme and we will describe how it will be applied to solve the linear advection equation and how it has been extended to Hamilton–Jacobi equation. Finally, Sect. 4 will be devoted to the numerical tests in one dimension.

2 Semi-lagrangian Schemes [10]

A semi-Lagrangian (SL) method is based on two basic steps: the reconstruction of the solution on a fixed grid and numerical integration along the lines of the same characteristics The idea of using the aspect numerical method of characteristics was proposed for the first time by Courant, Isaacson and Rees in the [8]. In dimension one, the CIR scheme precisely gives the first-order upwind scheme when applied to the advection equation imposing the CFL condition $c_{max}\Delta t/\Delta x \leq 1$ where c_{max} is the upper bound for the modulus of the velocity. However, the main advantage of these methods is that they are still stable for large time steps so they do not need the typical CFL condition required by finite difference methods. This helps particularly to run simulations to investigate the long-time behaviour of the solution. In the framework of HJ, SL schemes have been developed initially for the solution of Bellman's equations associated with optimal control problems. This schemes can be interpreted as a discretization of the dynamic programming principle.
The typical *assumptions on H* are as follows:

1. $H(\cdot,\cdot,\cdot)$ is uniformly continuous in all the variables.
2. $H(x,v,\cdot)$ is convex and coercive.
3. $H(x,\cdot,Dv)$ is monotone.

Under these assumptions, we have the representation Hopf–Lax formula for the solution of equation (1)

$$v(x,t+\Delta t)=\min_{a\in\mathbb{R}}\{v(x-a\Delta t,t)+\Delta t H^*(a)\},$$

where

$$H^*(a)=\sup\{a\cdot p-H(p)\}$$

is the Legendre transform of Hamiltonian H. Note that the formula is the extension of the classical representation formula for the linear advection equation. For simplicity, we set up everything in dimension one. Let $I_1[u]$ denote the P_1-interpolation of a function u in dimension one on the mesh $G = \{x_j\}$, i.e.

$$I_1[u](x) = \frac{x_{j+1} - x}{\Delta x} u_j + \frac{x - x_j}{\Delta x} u_{j+1} \quad \text{for } x \in [x_j, x_{j+1}]. \tag{3}$$

Hence, the SL scheme for (1) is

$$u_j^{n+1} = \min_{a \in R} \{I_1[u^n](x_j - a\Delta t) + \Delta t H^*(a)\}. \tag{4}$$

SL scheme is monotone stable and works for large the Courant number. Convergence and error estimate have been proved (see [10] for precise results).

2.1 Ultra-Bee Scheme for HJ Equations [4]

In this section, we recall "Ultra-Bee" (UB) scheme of Roe [14]. The Ultra-Bee scheme is nonmonotone, but it has the interesting property to transport exactly a particular space of step functions in the case of linear advection when the speed is constant. In [4], Bokanowski and Zidani have presented a modified Ultra-Bee scheme for the model problem (5) and anti-dissipative properties of the scheme have been shown in [3, 5, 6]. A first-order convergence result has been proved for the modified Ultra-Bee scheme, in L^1-norm, towards the viscosity solution for the model problem (5) (for more details and proof, we refer reader to see [3, 5]). Here, we will recall the scheme for the model problem:

$$\partial_t v + \max_{a \in \mathscr{A}} (f(x, a) v_x(t, x)) = 0, \quad (t, x) \in [0, T] \times \mathbb{R}, \tag{5}$$

$$v(0, x) = v_0(x), \quad x \in \mathbb{R}. \tag{6}$$

In optimal control theory, the solution of above equation corresponds to the value function of an optimization problem [2]. It is usual that this function, as well as the final cost v_0, is discontinuous (for instance for target or rendezvous problems). Let Δt be a constant time step and $t_n = n\Delta t$ for $n \geq 0$. Given two velocity functions $f^g : \mathbb{R} \to \mathbb{R}$, $g = m,\ M$, we set the following notation for the corresponding CFL numbers at a node x_j:

$$\nu_j^m := \frac{\Delta t}{\Delta x} f_m(x_j) \text{ and } \nu_j^M := \frac{\Delta t}{\Delta x} f_M(x_j), \ \ j \in \mathbb{Z}. \tag{7}$$

Then, we can define the vectors, $v^m = \{v_j^m,\ j \in \mathbb{Z}\}$, $v^M = \{v_j^M,\ j \in \mathbb{Z}\}$. Let us define the exact average values of the approximate solution at time t_n:

$$\overline{u}_j^n = \frac{1}{\Delta x} \int_{j-1/2}^{j+1/2} u(t_n, x)dx\ ,\ j \in \mathbb{Z},\ n \in \mathbb{N}. \tag{8}$$

Denoting by $\|f\|_\infty$ the L^∞ norm of a bounded function defined on $\mathbb{R}$, we define the CFL condition

$$\max\left(\|f_m\|_\infty, \|f_M\|_\infty\right) \frac{\Delta t}{\Delta x} \leq 1. \tag{9}$$

Here, we recall the steps of the algorithm for the UB scheme.

Algorithm for the UB scheme

Initialization. Compute the initial averages $\{\overline{u}_j^0\}_{j\in\mathbb{Z}}$ as above.
For $n \geq 0$.
Main cycle:
Step 1. Compute $u^{n+1} = \{u_j^{n+1}\}_{j\in\mathbb{Z}}$ by:
Step 2. For every $j \in \mathbb{Z}$, we define the "fluxes" $u_{j\pm1/2}^n(v_j)$ for $v_j \in \{v_j^m, v_j^M\}$ as follows:
if $v_j \geq 0$, we set

$$u_{j+1/2}^n(v) := \begin{cases} \min\left(\max\left(\overline{u}_{j+1}^n, b_j^+(v_j)\right), B_j^+\right) & \text{if } v_j > 0 \\ \overline{u}_{j+1}^n & \text{if } v_j = 0 \text{ and } \overline{u}_j^n \neq \overline{u}_{j-1}^n \\ \overline{u}_j^n & \text{if } v_j = 0 \text{ and } \overline{u}_j^n = \overline{u}_{j-1}^n, \end{cases} \tag{10}$$

where

$$\begin{cases} b_j^+(v) := \max\left(\overline{u}_j^n, \overline{u}_{j-1}^n\right) + \frac{1}{v_j}\left(\overline{u}_j^n - \max\left(\overline{u}_j^n, \overline{u}_{j-1}^n\right)\right), \\ B_j^+(v) := \min\left(\overline{u}_j^n, \overline{u}_{j-1}^n\right) + \frac{1}{v_j}\left(\overline{u}_j^n - \min\left(\overline{u}_j^n, \overline{u}_{j-1}^n\right)\right), \end{cases} \tag{11}$$

if $v_j \leq 0$, we set

$$u_{j-1/2}^n(v) := \begin{cases} \min\left(\max\left(\overline{u}_{j-1}^n, b_j^-(v_j)\right), B_j^-\right) & \text{if } v_j < 0 \\ \overline{u}_{j-1}^n & \text{if } v_j = 0 \text{ and } \overline{u}_j^n \neq \overline{u}_{j+1}^n \\ \overline{u}_j^n & \text{if } v_j = 0 \text{ and } \overline{u}_j^n = \overline{u}_{j+1}^n, \end{cases} \tag{12}$$

where

$$\begin{cases} b_j^-(v) := \max\left(\overline{u}_j^n, \overline{u}_{j+1}^n\right) + \frac{1}{v_j}\left(\overline{u}_j^n - \max\left(\overline{u}_j^n, \overline{u}_{j+1}^n\right)\right), \\ B_j^-(v) := \min\left(\overline{u}_j^n, \overline{u}_{j+1}^n\right) + \frac{1}{v_j}\left(\overline{u}_j^n - \min\left(\overline{u}_j^n, \overline{u}_{j+1}^n\right)\right), \end{cases} \tag{13}$$

Step 3. For $\nu_j \in \left\{\nu_j^m, \nu_j^M\right\}$, we define

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \nu_j \left(u_{j+1/2}^n(\nu) - u_{j-1/2}^n(\nu)\right). \tag{14}$$

Step 4. Finally, we set $\overline{u}_j^{n+1} := \min\left(\overline{u}_j^{n+1}(\nu_j^m), \overline{u}_j^{n+1}(\nu_j^M)\right)$, $j \in \mathbb{Z}$.

For simplicity and considering all the cases, we will use the following short notation for the Ultra-Bee scheme

$$\overline{u}_j^{n+1} = S_j^{UB}(\overline{u}^n) := \left(\min\left(\overline{u}_j^{n+1}(\nu^m), \overline{u}_j^{n+1}(\nu^M)\right)\right)_{j\in\mathbb{Z}}. \tag{15}$$

For the advection equation in [9], it has been proved that under the CFL condition $0 \le \nu_j \le 1$, for all j, UB scheme is consistent, L^∞ stable and TVD. Let us also mention the form of flux which is used in [9], i.e.

$$u_{j+1/2}^n := \overline{u}_j^n + \frac{1-\nu_j}{\phi_j}(\overline{u}_{j+1}^n - \overline{u}_j^n), \tag{16}$$

where ϕ_j is defined as

$$\phi_j = \begin{cases} \max\left(0, \min\left(\frac{2r_j}{\nu_j}, \frac{2}{1-\nu_j}\right)\right), & \text{if } \overline{u}_{j+1}^n = \overline{u}_j^n \text{ and } \nu_j \neq 1 \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$

where $r_j = \frac{\overline{u}_j^n - \overline{u}_{j-1}^n}{\overline{u}_{j+1}^n - \overline{u}_j^n}$ and by replacing $j = j-1$ we can compute $u_{j-1/2}^n$.

3 Construction of the Coupled Scheme (CS)

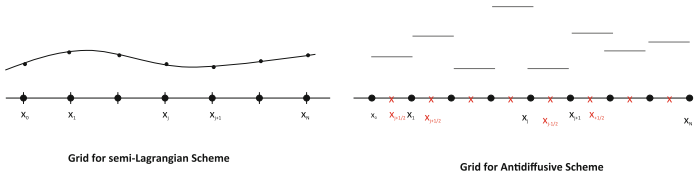
As we said, Ultra-Bee schemes are based on previous results for conservation laws and they typically require a projection onto a discontinuous reconstruction at every step. This choice seems to be clever for the regions where the solution is nonregular but rather unfortunate where the solution is regular. Then, a natural idea is to couple the features of two schemes: a scheme (SL) well adepted for regular (at least Lipschitz continuous) solutions with an Ultra-Bee scheme (UB) which provides a better solution profile at the jumps. Thus, we expect to get advantages coupling the two schemes; to this end, we should be able to detect the regularity regions and the singular regions.

SL scheme uses a local interpolation operator to recover the value of the numerical solution at the foot of characteristics which are not grid points themselves. In their standard version, SL schemes do not use cell averages. On the contrary, AD schemes

are based on cell average values. For the coupling, we need values on two different grids G^{SL} and G^{AD} with space step Δx which are defined below:

$$G^{SL} = \{x_j = j\Delta x : j = \mathbb{Z}\},\ G^{AD} = \{\overline{x}_j : \overline{x}_j = x_j + \frac{\Delta x}{2},\ j \in \mathbb{Z}\}$$

For simplicity, we will often use the shorthand notation for the nodes of the two grids which are shifted by $\Delta x/2$, denoting as $\bullet$-nodesthe nodes of G^{SL} and $\times$-nodesthe nodes of G^{AD}. In the sequel, u_j^n denotes an approximation of $u(x_j, t_n)$, and $\overline{u}_j^n$ denotes an approximation of $\overline{u}(\overline{x}_j, t_n)$, where $t_n = n\Delta t$, $\Delta t > 0$. Moreover, we will drop the time index n and denote for simplicity $u_j = u_j^n$ whenever the time dependence is not necessary. At every step, we divide our domain into two regions, one where our approximate solution is "regular" and the other where we detect discontinuities.



To construct the coupled scheme, we need to introduce some indicators. Let us start defining the approximate derivatives as left and right derivatives for every node $x_j \in G^{SL}$

$$D^- u_j := \frac{u_j - u_{j-1}}{\Delta x} \text{ and } D^+ u_j := \frac{u_{j+1} - u_j}{\Delta x}. \tag{18}$$

Definition 1. Let δ be a positive threshold parameter. A cell $C_j = [x_j, x_{j+1})$ is said to be a *regular cell* if we have $|D^- u_j| < \delta$, $D^- u_j D^- u_{j-1} > 0$ and $D^- u_j$ $D^- u_{j+1} > 0$.

This means that a derivative below a given threshold as well as a constant sign in the derivatives just before and after the node x_j is considered to be a regularity indicator. For the choice of the threshold δ, we can use a previous knowledge of the bounds for the exact solution. For example, in the case of transport equation with the constant velocity, we know the solution which is $u(x,t) = u_o(x - ct)$, so we can set our threshold with the help of the initial condition $\delta = ||D^- u_j^0||_\infty - \varepsilon, \varepsilon > 0$.

Definition 2. A cell C_j is said to be *a singular cell* if it is not a regular cell. We denote the set of singular cells by $\mathscr{C}_s$.

Definition 3. The *singular region* Ω_{sin} is defined by the union of all the singular cells.

Definition 4. The set $\Omega_{reg} = \mathbb{R} \setminus \Omega_{sin}$ is called the *regular region*.

Definition 5. The set $\Omega_{reg} = \mathbb{R} \setminus \Omega_{sin}$ is called the *regular region*.

We need to distinguish between the nodes $x_j \in G^{SL}$ belonging to one of the above regions in order to apply the more adept scheme there. To this end, we define the *regularity indicator*, which will govern the switching between the two schemes: $\sigma_j \equiv 0$ for $x_j \in \Omega_{sin}$ and $\sigma_j \equiv 1$ for $x_j \in \Omega_{reg}$. It is important to note that the quantities used in the definition of the singular cells rely on the pointwise values u_j and not on the averaged values $\overline{u}_j$ so we will always need the pointwise values everywhere on the G^{SL} grid (see Step 6 of the coupled algorithm below). Another important point is that we are not applying the two schemes everywhere at every time step but we apply only one scheme in every cell, simply switching from one to the other whenever the regularity indicator σ_j^n switches from 0 to 1 or vice versa. Note that the schemes are working on different grids and that during the evolution the singularity can move from cell to cell, so when the indicator says that we have to switch from *SL*to *UB*(or vice versa) we will need to define the necessary values on the $\times$-nodes(or $\bullet$-nodes) which are neighbours of the node x_j. The coupled scheme will construct them by means of two local projection operators defined below.

Definition 6 ***(Local Projection Operator for SL).*** We define the local projection operator $P^{SL} : \mathbb{R}^2 \to \mathbb{R}$ by a map which defines the new value u_j at x_j starting from the values $(\overline{u}_{j-1/2}, \overline{u}_{j+1/2})$,

$$P^{SL}(\overline{u}_{j-1/2}, \overline{u}_{j+1/2}) := \frac{\overline{u}_{j-1/2} + \overline{u}_{j+1/2}}{2} = u_j \tag{19}$$

The P^{SL} operator constructs the point value at x_j as the average of the averaged values at $\overline{x}_{j-1}$ and $\overline{x}_j$.

Definition 7 ***(Local Projection Operator for AD).*** We define the local projection operator $P^{AD} : \mathbb{R}^2 \to \mathbb{R}$ by a map which defines the new value $\overline{u}_j$ at $x_{j+1/2}$ starting from the values (u_j, u_{j+1}),

$$P^{AD}(u_j, u_{j+1}) := \frac{u_j + u_{j+1}}{2} = \overline{u}_{j+1/2}. \tag{20}$$

The P^{AD} operator constructs the averaged value at $\overline{x}_j$ as the average of the point values at x_j and x_{j+1}.

As we said, the projection operators will be used locally whenever in a cell we switch from one scheme to the other and we need new values which were not available before. The P^{SL} operator will also be used at Step 6 to allow the update of the regularity indicator which is computed on the $\bullet$-nodes. In the sequel, we will consider an initial condition w^0 with compact support Q and define the subset $J := [j_{min}, j_{max}] \subset \mathbb{Z}$ containing the node indices of an interval containing Q.

Algorithm for the Coupled Scheme SL+UB

Step 1 (Initialization).

We compute the initial data $w_j^0 = u_j^0$ on every x_j, $j \in J$.

We compute $D^- w_{j-1}^0$, $D^- w_j^0$ and $D^- w_{j+1}^0$ and check the condition

$$|D^- w_j^0| < \delta \text{ and } D^- w_{j-1}^0 D^- w_j^0 > 0, \quad D^- w_j^0 D^- w_{j+1}^0 > 0 \,. \tag{21}$$

if this condition is true then we set $\sigma_j^0 = 1$ else we set $\sigma_j^0 = 0$.

For $n > 0$.

Main cycle on $j \in J$

Step 2. We compute $D^- w_{j-1}^n$, $D^- w_j^n$ and $D^- w_{j+1}^n$ and check the condition

$$|D^- w_j^n| < \delta \text{ and } D^- w_{j-1}^n D^- w_j^n > 0, \quad D^- w_j^n D^- w_{j+1}^n > 0 \,. \tag{22}$$

go to Step 3.

Step 3. If condition (22) is true then go to Step 4 else we go to Step 5.

Step 4. We want to apply the SL-scheme at x_j, so we set $\sigma_j^n = 1$ at the node x_j.

If $\sigma_j^n = \sigma_j^{n-1}$ we directly compute the new value according to the SL-scheme

$$w_j^{n+1} = \sigma_j^n S_j^{SL}[w^n] + (1 - \sigma_j^n) S_j^{AD}[w^n] = S_j^{SL}[w^n]. \tag{23}$$

If $\sigma_j^n \neq \sigma_j^{n-1}$, we have to switch from the AD-scheme to the SL-scheme and we need the projection P^{SL}. Then, we set for $k = j, j+1$

$$w_k^n = P^{SL}(\overline{u}_{k-1/2}^n, \overline{u}_{k+1/2}^n) := \frac{\overline{u}_{k-1/2}^n + \overline{u}_{k+1/2}^n}{2} = u_k^n,$$

and we compute

$$w_j^{n+1} = \sigma_j^n S_j^{SL}[w^n] + (1 - \sigma_j^n) S_j^{AD}[w^n] = S_j^{SL}[w^n]. \tag{24}$$

Step 5. The condition (22) is not satisfied, then we set $\sigma_j^n = 0$.

If $\sigma_j^n = \sigma_j^{n-1}$ we directly compute the new value according to the AD-scheme

$$\overline{w}_j^{n+1} = \sigma_j^n S_j^{SL}[w^n] + (1 - \sigma_j^n) S_j^{AD}[w^n] = S_j^{AD}[w^n]. \tag{25}$$

If $\sigma_j^n \neq \sigma_j^{n-1}$, we have to switch from the SL-scheme to the AD-scheme and we need the projection P^{AD}. Then, we set for $k = j-1, j, j+1$

$$\overline{w}_k^n = P^{AD}(u_k^n, u_{k+1}^n) := \frac{u_k^n + u_{k+1}^n}{2} = \overline{u}_{k+1/2}^n$$

and we compute

$$\overline{w}_j^{n+1} = \sigma_j^n S_j^{SL}[w^n] + (1 - \sigma_j^n) S_j^{AD}[\overline{w}^n] = S_j^{AD}[\overline{w}^n]. \tag{26}$$

End of the j *cycle*.

Step 6 (Filling the holes procedure).
At the •-nodeswhere $\sigma_j^n = 0$ we need to project by P^{SL} defined in (19) using the intermediate values at $\overline{w}_{j-1}^{n+1}$ and $\overline{w}_j^{n+1}$, i.e.

$$w_j^{n+1} = P^{SL}(\overline{w}_{j-1}^{n+1}, \overline{w}_j^{n+1}), \text{ for } \sigma_j = 0.$$

(the values w_j^{n+1} for the •-nodeswhere $\sigma_j = 1$ are already available by Step 4). This will finally produce the new approximate solution w_j^{n+1}.
Step 7.
Set $n = n + 1$, $j = j_{min}$ and go back to the main cycle. □

Note that at the •-nodeswhere $\sigma_j^n = 1$, we always have a value which is computed by the *SL*scheme and that the switching indicator is chosen on the basis of the values at the •-nodes. Several properties of the coupled scheme have been proved in [11].

4 Numerical Tests

In this section, we present two numerical tests (advection equation and HJ equation) in dimension one to check the efficiency of the method and to verify that the use of the switching indicator σ_j^n actually allows to improve the accuracy with respect to the two original methods used in the coupled scheme in many interesting cases. We will always compare the proposed coupled scheme with the two schemes used as building blocks. To this end, we will consider several initial conditions with various regularity properties and we follow their evolutions in time over an interval Ω. It is important to note that in coupled scheme, we have additional computational cost to calculate the indicator function. However, we are not projecting whole grid every time. When our indicator function detects that we are switching from one scheme to another, then we use projection operator only in that cell not everywhere so this computational cost is minimal.

Example 1. **Advection equation with constant velocity.**

$$v_t + cv_x = 0, \quad (t, x) \in [0, T] \times \Omega, \tag{27}$$

$$v(0, x) = v_0(x) = \begin{cases} 1 - |x| & \text{if } |x| \le 1 \\ 0 & \text{otherwise,} \end{cases}, \tag{28}$$

where $c \equiv 1$ is the velocity and $v_0(x)$ is the initial condition with bounded support. We define $\Omega := [-2, 2]$, $T = 1$ and the Courant number $\nu = \Delta t / \Delta x$ equal to 0.321. In Fig. 1, it is clear that UB scheme has the typical behaviour in regularity reason but keeps the support correctly. For this example, SL scheme is working more accurately. We expect from the couple scheme to switch to the SL scheme. In Fig. 2, $\sigma = 1$ which means coupled scheme switches to SL scheme; hence, the SL scheme and coupled

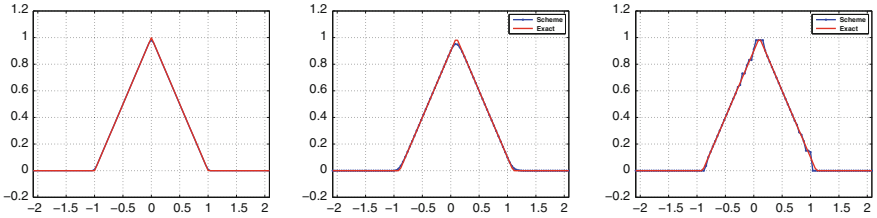


Fig. 1 (Example 1), on the left is the plot of initial data (28) and on the middle $SL\text{-}P^1$ or $SL\text{-}P^1$ $+UB$coupled scheme on the right UB scheme at $t = 20\,\Delta t$ where $\Delta t = 0.010$

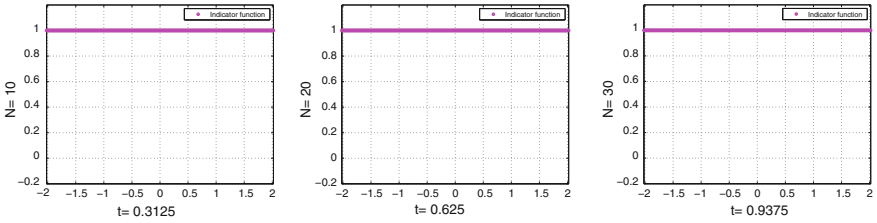


Fig. 2 (Example 1), the plot of the indicator function σ for (28) initial data at $t = 10\,\Delta t,\ 20\,\Delta t,\ 30\,\Delta t$ where $\Delta t = 0.010$

Table 1 (Example 1), errors for the Ultra-Bee scheme with initial condition (28)

Δt	Δx	L^1 Error	L^2 Error	L^∞ Error
0.0400	0.1250	0.031855	0.028728	0.071408
0.0200	0.0625	0.017081	0.015496	0.040000
0.0100	0.0312	0.008840	0.008120	0.025000
0.0050	0.0156	0.004631	0.004691	0.020000
0.0025	0.0078	0.003663	0.003197	0.011250

scheme have exactly same error table i.e. Table 2 and errors of Ultra-Bee are given in Table 1.

Example 2. **In the example below, we solve the HJ equation**

$$v_t + |f(x)v_x|, \quad (t, x) \in [0, T] \times \Omega, \tag{29}$$

$$v(0, x) = v_0(x) = \begin{cases} 1 & \text{if } |x| \le 1 \\ 0 & \text{otherwise} \end{cases}, \tag{30}$$

where $f(x) \equiv 1$. We solve the above HJ equation for the initial data (30). We fix the CFL number to 0.321 and the domain $\Omega = [-2, 2]$ and $T = 1$. All error calculations are global as before. However, we also added one column of local L^∞-errors in

Table 2 (Example 1), errors for the $SL\text{-}P^1$ or $SL\text{-}P^1 + UB$coupled scheme with initial condition (28)

Δt	Δx	L^1 Error	L^2 Error	L^∞ Error
0.0400	0.1250	0.026628	0.018858	0.027079
0.0200	0.0625	0.013588	0.009813	0.018690
0.0100	0.0312	0.007032	0.005338	0.015603
0.0050	0.0156	0.003597	0.002924	0.012181
0.0025	0.0078	0.001827	0.001616	0.009171

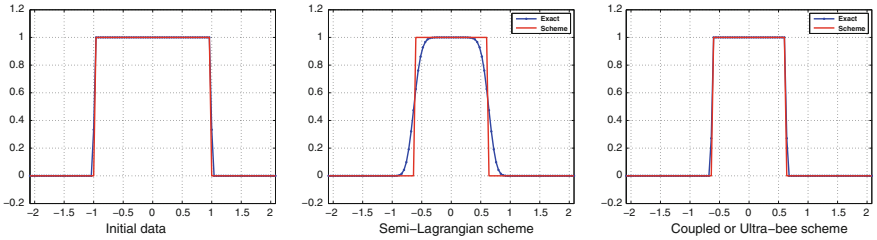


Fig. 3 (Example 2), on the left is the plot of initial data (30) and on the middle $SL\text{-}P^1$ and coupled scheme or $SL\text{-}P^1 + UB$or UB on the right $SL\text{-}P^1 + UB$ at $t = 20\Delta t$, where $\Delta t = 0.010$

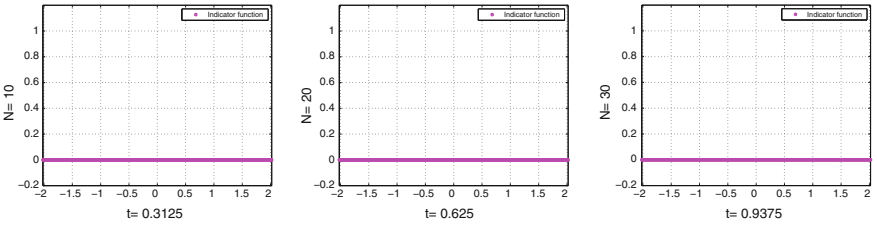


Fig. 4 (Example 2), the plot of the indicator function σ for (30) initial data at $t = 10\,\Delta t,\ 20\,\Delta t,\ 30\,\Delta t$, where $\Delta t = 0.010$

Tables 3 and 4. Local error in the L^∞ norms is computed in some subdomain D, which, at a given time t_n, corresponds to

$$e_{L^\infty_{loc}} := \max_{\{i,\ x_i \in D\}} |v(t_n, x_i) - u_i^n|$$

For this example, $D = [-2, -0.96] \cup [-0.84, 0.86] \cup [0.9, 2]$. Here, SL scheme is diffusive (see Figs. 3 and 4); on the other hand, UB scheme is nondiffusive and has very nice behaviour. We can see in Table 3 that coupled scheme switches to UB scheme as expected. Hence, our indicator is able to use the right scheme.

Table 3 (Example 2), errors for the SL-P^1+ UB coupled scheme or UB scheme with initial condition (30)

Δt	Δx	L^1 Error	L^2 Error	L^∞ Error	L^∞_{Loc}-Error
0.0400	0.1250	0.136882	0.187075	0.480895	0.088388
0.0200	0.0625	0.101730	0.163648	0.486520	0.062500
0.0100	0.0312	0.074883	0.142337	0.490479	0.044194
0.0050	0.0156	0.055000	0.123685	0.493272	0.031250
0.0025	0.0078	0.039634	0.105689	0.495244	0.022097

Table 4 (Example 2), errors for the SL-P^1 scheme with initial condition (30)

Δt	Δx	L^1 Error	L^2 Error	L^∞ Error	L^∞_{Loc}-Error
0.0400	0.1250	0.025000	0.088388	0.312500	0.187075
0.0200	0.0625	0.012500	0.062500	0.312500	0.163648
0.0100	0.0312	0.006250	0.044194	0.312500	0.142337
0.0050	0.0156	0.003125	0.031250	0.312500	0.123685
0.0025	0.0078	0.001563	0.022097	0.312500	0.105689

5 Conclusion and Future Work

We have presented a coupling between SL scheme and UB scheme for advection and HJ equations. The construction of the coupling is based on the computation of a switching indicator which allows to choose one of the two schemes in every cell according to the regularity properties of the solution and to some stability considerations. The technique behind the coupling can be applied also to other schemes and can be simplified when the two original schemes work on the same grid and have the same type of approximate values because in this situation we will not need to project the values on two different grids. The analysis for the advection problem shows that the coupled has some good properties which hopefully can be extended also to nonlinear Hamilton–Jacobi-type equations as the last example seems to suggest. This analysis as well as the extension to 2D problems will be the focus of a future work.

Acknowledgements I would like to thank Prof. Maurizio Falcone for the useful suggestions and discussion.

References

1. G. Barles, *Solution de Viscositè des èquations de Hamilton-Jacobi, Mathematiques et Applications* (Springer, Paris, 1994)
2. M. Bardi, D.I. Capuzzo, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations, Systems and Control Foundations and Applications* (Birkhäuser, Boston, 1997)

3. O. Bokanowski, N. Megdich, H. Zidani, An adaptive antidissipative method for optimal control problems. Rev. ARIMA **5**, 256–271 (2006)
4. O. Bokanowski, H. Zidani, Anti-dissipative schemes for advection and application to HJB equations. J. Sci. Comput. **30**(1), 1–33 (2007)
5. O. Bokanowski, N. Forcadel, H. Zidani, L^1-error estimates for numerical approximation of Hamilton-Jacobi-Bellman equation in dimension 1. Math. Comp. **79**, 1395–1426 (2010)
6. O. Bokanowski, N. Megdich, H. Zidani, Convergence of a non-monotone scheme for HJB equations with discontinuous initial data. Numer. Math. **115**(1), 1–44 (2010)
7. O. Bokanowski, M. Falcone, S. Sahu, An efficient filtered scheme for some first order time-dependent Hamilton-Jacobi equations. SIAM J. Sci. Comput. **38**(1), 171–195 (2016)
8. R. Courant, E. Isaacson, M. Rees, On the solution of nonlinear hyperbolic differential equations by finite differences. Comm. Pure Appl. Math. **5**, 243–255 (1952)
9. B. Desprè, F. Lagoutière, Contact discontinuity capturing schemes for linear advection and compressible gas dynamics. J. Sci. Comput. **16**, 479–524 (1999)
10. M. Falcone, R. Ferretti, *Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations* (SIAM-Society for Industrial and Applied Mathematics, Philadelphia, 2014)
11. M. Falcone, S. Sahu, *Coupled Scheme for Linear and Hamilton-Jacobi-Bellman Equations*. Submitted to Communications in Applied and Industrial Mathematics (CAIM) (2016)
12. B.D. Froese, A.M. Oberman, Convergent filtered schemes for the Monge-Ampre partial differential equation. SIAM J. Numer. Anal. **51**, 423–444 (2013)
13. C.B. Laney, *Computational Gasdynamics* (Cambridge University Press, New York, 1998)
14. P.L. Roe, Some contributions to the modeling of discontinuous flows. Lect. App. Math. **22**, 163–193 (1985)
15. S. Sahu, High order filtered scheme for front propagation problems. Bull. Braz. Math. Soc. **47**(2), 727–744 (2016)
16. S. Sahu, High-order filtered schemes for first order time dependent linear and non-linear partial differential equations. Math. Comput. Simul. **147**, 250–263 (2018)

Compressible Heterogeneous Two-Phase Flows

Nicolas Seguin

Abstract The modeling and the numerical simulation of two-phase flows are investigated for several decades. When dealing with very heterogeneous problems, for instance a water flow with many bubbles, one has to make use of averaged models since the description of each phase and interface is out of reach. Whatever the average is, the resulting models often suffer from severe mathematical pathologies: lack of hyperbolicity, non-conservative products, non-preservation of admissible states… In 1986, Baer and Nunziato proposed an original model which possesses interesting features from the mathematical point of view. Our goal is to provide a (partial) state of art on this model and its derivatives, but also to list some open questions.

Keywords Two-phase flows · Hyperbolicity · Nonconservative products Well-posedness

1 General Problems for the Modeling of Two-Phase Flows

We are interested in a flow of a fluid which is composed of two different phases. These phases are considered immiscible; i.e., with a perfect description, at a given point x, and for a given time t, only one phase is present. In other words, the spatial domain $\Omega \subset \mathbb{R}^d$, $d \geqslant 1$, can be divided in two disjoint regions $\Omega_1(t)$ and $\Omega_2(t)$ where 1 and 2 denote the label of the phase: $\Omega = \bar{\Omega}_1(t) \cup \bar{\Omega}_2(t)$ and $\Omega_1(t) \cap \Omega_2(t) = \emptyset$. Within each of these regions, the phase is governed by some classical equations of fluid dynamics, for instance the compressible Navier–Stokes equations. Moreover, the evolution of the interfaces $\Sigma(t) = \bar{\Omega}_1(t) \cap \bar{\Omega}_2(t)$ is deduced from some interfacial laws, such that the continuity of the velocity, the pressure… When the flow is more or less still, such a description is satisfactory and is relevant for numerical simulations. However, in many industrial contexts, the flow is very heterogeneous so that its numerical approximation is unfeasible. An example, already mentioned in the

N. Seguin (✉)
IRMAR, Université de Rennes 1, 35042 Rennes Cédex, France
e-mail: nicolas.seguin@univ-rennes1.fr

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_43

abstract, is a water flow with numerous bubbles of gas. For a complete description, the mesh size must be smaller that the size of the smallest bubble and the interfaces must be accurately discretized. This is clearly out of reach from a practical point of view. Moreover, a detailed knowledge of the structure of the flow is unnecessary since only the global behavior is interesting.

1.1 Averaged Models

In order to avoid the use of a full description of the two-phase flows, averaged models have been proposed. The average is done with respect to the separation of the two phases: at a given point x and for a given time t, the two phases may be simultaneously present. Whereas in the complete description the characteristic functions of the domains $\Omega_1(t)$ and $\Omega_2(t)$, say $\mathbf{1}_{\Omega_1(t)}(x)$ and $\mathbf{1}_{\Omega_2(t)}(x)$, are the unknowns which represent the repartition of the two phases, averaged models are based on new unknowns $\alpha_1(t,x)$ and $\alpha_2(t,x)$ which lie in $[0,1]$ (and not in $\{0,1\}$ as $\mathbf{1}_{\Omega_1(t)}(x)$ and $\mathbf{1}_{\Omega_2(t)}(x)$). They satisfy the relation

$$\alpha_1(t,x) + \alpha_2(t,x) = 1 \tag{1}$$

for all $t \geqslant 0$ and $x \in \mathbb{R}^d$ which corresponds to the immiscibility of the two phases. According to the average, $\alpha_k(t,x)$ may represent the probability of presence or the volume fraction of phase k and time t and at point x for example. Some processes of averaging of two-phase flows can be found in [24] or in [14].

After these averages, the quantities which characterize each phase also become averaged quantities: density, velocity, pressure, temperature… In order to keep the most accurate description, one may assume that these averaged quantities are different between the phases. For example, if u_1 and u_2 denote the averaged velocities of phases 1 and 2, respectively, then $u_1(t,x)$ and $u_2(t,x)$ are a priori different. One may hope to obtain some closed averaged model, which should differ according to the average:

- Time or space average. These are based on the existence of a small scale (typically the size of a bubble) and one performs some homogeneization process. See for instance [24, 25].
- Space dimension reduction. This case occurs when considering particular configuration. A classical example is a flow inside a pipe. If its length is very large with respect to the section, one may only consider the effects in the direction of the axis of the pipe. Therefore, the average is done by integration in the orthogonal direction, corresponding to the cross section. See for instance [35, 36, 41].
- Ensemble or statistical average. This is this kind of average which leads to the notion of probability of presence. A random variable is introduced, representing the uncertainty of the description of the actual flow. Therefore, a statistical average is done according to this random variable (this can be related to the notion of Young measures). See for instance [14] or [7] (the latter reference deals with moderately

heterogeneous flows). One can also refer to [1] where this approach is used to design numerical schemes.

Let us emphasize that this presentation is very incomplete and arbitrary…

Up to our knowledge, the paper [7] (with related works done by these authors) is the only one where a mathematically rigorous derivation of the averaged model is proposed. Unfortunately, it does not cover heterogeneous flows, where averaged velocities, pressures… are kept different. With the present knowledge in homogeneization of PDE's or on space dimension reduction, in particular for shallow water models, one may hope to derive properly averaged models.

1.2 The Baer–Nunziato Model

Here, we are interested in compressible inviscid phases, which may be described separately by the compressible Euler equations (or Navier–Stokes equations, if the viscosity of both phases tends to zero during the process of averaging). After averaging, each phase possesses its own (averaged) density ρ_k, velocity u_k, pressure p_k, specific total energy E_k, in addition to the fraction α_k. Surprisingly, the general structure of the averaged models is very similar whatever the averaging process is considered (see the references above). They all may be written as

$$\begin{cases} \partial_t \alpha_k + V_I(\mathbf{u})\partial_x \alpha_k = R_\alpha^k(\mathbf{u}), \\ \partial_t\, \alpha_k \begin{bmatrix} \rho_k \\ \rho_k u_k \\ \rho_k E_k \end{bmatrix} + \partial_x\, \alpha_k \begin{bmatrix} \rho_k u_k \\ \rho_k u_k^2 + p_k \\ (\rho_k E_k + p_k) u_k \end{bmatrix} - \begin{bmatrix} 0 \\ P_I(\mathbf{u}) \\ P_I V_I(\mathbf{u}) \end{bmatrix} \partial_x \alpha_k = \begin{bmatrix} R_\rho^k(\mathbf{u}) \\ R_m^k(\mathbf{u}) \\ R_E^k(\mathbf{u}) \end{bmatrix}, \end{cases} \tag{2}$$

where $\mathbf{u} = (\alpha_1, \alpha_1\rho_1, \alpha_1\rho_1 u_1, \alpha_1\rho_1 E_1, \alpha_2\rho_2, \alpha_2\rho_2 u_2, \alpha_2\rho_2 E_2)$ is the vector of the unknowns (α_2 can be expressed in function of α_1 thanks to (1)). Therefore, this model is composed by seven equations in one space dimension (note that we only present the one-dimensional version of this model but one can easily write it in the multidimensional setting).

The difficulty on this type of models lies in the definition of the terms which describe the interaction between the phases. The so-called interfacial velocity V_I and interfacial pressure P_I have no unique form, and many definitions can be found in the literature [5, 9, 17, 37]… The same remark holds for the source terms $R_\alpha^k(\mathbf{u})$, $R_\rho^k(\mathbf{u})$, $R_m^k(\mathbf{u})$, and $R_E^k(\mathbf{u})$. As mentioned above, even for given type of flows, there exists no rigorous derivation of averaged models (2) which should enable us to define these quantities without ambiguity. Nonetheless, in order to recover the conservation of the global mass, momentum, and energy, we have

$$\text{for } \zeta = \rho, m, E, \quad \sum_{k=1,2} R_\zeta^k = 0.$$

We also have, as a consequence of (1), $R_\alpha^1 + R_\alpha^2 = 0$.

The model (2) involves a (mean) specific internal energy e_k, such that

$$E_k = e_k + \frac{1}{2}|u|^2$$

together with an equation of state which writes, T_k denoting the temperature of phase k,

$$T_k \mathrm{d}s_k = \mathrm{d}e_k + p_K \mathrm{d}(1/\rho_k).$$

Note that one assumes in general that these equations of state belong to the class of classical equations of state of fluids. However, they are written for averaged quantities and should a priori differ from the original equations of state of pure phases, since the averaging processes have a little chance to commute with the nonlinearity of the equations of state.

Let us now focus on the structure of this model and on its properties.

2 Analysis of Baer–Nunziato Type Models

In order to provide a complete overview, we should compare the properties of the solutions of models of the form (2) with other two-phase flows models. However, this would be a very hard task due to the wide variety of such models but, up to our knowledge, Baer–Nunziato type models possess the most rational properties.

2.1 Basic Properties

Let us first present the properties which are independent of the definition of the interfacial quantities V_I and P_I and of the source terms.

Proposition 1. *1. If α_1 is constant and if the source terms are set to zero, then system (2) becomes two independent systems composed by the Euler equations for each phase.*

2. *If the source term R_α^1 vanishes when $\alpha = 0$ and 1, then $\alpha_1(t, x) \in [0, 1]$ (idem for α_2) for all $x \in \mathbb{R}$ and $t > 0$ as soon as $\alpha_1(0, x) \in [0, 1]$.*
3. *The differential part of system (2) (i.e., without source term) possesses the following structure:*
   - *the eigenvalues are V_I, u_k, and $u_k \pm c_k$ for $k = 1, 2$ (c_k being the classical sound speed computed from the equation of state of phase k),*
   - *the eigenvectors form a basis of $\mathbb{R}^7$ except if $V_I = u_k \pm c_k$ (this is the so-called resonance),*

- *the characteristic fields associated with the waves $u_k \pm c_k$ are genuinely non-linear and the characteristic fields associated with the waves u_k are linearly degenerate.*

These properties have been shown in several works; see for instance [15, 17]. One can remark the robustness and the relative simplicity of the structure of the model.

Moreover, let us consider the Riemann problem, that is to say system (2) without source term, which may be written shortly (with obvious notations)

$$\partial_t \mathbf{u} + \partial_x f(\mathbf{u}) + c(\mathbf{u})\partial_x \alpha_1 = 0, \tag{3}$$

with initial condition

$$\mathbf{u}(0, x) = \begin{cases} \mathbf{u}_L & \text{if } x < 0 \\ \mathbf{u}_R & \text{if } x > 0 \end{cases} \tag{4}$$

where $\mathbf{u}_L$ and $\mathbf{u}_R$ belong to the set of admissible states $\mathscr{A} = \{\mathbf{u} \in \mathbb{R}^7, \alpha_1 \in (0, 1), \rho_k > 0, T_k > 0\}$. Once again, even without the knowledge of V_I and P_I, one can obtain the following properties of self-similar solutions (see for instance [15] or [17]).

Proposition 2. *Consider self-similar solutions of the Riemann problem (3–4). The void fractions α_k are constant on each part away from the wave V_I. Moreover, all the waves which are not superposed with the wave V_I are defined by the classical relations from the Euler equations (using Rankine–Hugoniot jump relations and Riemann invariants).*

As a consequence, up to an appropriate definition of the wave V_I, one can expect that the solution(s) of the Riemann problem (3–4) lies in $\mathscr{A}$. Let us emphasize that this kind of property is rather difficult to prove in general with other two-phase flows models (and actually, it is not verified for many models).

2.2 Entropy and Symmetric Form

Let us continue to present the properties of (2) which are independent of the definitions of V_I and P_I. The classical mathematical entropy for such models of two-phase flows is

$$\eta(\mathbf{u}) = -\alpha_1 \rho_1 s_1 - \alpha_2 \rho_2 s_2. \tag{5}$$

One can check that

Proposition 3 ([13])**.** *The entropy η defined by (5) is non-strictly convex, in the sense that the hessian matrix $\mathrm{D}^2\eta$ is positive semi-definite on $\mathscr{A}$.*

This property is difficult to use in practice, due to the lack of strict convexity of η. Indeed, as far as V_I and P_I are not given, one cannot derive the PDE satisfied by η.

Even forgetting the singularity of $\mathrm{D}^2\eta$, since system (2) is not in a conservative form, one cannot apply the Godunov–Mock results to deduce a symmetric form of (3), that would be

$$P(\mathbf{y})\partial_t \mathbf{y} + Q(\mathbf{y})\partial_x \mathbf{y} = 0 \tag{6}$$

where P is a symmetric positive-definite matrix and Q is a symmetric matrix, while $\mathbf{y}$ is obtained from $\mathbf{u}$ by a $\mathscr{C}^1$-diffeomorphism. Nonetheless, this can be done separately:

Lemma 1 ([13])**.** *System* (3) *admits a symmetric form* (6) *if and only of* $\mathbf{u}$ *is not resonant, i.e.,* $V_I \neq u_k \pm c_k$, $k = 1, 2$.

This property is crucial to obtain the local well-posedness of small smooth solutions (assuming smooth pressure laws interfacial velocity and pressure, and source terms) of the Cauchy problem. Following Kato [28], we have:

Theorem 1. *Assume that the initial data* $\mathbf{u}_0$ *are continuously differentiable on* $\mathbb{R}$*, take values in some compact subset of* $\mathscr{A}$ *far from the set of resonant states and such that* $\mathbf{u}_0'(\cdot) \in \mathbf{H}^\ell(\mathbb{R})$ *with* $\ell \geqslant 1$*. Then there exists* $T \leqslant +\infty$ *and a unique continuously differentiable function* $\mathbf{u}$ *on* $[0, T) \times \mathbb{R}$ *taking values in* $\mathscr{A}$*, which is a classical solution of the Cauchy problem for* (2) *on* $[0, T)$*. Furthermore,*

$$\partial_x \mathbf{u}(t, \cdot) \in \bigcap_{k=0}^{\ell} \mathscr{C}^k([0, T), \mathbf{H}^{\ell-k}(\mathbb{R})).$$

Since we are dealing with nonlinear hyperbolic equations, one cannot expect a global well-posedness theorem, i.e., that $T = +\infty$, at least without more assumptions.

Remark 1. Let us make some comments on the resonance case. Even if the interfacial velocity is not defined, in general it is a combination of the phase velocities u_1 and u_2. As a consequence, the interfacial velocity is more likely to be very small in comparison to the acoustic velocities $u_k \pm c_k$, $k = 1, 2$. As a consequence, this theorem is relevant for most realistic configurations since the fluid velocities are about 10 m/s while the the sound speeds are about several hundreds of m/s. Besides, the source terms R_m^k include drag forces which make the relative velocity $|u_1 - u_2|$ tend to zero. This should prevent even more the appearance of resonant states.

3 Mathematical Closure of Baer–Nunziato Type Models

As we mentioned before, many different propositions of the terms which remain to define exist in the literature. In [10, 17], a mathematical point of view is adopted, which means that the definitions of V_I and P_I are deduced from mathematical properties of solutions.

3.1 Interfacial Velocity

Let us first investigate the definition of V_I. In order to consider it, one may assume for the moment that the source terms are null. In this case, this velocity corresponds to the speed of transport of the composition of the two-phase mixture since it appears in the propagation of the void fraction α_1, which describes the repartition of the two phases.

Proposition 4 ([10, 17])**.** *Assume that the interfacial velocity writes, with* $\beta_V(\mathbf{u}) \in [0, 1]$,

$$V_I(\mathbf{u}) = \beta_V(\mathbf{u})u_1(\mathbf{u}) + (1 - \beta_V(\mathbf{u}))u_2(\mathbf{u}). \tag{7}$$

The following assertions are equivalent:

1. *An discontinuity of* α_1 *in the initial data remains a discontinuity for all* $t > 0$.
2. *The characteristic field associated with* V_I *is linearly degenerate.*
3. *The weight* β_V *is given by*

$$\beta_V(\mathbf{u}) = 1, \quad \textit{either} \quad \beta_V(\mathbf{u}) = 0, \quad \textit{or} \quad \beta_V(\mathbf{u}) = \frac{\alpha_1\rho_1}{\alpha_1\rho_1 + \alpha_2\rho_2}.$$

As a consequence, the interfacial velocities

$$V_I(\mathbf{u}) = u_1, \quad V_I(\mathbf{u}) = u_2, \quad \text{and} \quad V_I(\mathbf{u}) = \frac{\alpha_1\rho_1u_1 + \alpha_2\rho_2u_2}{\alpha_1\rho_1 + \alpha_2\rho_2}$$

play very particular roles. Actually, these closures are very common in the literature, but in general, they are motivated only by heuristic arguments.

Remark 2. Actually, the statement of this proposition is not fully exact, some additional assumptions on β have to be added, such as restricting to definitions which preserve the Galilean invariance.

Remark 3. Assume that we are in the case of the two first choices, say $V_I(\mathbf{u}) = u_1$. This wave corresponds to an eigenvalue of multiplicity 2 in 1D, decreases the number of possible resonant cases decreases: V_I may only interact with $u_2 \pm c_2$ and no longer with $u_1 \pm c_1$. Moreover, the system (3) has a particular structure which may be used for operator splitting and numerical approximation [12].

3.2 Interfacial Pressure

We already have introduce the Lax mixture entropy η in Eq. (5). The natural entropy flux F associated with η is

$$F(\mathbf{u}) = -\alpha_1\rho_1s_1u_1 - \alpha_2\rho_2s_2u_2. \tag{8}$$

Even if the entropy η is not strictly convex, it is natural to assume that it satisfies a balance law (or a conservation law for zero source terms) for smooth solutions.

Proposition 5 ([10, 17])**.** *Assume that the interfacial pressure writes, with* $\beta_P(\mathbf{u}) \in [0, 1]$,

$$P_I(\mathbf{u}) = \beta_P(\mathbf{u})p_1(\mathbf{u}) + (1 - \beta_P(\mathbf{u}))p_2(\mathbf{u}). \tag{9}$$

Then, smooth solutions of Baer–Nunziato type models (2) *with interfacial velocities of the form* (7) *satisfy the balance law*

$$\partial_t \eta(\mathbf{u}) + \partial_x F(\mathbf{u}) = \nabla_{\mathbf{u}} \eta(\mathbf{u}) \cdot r(\mathbf{u}) \tag{10}$$

if and only if

$$\beta_P(\mathbf{u}) = \frac{(1 - \beta_V)T_2}{\beta_V T_1 + (1 - \beta_V)T_2}.$$

It is worth noting the two couples $(V_I, P_I) = (u_1, P_2)$ and $(V_I, P_I) = (u_2, P_1)$ which often appear in the literature, in particular in the original paper by Baer and Nunziato [5], while the last couple appeared for the first time in [10, 17].

3.3 The Interfacial Wave and the Riemann Problem

We assume now that V_I and P_I are defined by one of the three couples provided by the last two propositions. As a consequence, the wave associated with the eigenvalue V_I (the interfacial wave) is linearly degenerate. This may be defined using Riemann invariants and Rankine–Hugoniot jump relations. In the case of resonance, the situation becomes more complicated since the relations of the acoustic wave which interacts with the interfacial wave are incompatible with the relations of the interfacial wave. In order to define this interaction, a blowup in α_1 is usually used. In other words, a single interfacial wave interacting with an acoustic shock wave is defined as the limit traveling waves, provided by a regularized void fraction. This regularization is chosen smooth and monotone. Such an approach is widely used for scalar conservation laws with singularities in space [4, 39].

The homogeneous system (3) has a very particular structure, first described in the scalar case by Issacson and Temple [23] and extended to systems in [19]. Basically, the vector $c(\mathbf{u})$ does not vanish when the resonance occurs. Therefore, one has the following negative result:

Proposition 6. *The Riemann problem* (*3–4*) *may have up to three self-similar solutions.*

As a consequence, the Cauchy problem may admit an infinite number of solutions. Nonetheless, one can check that, as soon as we are interested in solutions with moderate Mach numbers, one may recover uniqueness, but finding explicit conditions on initial data to fulfill this requirement seems difficult.

4 Dissipative Source Terms

In the previous section, we have mainly addressed homogeneous Baer–Nunziato type models, and we now focus on the source terms. We do not plan to provide explicit formulas but only their general structure and the impact of such choices. We are only concerned with internal effects due to exchanges between the two phases.

4.1 Derivation and Basic Properties

Let us comment their origin. For instance, it is natural to assume that, before averaging, the pressure is continuous through the interfaces between phases. When the average is applied, the averaged pressures are not equal in general and the notion of interface disappear. Then, the impact of the continuity of the pressure through interfaces appears in relaxation source terms which, in absence of other effects, make the the relative pressure tend to zero, $|p_1 - p_2| \to 0$ when $t \to +\infty$. This source term appears at least in the equation of α_1 since a difference of pressure leads to variations of volume (e.g., α_1 increases in time if p_1 is greater than p_2).

Actually, this process of occurrence of source terms may be obtained in heuristic ways, see [14] for instance, or in some cases by rigorous derivation [7]. Let us list all the possible source terms which may appear in Baer–Nunziato type models:

- Mechanical exchanges. This provides a relaxation of the relative pressure $p_1 - p_2$ with equilibrium $p_1 = p_2$.
- Drag force. It comes from the friction between the two phases at the interfaces. It leads to a relaxation of the relative velocity $u_1 - u_2$, with equilibrium $u_1 - u_2$. The term appears in the momentum equations and is quadratic, i.e., of the form $|u_1 - u_2|(u_1 - u_2)$.
- Temperature exchanges. This term appears in the equations of energies and corresponds to the relaxation in temperature such the the equilibrium is $T_1 = T_2$.
- Mass transfer. This effect only appears when two phases of the same fluid are considered, such as liquid water and steam. It corresponds to the phase transition and the associated equilibrium is the equality of chemical potentials, $\mu_1 = \mu_2$, which are defined by
$$\mu_k = e_k + p_k/\rho_k - T_k s_k.$$

The source terms, R_ζ^k, $\zeta = \alpha, \rho, m, E$, are combination of these effects, with nonlinear factors. Let us recall the notation $r(\mathbf{u}) = (R_\alpha^1, R_\rho^1, R_m^1, R_E^1, R_\rho^2, R_m^2, R_E^2)^\top$. Since they lead to some equilibria, they have to be entropy dissipative, that is to say

$$\nabla_{\mathbf{u}}\eta(\mathbf{u}) \cdot r(\mathbf{u}) \leqslant 0.$$

The equality

$$\nabla_{\mathbf{u}}\eta(\mathbf{u}) \cdot r(\mathbf{u}) = 0$$

holds if and only if $\mathbf{u}$ is an equilibrium state; i.e., it belongs to the equilibrium manifold

$$\mathscr{E} = \{\mathbf{u} \in \mathscr{A} \mid p_1 = p_2, T_1 = T_2, u_1 = u_2\}$$

if there is no mass transfer, or to

$$\mathscr{E}_{\mathrm{mt}} = \{\mathbf{u} \in \mathscr{A} \mid p_1 = p_2, T_1 = T_2, u_1 = u_2, \mu_1 = \mu_2\}$$

with mass transfer. In the space homogeneous case, i.e.,

$$\partial_t \mathbf{u} = r(\mathbf{u}), \tag{11}$$

the entropy is not a Lyapunov function due to the lack of strict convexity. Moreover, the equilibrium sets $\mathscr{E}$ and $\mathscr{E}_{\mathrm{mt}}$ only correspond to the case when the two phases are present. Actually, appearance or disappearance of phases should be also taken into account, but in this case the solution leaves the set $\mathscr{A}$ and the phasic variables are no longer defined. Some tentatives with a simpler PDE model are done for instance in [2, 21]; see also [26]. But, up to our knowledge, this problem is far from being understood.

Remark 4. In general, in order to ensure that the set of admissible states $\mathscr{A}$ is an invariant domain for the differential system (11), one includes the factor $\alpha_1(1-\alpha_1)$ in the definition of the source terms. Together with smoothness assumptions, this ensures that the void fractions remain in $[0, 1]$. However, this can create new stable equilibrium states.

4.2 Relaxation and Hierarchy of Models

Since the source terms are entropy stable, one can hope a deeper impact on the solutions of the Cauchy problem.

Following the pioneer work by Chen, Levermore, and Liu [8], one can study the structure of the relaxation of model (2). One of the main difference is due to the non-conservative form of the equations. Nevertheless, some properties such as the sub-characteristic condition and the parabolic behavior after Chapman–Enskog expansion can be investigated (see for instance [6, 33] for discussions in the conservative case). This approach has been followed by several authors, for instance in [27], and has deserved a detailed attention by a Norwegian group (see in particular [16, 29, 32]).

An important question, studied in several works mentioned here, is the form of reduced equations, when characteristic times of some relaxation effects become infinitely fast. An important step has been done by Kapila et al. in [27]. They obtain a limit model which has many interesting properties. However, it is written in a non-conservative form, and up to now, the definition of the non-conservative products

remains unclear. An important assumption in the derivation provided in [27] is that the characteristic times of all relaxation terms are of the same order. In some configurations, this assumption is no longer true and other limit models can be obtained, the so-called drift models; see [3, 20] for instance. This kind of asymptotics involves parabolic term and is far from being understood in this context (see [34] for first results in the classical setting).

4.3 Nonlinear Stability

Indeed, when studying system of balance laws with entropy dissipative source terms, the global existence of small and smooth solutions holds. In other terms, $T = +\infty$ in Theorem 1 for small data (i.e., initial data close to a constant state of the equilibrium manifold). This result is valid under some assumptions, which we recall here in an approximate way:

- The source term is entropy dissipative.
- The system of balance laws admits a symmetric form.
- The Jacobian of the source term is non-singular in the equilibrium manifold.
- The source term does not vanish in the eigenspaces of the convection matrix.

As mentioned above, the first assumption is satisfied by construction of the source terms. The second one holds if resonant states are not considered [13]. As we discussed before, starting with initial data in the vicinity of the equilibrium manifold $\mathscr{E}$ or $\mathscr{E}_{\mathrm{mt}}$ should prevent the occurrence of resonant states. The third assumption ensures a linear local behavior of the relaxation terms in the equilibrium manifold. This is the case of all the source terms described above, except for the drag force. Up to our knowledge, quadratic relaxation terms have not been studied yet in a general setting.

The last assumption is called the Kawashima condition. It has been introduced first in [40], for parabolic perturbations of systems of conservation laws. It ensures that all the waves of the system are impacted by the relaxation. Unfortunately, this is not the case for Baer–Nunziato type models, where non-trivial equilibrium solutions exists (traveling contact discontinuities with constant and equal velocities, pressures and temperatures). In [30], Mascia and Natalini investigate systems where the Kawashima condition is violated, and one may hope to apply their analysis to (2) (but some work remains…).

5 Some Additional Remarks

Let us recall that this note is only a partial review of the works on Baer–Nunziato type models. The bibliographical list is far from being exhaustive. Moreover, the question of extensions of this type of models is important and is not addressed here (see for examples [18, 22]). Concerning the numerical approximation, since this models

correspond to hyperbolic systems of PDEs, finite volume scheme is generally used. Actually, very few preserve the set of admissible states and satisfy a discrete version of the entropy inequality (10) (see [11, 12, 38], and also [1]). We did not insist on the different models which are involved in the hierarchy obtained in the cascade of asymptotics. Most of them should also deserve attention, as well as the measure of the actual difference between each model of the hierarchy. Such a study would be of great importance from a practical point of view, since the complete model (2) could be locally replaced by simpler ones without altering the global accuracy (see for instance [31] for a first attempt).

Acknowledgements During his position in UPMC-Paris 6, the author has been supported by the LRC Manon (Modélisation et approximation numérique orientées pour l'énergie nucléaire—CEA/DM2S-LJLL).

References

1. R. Abgrall, R. Saurel, Discrete equations for physical and numerical compressible multiphase mixtures. J. Comput. Phys. **186**(2), 361–396 (2003)
2. G. Allaire, G. Faccanoni, S. Kokh, Modelling and simulation of liquid-vapor phase transition in compressible flows based on thermodynamical equilibrium. M^2 *AN* **46**:1029–1054 (2012)
3. A. Ambroso, C. Chalons, F. Coquel, T. Galié, E. Godlewski, P.-A. Raviart, N. Seguin, The drift-flux asymptotic limit of barotropic two-phase two-pressure models. Commun. Math. Sci. **6**(2), 521–529 (2008)
4. B. Andreianov, N. Seguin, Analysis of a Burgers equation with singular resonant source term and convergence of well-balanced schemes. Discrete Contin. Dyn. Syst. **32**(6), 1939–1964 (2012)
5. M.R. Baer, J.W. Nunziato, A two-phase mixture theory for the deflagration-to-detonation transition (DDT) in reactive granular materials. Int. J. Multiph. Flow **12**, 861–889 (1986)
6. F. Bouchut, A reduced stability condition for nonlinear relaxation to conservation laws. J. Hyperb. Diff. Equ. **1**(1), 149–170 (2004)
7. D. Bresch, M. Hillairet, Note on the derivation of multi-component flow systems. Proc. Am. Math. Soc. **143**(8), 3429–3443 (2015)
8. G.Q. Chen, C.D. Levermore, T.P. Liu, Hyperbolic conservation laws with stiff relaxation terms and entropy. Comm. Pure Appl. Math. **47**(6), 787–830 (1994)
9. A. Chinnayya, E. Daniel, R. Saurel, Modelling detonation waves in heterogeneous energetic materials. J. Comput. Phys. **196**(2), 490–538 (2004)
10. F. Coquel, T. Gallouët, J.-M. Hérard, N. Seguin, Closure laws for a two-fluid two-pressure model. C. R. Math. Acad. Sci. Paris **334**(10), 927–932 (2002)
11. F. Coquel, J.-M. Hérard, K. Saleh, A positive and entropy-satisfying finite volume scheme for the Baer–Nunziato model. J. Comput. Phys. (to appear)
12. F. Coquel, J.-M. Hérard, K. Saleh, N. Seguin, A robust entropy-satisfying finite volume scheme for the isentropic Baer-Nunziato model. M2AN Math. Model. Numer. Anal. **48**, 165–206 (2013)
13. F. Coquel, J.-M. Hérard, K. Saleh, N. Seguin, Two properties of two-velocity two-pressure models for two-phase flows. Commun. Math. Sci. **12**(3), 593–600 (2014)
14. D.A. Drew, S. Passman, *Theory of Multicomponent Fluids* (Springer, New York, 1998)
15. P. Embid, M. Baer, Mathematical analysis of a two-phase continuum mixture theory. Contin. Mech. Thermodyn. **4**(4), 279–312 (1992)
16. S. Evje, T. Flåtten, On the wave structure of two-phase flow models. SIAM J. Appl. Math. **67**(2), 487–511 (2006/07)

17. T. Gallouët, J.-M. Hérard, N. Seguin, Numerical modeling of two-phase flows using the two-fluid two-pressure approach. Math. Models Methods Appl. Sci. **14**(5), 663–700 (2004)
18. S. Gavrilyuk, R. Saurel, Mathematical and numerical modelling of two phase compressible flows with micro-inertia. J. Comp. Phys. **175**(1), 326–360 (2002)
19. P. Goatin, P.G. LeFloch, The Riemann problem for a class of resonant hyperbolic systems of balance laws. Ann. Inst. H. Poincaré Anal. Non Linéaire **21**(6), 881–902 (2004)
20. H. Guillard, F. Duval, A Darcy law for the drift velocity in a two-phase flow model. J. Comput. Phys. **224**(1), 288–313 (2007)
21. P. Helluy, N. Seguin, Relaxation models of phase transition flows. M2AN Math. Model. Numer. Anal. **40**(2), 331–352 (2006)
22. J.-M. Hérard, A three-phase flow model. Math. Comput. Model. **45**(5–6), 732–755 (2007)
23. E. Isaacson, B. Temple, Convergence of the 2×2 Godunov method for a general resonant nonlinear balance law. SIAM J. Appl. Math. **55**(3), 625–640 (1995)
24. M. Ishii, *Thermo-Fluid Dynamic Theory of Two-Phase Flows* (Collection de la Direction des Etudes et Recherches d'Électicité de France, 1975)
25. M. Ishii, T. Hibiki, *Thermo-Fluid Dynamics of Two-Phase Flow* (Springer, New York, 2006). With a foreword by Lefteri H. Tsoukalas
26. F. James, H. Mathis, Modeling phase transition and metastable phases, in *Finite Volumes for Complex Applications. VII. Elliptic, Parabolic and Hyperbolic Problems*. Springer Proceedings in Mathematics and Statistics, vol. 78 (Springer, Cham, 2014), pp. 865–872
27. A.K. Kapila, R. Menikoff, J.B. Bdzil, S.F. Son, D.S. Stewart, Two-phase modeling of deflagration-to-detonation transition in granular materials: reduced equations. Phys. Fluids **13**(10), 3002–3024 (2001)
28. T. Kato, The Cauchy problem for quasi-linear symmetric hyperbolic systems. Arch. Ration. Mech. Anal. **58**(3), 181–205 (1975)
29. P.J.M. Ferrer, T. Flåtten, S.T. Munkejord, On the effect of temperature and velocity relaxation in two-phase flow models. ESAIM Math. Model. Numer. Anal. **46**(2), 411–442 (2012)
30. C. Mascia, R. Natalini, On relaxation hyperbolic systems violating the Shizuta-Kawashima condition. Arch. Ration. Mech. Anal. **195**(3), 729–762 (2010)
31. H. Mathis, C. Cancès, E. Godlewski, N. Seguin, Dynamic model adaptation for multiscale simulation of hyperbolic systems with relaxation. J. Sci. Comput. **63**(3), 820–861 (2015)
32. A. Morin, T. Flåtten, A two-fluid four-equation model with instantaneous thermodynamical equilibrium. ESAIM Math. Model. Numer. Anal. **50**(4), 1167–1192 (2016)
33. R. Natalini, Recent results on hyperbolic relaxation problems, in *Analysis of Systems of Conservation Laws (Aachen, 1997)*. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics, vol. 99 (Chapman & Hall/CRC, Boca Raton, FL, 1999), pp. 128–198
34. Y.-J. Peng, V. Wasiolek, Uniform global existence and parabolic limit for partially dissipative hyperbolic systems. J. Diff. Eq. **260**(9), 7059–7092 (2016)
35. V.H. Ransom, D.L. Hicks, Hyperbolic two-pressure models for two-phase flow. J. Comp. Phys. **53**(1), 124–151 (1984)
36. V.H. Ransom, D.L. Hicks, Hyperbolic two-pressure models for two-phase flow revisited. J. Comp. Phys. **75**(2), 498–504 (1988)
37. R. Saurel, R. Abgrall, A multiphase Godunov method for compressible multifluid and multiphase flows. J. Comput. Phys. **150**(2), 425–467 (1999)
38. D.W. Schwendeman, C.W. Wahle, A.K. Kapila, The Riemann problem and a high-resolution Godunov method for a model of compressible two-phase flow. J. Comput. Phys. **212**(2), 490–526 (2006)
39. N. Seguin, J. Vovelle, Analysis and approximation of a scalar conservation law with a flux function with discontinuous coefficients. Math. Mod. Meth. Appl. Sci. (M^3 AS) **13**(2), 221–250 (2003)
40. Y. Shizuta, S. Kawashima, Systems of equations of hyperbolic-parabolic type with applications to the discrete boltzmann equation. Hokkaido Math. J. **14**, 249–275 (1985)
41. H.B. Stewart, B. Wendroff, Two-phase flow: models and methods. J. Comput. Phys. **56**(3), 363–409 (1984)

Bound-Preserving High-Order Schemes for Hyperbolic Equations: Survey and Recent Developments

Chi-Wang Shu

Abstract Solutions to many hyperbolic equations have convex invariant regions, for example, solutions to scalar conservation laws satisfy the maximum principle, solutions to compressible Euler equations satisfy the positivity-preserving property for density and internal energy. It is, however, a challenge to design schemes whose solutions also honor such invariant regions. This is especially the case for high-order accurate schemes. In this contribution, we survey strategies in the recent literature to design high-order bound-preserving schemes, including a general framework in constructing high-order bound-preserving finite volume and discontinuous Galerkin schemes for scalar and systems of hyperbolic equations through a simple scaling limiter and a convex combination argument based on first-order bound-preserving building blocks, and various flux limiters to design high-order bound-preserving finite difference schemes. We also discuss a few recent developments, including high-order bound-preserving schemes for relativistic hydrodynamics, high-order discontinuous Galerkin Lagrangian schemes, and high-order discontinuous Galerkin methods for radiative transfer equations.

Keywords Bound-preserving · High order schemes · Hyperbolic equations

1 Introduction

We are interested in numerically solving hyperbolic conservation laws

$$u_t + \nabla \cdot \mathbf{F}(u) = 0, \qquad u(\mathbf{x}, 0) = u_0(\mathbf{x}) \tag{1}$$

in a bounded domain with periodic or other types of boundary conditions. We are also interested in other related hyperbolic or convection dominated equations. In par-

C.-W. Shu (✉)
Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
e-mail: shu@dam.brown.edu

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_44

ticular, we are interested in the bound-preserving properties of high-order numerical schemes solving such equations.

We assume the exact solution of the PDE (1) has a *convex* invariant region G:

- If $u(\cdot, 0) \in G$, then $u(\cdot, t) \in G$ for all $t > 0$.

For a convex region G, if $u_1, \cdots, u_m \in G$, $\alpha_i \geq 0$, $\sum_{i=1}^m \alpha_i = 1$, then $u = \sum_{i=1}^m \alpha_i u_i \in G$. We will heavily use this property when building our high-order bound-preserving schemes.

We now give several examples of invariant regions:

1. If Eq. (1) is a scalar conservation law, an important property of its entropy solution (which may be discontinuous) is that it satisfies a strict maximum principle:
$$m = \min_{\mathbf{x}} u_0(\mathbf{x}), \qquad M = \max_{\mathbf{x}} u_0(\mathbf{x}), \tag{2}$$
If, then $u(\mathbf{x}, t) \in [m, M]$ for any $\mathbf{x}$ and t. Therefore, $G = [m, M]$ is an invariant region. It is clearly convex.
2. Consider the Eq. (1) as the compressible Euler system:
$$u_t + f(u)_x = 0$$
with
$$u = \begin{pmatrix} \rho \\ \rho v \\ E \end{pmatrix}, \qquad f(u) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ v(E + p) \end{pmatrix},$$
where $E = e + \frac{1}{2}\rho v^2$, and the internal energy e is related to density and pressure through an *equation of states (EOS)*. For the ideal gas, we have $e = \frac{p}{\gamma - 1}$ with $\gamma = 1.4$ for air. In this case, we can verify that the set
$$G = \{u : \ \rho \geq 0, \ \ e \geq 0\} \tag{3}$$
is invariant. It is also easy to check that G is convex (for this we need to check that the internal energy e is a concave function of the conservative variable u, then Jensen's inequality implies the convexity of G).
For many EOS's, e.g., that for the ideal gas, the region G defined in (3) is equivalent to
$$G = \{u : \ \rho \geq 0, \ \ p \geq 0\}.$$
In such cases, we can talk about the positivity of density and pressure, instead of the positivity of density and internal energy.
Even though we discuss the one-dimensional case here for simplicity, G defined in (3) is also a convex invariant region in the multi-dimensional case.
3. Consider the relativistic hydrodynamics
$$u_t + f(u)_x = 0$$

with

$$u = \begin{pmatrix} D \\ m \\ E \end{pmatrix}, \qquad f(u) = \begin{pmatrix} Dv \\ mv + p \\ m \end{pmatrix}$$

where p, D, m, and E are the thermal pressure, mass density, momentum, and energy, respectively. v is the velocity. Moreover, units are normalized such that the speed of light is $c = 1$. If we denote ρ to be the proper rest-mass density, then the conservative variable u can be written as

$$\begin{aligned} D &= \gamma\rho, \\ m &= Dh\gamma v, \\ E &= Dh\gamma - p, \end{aligned}$$

where $\gamma = (1 - v^2)^{-1/2}$ is the Lorentz factor and h is the specific enthalpy. To close the system, we specify an equation of state $h = h(p, \rho)$. For ideal gas

$$\rho h = \rho + p\Gamma/(\Gamma - 1)$$

with Γ being the specific heat ratio, such that $1 < \Gamma \leq 2$.
It can be shown that the density D and pressure p are positive, and the velocity satisfies $v^2 \leq 1$, if they are initially in these cases. Therefore,

$$G = \{u : \ D > 0, \ E > 0, \ p > 0, \ v^2 \leq 1\} \tag{4}$$

is an invariant region. It is convex and can be represented as

$$G = \{u : \ D > 0, \ E > \sqrt{D^2 + m^2}\}.$$

See [12, 21] for more details.

2 Bound-Preserving First-Order Schemes

It is of course desirable to have the invariant region G also to be an invariant region for the numerical solution. That is, we wish that, if the initial condition $u(\cdot, 0) \in G$ then $u(\cdot, t) \in G$ for later time $t > 0$. This time u stands for the numerical solution.

We first consider fulfilling this task for first-order schemes.

For scalar conservation laws, first-order monotone schemes can easily maintain the maximum principle. For example, for the one-dimensional scalar conservation law

$$u_t + f(u)_x = 0,$$

the first-order monotone scheme

$$\begin{aligned} u_j^{n+1} &= H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n) \\ &= u_j^n - \lambda[h(u_j^n, u_{j+1}^n) - h(u_{j-1}^n, u_j^n)] \end{aligned}$$

where $\lambda = \frac{\Delta t}{\Delta x}$ and $h(u^-, u^+)$ is a monotone flux (h is non-decreasing in its first argument and non-increasing in its second argument, symbolically $h(\uparrow, \downarrow)$), satisfies the monotonically non-decreasing property in all its arguments

$$H_\lambda(\uparrow, \uparrow, \uparrow)$$

under a suitable CFL condition

$$\lambda \le \lambda_0. \tag{5}$$

Also, for any constant c, we have

$$H_\lambda(c, c, c) = c - \lambda[h(c, c) - h(c, c)] = c.$$

Therefore, if

$$m \le u_{j-1}^n, u_j^n, u_{j+1}^n \le M$$

, then

$$u_j^{n+1} = H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n) \ge H_\lambda(m, m, m) = m,$$

and

$$u_j^{n+1} = H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n) \le H_\lambda(M, M, M) = M.$$

Thus, the scheme satisfies the maximum principle under the CFL condition (5).

For compressible Euler equations, there are several first-order schemes, including the Godunov scheme, Lax–Friedrichs scheme, kinetic scheme, HLLC scheme, which satisfy the bound-preserving property for positive density and internal energy (or positive density and pressure for certain EOS), under suitable CFL condition (5).

For relativistic hydrodynamics, the first-order Lax–Friedrichs scheme is bound-preserving for the invariant region G defined in (4), under suitable CFL condition (5). See [12, 21] for more details.

We emphasize that it is often already non-trivial to find first-order schemes which are bound-preserving, e.g., for MHD equations. Since our high-order bound-preserving schemes discussed later are built upon first-order bound-preserving schemes, the very first task when one would like to solve a new PDE is to find a first-order bound-preserving scheme.

3 Bound-Preserving High-Order Schemes

For higher-order *linear* schemes, i.e., schemes which are linear for a linear PDE

$$u_t + au_x = 0, \tag{6}$$

for example, the second-order accurate Lax–Wendroff scheme

$$u_j^{n+1} = \frac{a\lambda}{2}(1 + a\lambda)u_{j-1}^n + (1 - a^2\lambda^2)u_j^n - \frac{a\lambda}{2}(1 - a\lambda)u_{j+1}^n$$

where $\lambda = \frac{\Delta t}{\Delta x}$ and $|a|\lambda \leq 1$, the maximum principle is *not* satisfied. In fact, no linear schemes with order of accuracy higher than one can satisfy the maximum principle (the Godunov Theorem).

Therefore, nonlinear schemes, namely schemes which are nonlinear even for the linear PDE (6), have been designed to overcome this difficulty. These include roughly two classes of schemes:

- *TVD schemes*. Most TVD (total variation diminishing) schemes also satisfy strict maximum principle, even in multi-dimensions. TVD schemes can be designed for any formal order of accuracy for solutions in smooth, monotone regions. However, all TVD schemes will degenerate to first-order accuracy at smooth extrema.
- *TVB schemes, ENO schemes, WENO schemes*. The TVB (total variation bounded), ENO (essentially non-oscillatory), and WENO (weighted ENO) schemes do not insist on strict TVD properties; therefore, they do *not* satisfy strict maximum principles, although they can be designed to be arbitrarily high-order accurate for smooth solutions.

A high-order finite volume scheme has the following algorithm flowchart:

(1) Given the cell averages $\{\bar{u}_j^n\}$
(2) reconstruct $u^n(x)$ (piecewise polynomial with cell average $\bar{u}_j^n$)
(3) evolve by, e.g., Runge–Kutta time discretization to get $\{\bar{u}_j^{n+1}\}$
(4) return to (1)

A high-order discontinuous Galerkin scheme has a similar algorithm flowchart:

(1) Given $u^n(x)$ (piecewise polynomial with the cell average $\bar{u}_j^n$)
(2) evolve by, e.g., Runge–Kutta time discretization to get $u^{n+1}(x)$ (with the cell average $\{\bar{u}_j^{n+1}\}$)
(3) return to (1)

Take scalar one-dimensional conservation law as an example. We will call a finite volume or DG scheme bound-preserving, if we have

$$m \le u^{n+1}(x) \le M, \qquad \forall x$$

provided

$$m \le u^{n}(x) \le M, \qquad \forall x.$$

A suitable modification to evaluate the bounds only at certain quadrature points will be given later to facilitate easy implementation.

The flowchart for designing a high-order finite volume or DG scheme which obeys a strict maximum principle is as follows:

1. Start with $u^n(x)$ which is high-order accurate

$$|u(x, t^n) - u^n(x)| \le C\Delta x^p$$

and satisfies

$$m \le u^{n}(x) \le M, \qquad \forall x$$

therefore of course we also have

$$m \le \bar{u}_j^n \le M, \qquad \forall j.$$

2. *Evolve for one time step to get*

$$m \le \bar{u}_j^{n+1} \le M, \qquad \forall j. \tag{7}$$

3. Given (7) above, obtain $u^{n+1}(x)$ (reconstruction or evolution) which
   - satisfies the maximum principle

$$m \le u^{n+1}(x) \le M, \qquad \forall x;$$

   - is high-order accurate

$$|u(x, t^{n+1}) - u^{n+1}(x)| \le C\Delta x^p.$$

There are three major difficulties.

The first difficulty is how to evolve in time for one time step to guarantee the bound for the new cell averages at the next time level (7). This must be achieved by the original high-order DG or finite volume evolution, before using any nonlinear limiters, in order to assure that these new cell averages are both high-order accurate and satisfy the boundedness (7). *This is very difficult to achieve!* Previous works use one of the following two approaches:

- Use exact time evolution. This can guarantee the bound for the new cell averages at the next time level (7). However, it can only be implemented with reasonable cost for linear PDEs, or for scalar nonlinear PDEs in one dimension. This approach

was used in, e.g., Jiang and Tadmor [7], Liu and Osher [10], Sanders [16], Qiu and Shu [13], and Zhang and Shu [28], to obtain TVD schemes or maximum-principle-preserving schemes for linear and nonlinear PDEs in one dimension or for linear PDEs in multi-dimensions, for second-, third-, or higher-order accurate schemes.

- Use simple time evolution such as SSP Runge–Kutta or multi-step methods [5, 17]. However, additional limiting will be needed on $u^n(x)$ which may destroy accuracy near smooth extrema.

In Zhang and Shu [29], a procedure is designed to prove the bound for the new cell averages at the next time level (7), with simple Euler forward or SSP Runge–Kutta or multi-step methods without losing accuracy on the limited $u^n(x)$, as described below.

The evolution of the cell average for a higher-order finite volume or DG scheme satisfies

$$\begin{aligned}\bar{u}_j^{n+1} &= G(\bar{u}_j^n, u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) \\ &= \bar{u}_j^n - \lambda[h(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) - h(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}})],\end{aligned}$$

where we can easily verify

$$G(\uparrow, \uparrow, \downarrow, \downarrow, \uparrow)$$

therefore there is no maximum principle. That is, even if we insist that all five arguments of the function G are in the range $[m, M]$, the cell average at the next time level $\bar{u}_j^{n+1}$ may still be outside this range, regardless of how small one takes the CFL number $\lambda > 0$. The problem is with the two arguments $u^+_{j-\frac{1}{2}}$ and $u^-_{j+\frac{1}{2}}$ which are values at points *inside* the cell I_j.

The polynomial $p_j(x)$ (either reconstructed in a finite volume method or evolved in a DG method) is of degree k, defined on I_j such that $\bar{u}_j^n$ is its cell average on I_j, $u^+_{j-\frac{1}{2}} = p_j(x_{j-\frac{1}{2}})$ and $u^-_{j+\frac{1}{2}} = p_j(x_{j+\frac{1}{2}})$.

We take a Legendre–Gauss–Lobatto quadrature rule which is exact for polynomials of degree k, then

$$\bar{u}_j^n = \sum_{\ell=0}^{m} \omega_\ell p_j(y_\ell)$$

with $y_0 = x_{j-\frac{1}{2}}$, $y_m = x_{j+\frac{1}{2}}$. The scheme for the cell average is then rewritten as

$$\begin{aligned}\bar{u}_j^{n+1} &= \omega_m \left[u^-_{j+\frac{1}{2}} - \frac{\lambda}{\omega_m} \left(h(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) - h(u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}) \right) \right] \\ &\quad + \omega_0 \left[u^+_{j-\frac{1}{2}} - \frac{\lambda}{\omega_0} \left(h(u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}) - h(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}) \right) \right] + \sum_{\ell=1}^{m-1} \omega_\ell p_j(y_\ell) \\ &= \omega_m H_{\lambda/\omega_m}(u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) + \omega_0 H_{\lambda/\omega_0}(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}) + \sum_{\ell=1}^{m-1} \omega_\ell p_j(y_\ell).\end{aligned}$$

Therefore, if

$$m \le p_j(y_\ell) \le M$$

at all Legendre–Gauss–Lobatto quadrature points and a reduced CFL condition

$$\lambda/\omega_m = \lambda/\omega_0 \le \lambda_0,$$

where λ_0 is the bound for the CFL condition of the first-order monotone scheme in (5), is satisfied, then

$$m \le \bar{u}_j^{n+1} \le M.$$

The second difficulty is: given

$$m \le \bar{u}_j^{n+1} \le M, \qquad \forall j$$

how to obtain an *accurate* $u^{n+1}(x)$ (reconstruction or limited DG evolution) which satisfies

$$m \le u^{n+1}(x) \le M, \qquad \forall x.$$

Previous work was mainly for relatively lower-order schemes (second or third-order accurate), and would typically require an evaluation of the extrema of the polynomial solution $u^{n+1}(x)$ before limiting, which, for a piecewise polynomial of higher degree, especially in high-dimension, could be quite costly.

Again in Zhang and Shu [29], a procedure is designed to obtain such $u^{n+1}(x)$ with a very simple scaling limiter, which only requires the evaluation of the unlimited $u^{n+1}(x)$ at certain predetermined quadrature points and does not destroy accuracy. The procedure involves replacing $p_j(x)$ by the limited polynomial $\tilde{p}_j(x)$ defined by

$$\tilde{p}_j(x) = \theta_j(p_j(x) - \bar{u}_j^n) + \bar{u}_j^n$$

where

$$\theta_j = \min\left\{ \left| \frac{M - \bar{u}_j^n}{M_j - \bar{u}_j^n} \right|, \left| \frac{m - \bar{u}_j^n}{m_j - \bar{u}_j^n} \right|, 1 \right\}, \tag{8}$$

with

$$M_j = \max_{x \in S_j} p_j(x), \qquad m_j = \min_{x \in S_j} p_j(x) \tag{9}$$

where S_j is the set of Legendre–Gauss–Lobatto quadrature points of cell I_j.

Clearly, this limiter is just a simple scaling of the original polynomial around its average. The computational cost is minimal, since it involves only the computation of θ_j by (8), which in turn only involves the computation of the local bounds m_j and M_j by (9), via evaluating the unlimited polynomial at the predetermined Legendre–Gauss–Lobatto quadrature points of cell I_j.

The following lemma, guaranteeing the maintenance of accuracy of this simple limiter, is proved in Zhang and Shu [29].

Lemma: Assume $\bar{u}_j^n \in [m, M]$ and $p_j(x)$ is an $O(\Delta x^p)$ approximation, then $\tilde{p}_j(x)$ is also an $O(\Delta x^p)$ approximation.

We have thus obtained a high-order accurate scheme satisfying the following maximum principle: If

$$m \le u^n(x) \le M, \qquad \forall x \in S_j,$$

then

$$m \le u^{n+1}(x) \le M, \qquad \forall x \in S_j.$$

Recall that S_j is the set of Legendre–Gauss–Lobatto quadrature points of cell I_j.

The third difficulty is how to generalize the algorithm and result to 2D (or higher dimensions). Algorithms which would require an evaluation of the extrema of the reconstructed polynomials $u^{n+1}(x, y)$ would not be easy to generalize at all. On the other hand, our algorithm uses only explicit Euler forward or SSP (also called TVD) Runge–Kutta or multi-step time discretizations, and a simple scaling limiter involving just evaluation of the polynomial at certain quadrature points, hence it easily generalizes to 2D or higher dimensions on structured or unstructured meshes, with strict maximum-principle-satisfying property and provable high-order accuracy.

The technique has been generalized to the following situations maintaining uniformly high-order accuracy:

- 2D scalar conservation laws on rectangular or triangular meshes with strict maximum principle, [29, 35].
- 2D incompressible equations in the vorticity-streamfunction formulation (with strict maximum principle for the vorticity), and 2D passive convections in a divergence-free velocity field, i.e.,

  $$\omega_t + (u\omega)_x + (v\omega)_y = 0,$$

  with a given divergence-free velocity field (u, v), again with strict maximum principle, [29, 35].
- One- and multi-dimensional compressible Euler equations maintaining positivity of density and pressure, [30, 35].
- One- and two-dimensional shallow water equations maintaining non-negativity of water height and well-balancedness for problems with dry areas, [22, 23].

- One- and multi-dimensional compressible Euler equations with source terms (geometric, gravity, chemical reaction, radiative cooling) maintaining positivity of density and pressure, [31].
- One- and multi-dimensional compressible Euler equations with gaseous detonations maintaining positivity of density, pressure, and reactant mass fraction, with a new and simplified implementation of the pressure limiter. DG computations are stable without using the TVB limiter, [20].
- A minimum entropy principle satisfying high-order scheme for gas dynamics equations, [32].
- Cosmological hydrodynamical simulation of turbulence in the intergalactic medium (IGM) involving kinetic energy dominated flows, [36].
- Ideal special relativistic hydrodynamics (RHD), [12].
- Positivity-preserving high-order finite difference WENO schemes for compressible Euler equations, [33].
- Simplified version for WENO finite volume schemes without the need to evaluate solutions at quadrature points inside the cell, [34].
- Positivity-preserving for PDEs involving global integral terms including a hierarchical size-structured population model [27], Vlasov–Boltzmann transport equations [2], and correlated random walk with density-dependent turning rates [11].
- Positivity-preserving semi-Lagrangian schemes, [14, 15].
- Positivity-preserving first-order and higher-order Lagrangian schemes for multi-material flows, [1, 18, 19].
- Positivity-preserving DG methods for radiative transfer equations, with iterative procedure for steady states or implicit time discretization for time-dependent equations, [26].

4 Another Approach: Flux Correction

Another approach to achieve bound-preserving schemes is through the traditional flux-correction method, namely modify the numerical flux by

$$\hat{f} = \theta \hat{f}^h + (1 - \theta) \hat{f}^l$$

where $\hat{f}^h$ is the high-order numerical flux and $\hat{f}^l$ is the first-order numerical flux (which does lead to a bound-preserving first-order scheme).

Many traditional TVD or bound-preserving schemes follow this approach. It is relatively easy to design θ to guarantee bound-preserving, but it is relatively more difficult to guarantee accuracy (and often accuracy is lost, especially near smooth extrema).

Recently, this approach has been revived. The limiter in [6] belongs to this class. We mention in particular the work of Xu [25]. This is one of the rare cases that such flux-correction method has been proved to maintain the original high-order

accuracy even near smooth extrema. However, the proof is via explicit and complicated algebraic verifications, thus limiting the scope that it can be applied. See Liang and Xu [9] for scalar conservation law, Xiong et al. [24] for incompressible flows, Christlieb et al. [3] for unstructured mesh, Christlieb et al. [4] for MHD, Jiang et al. [8] for correlated random walk, and Wu and Tang [21] for special relativistic hydrodynamics.

5 Conclusions and Future Work

We have surveyed a general framework to obtain uniformly high-order bound-preserving schemes for multi-dimensional nonlinear conservation laws and other hyperbolic equations including the radiative transfer equations, as well as another approach via flux correction to achieve the same purpose.

It would be interesting to carry out research in the future to design higher-order bound-preserving DG schemes for other types of PDEs and other types of time discretizations.

References

1. J. Cheng, C.-W. Shu, Positivity-preserving Lagrangian scheme for multi-material compressible flow. J. Comput. Phys. **257**, 143–168 (2014)
2. Y. Cheng, I.M. Gamba, J. Proft, Positivity-preserving discontinuous Galerkin schemes for linear Vlasov-Boltzmann transport equations. Math. Comput. **81**, 153–190 (2012)
3. A. Christlieb, L. Liu, Q. Tang, Z. Xu, High order parametrized maximum-principle-preserving and positivity-preserving WENO schemes on unstructured meshes. J. Comput. Phys. **281**, 334–351 (2015)
4. A. Christlieb, L. Liu, Q. Tang, Z. Xu, Positivity-preserving WENO schemes with constrained transport for ideal magnetohydrodynamic equations. SIAM J. Sci. Comput. **37**, A1825–A1845 (2015)
5. S. Gottlieb, D. Ketcheson, C.-W. Shu, *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations* (World Scientific, Singapore, 2011)
6. X.Y. Hu, N.A. Adams, C.-W. Shu, Positivity-preserving method for high-order conservative schemes solving compressible Euler equations. J. Comput. Phys. **242**, 169–180 (2013)
7. G.-S. Jiang, E. Tadmor, Nonoscillatory central schemes for multidimensional hyperbolic conservative laws. SIAM J. Sci. Comput. **19**, 1892–1917 (1998)
8. Y. Jiang, C.-W. Shu, M. Zhang, High order finite difference WENO schemes with positivity-preserving limiter for correlated random walk with density-dependent turning rates. Math. Models Methods Appl. Sci. (M^3AS) **25**, 1553–1588 (2015)
9. C. Liang, Z. Xu, Parametrized maximum-principle-preserving flux limiters for high order schemes solving multi-dimensional scalar hyperbolic conservation laws. J. Sci. Comput. **58**, 41–60 (2014)
10. X.-D. Liu, S. Osher, Non-oscillatory high order accurate self similar maximum principle satisfying shock capturing schemes. SIAM J. Numer. Anal. **33**, 760–779 (1996)

11. J. Lu, C.-W. Shu, M. Zhang, Stability analysis and a priori error estimate of explicit Runge-Kutta discontinuous Galerkin methods for correlated random walk with density-dependent turning rates. Sci. China Math. **56**, 2645–2676 (2013)
12. T. Qin, C.-W. Shu, Y. Yang, Bound-preserving discontinuous Galerkin methods for relativistic hydrodynamics. J. Comput. Phys. **315**, 323–347 (2016)
13. J.-M. Qiu, C.-W. Shu, Convergence of Godunov-type schemes for scalar conservation laws under large time steps. SIAM J. Numer. Anal. **46**, 2211–2237 (2008)
14. J.-M. Qiu, C.-W. Shu, Positivity preserving semi-Lagrangian discontinuous Galerkin formulation: theoretical analysis and application to the Vlasov-Poisson system. J. Comput. Phys. **230**, 8386–8409 (2011)
15. J.A. Rossmanith, D.C. Seal, A positivity-preserving high-order semi-Lagrangian discontinuous Galerkin scheme for the Vlasov-Poisson equations. J. Comput. Phys. **230**, 6203–6232 (2011)
16. R. Sanders, A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation law. Math. Comput. **51**, 535–558 (1988)
17. C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)
18. F. Vilar, C.-W. Shu, P.-H. Maire, Positivity-preserving cell-centered Lagrangian schemes for multi-material compressible flows: from first-order to high-orders. Part I: the one-dimensional case. J. Comput. Phys. **312**, 385–415 (2016)
19. F. Vilar, C.-W. Shu, P.-H. Maire, Positivity-preserving cell-centered Lagrangian schemes for multi-material compressible flows: from first-order to high-orders. Part II: the two-dimensional case. J. Comput. Phys. **312**, 416–442 (2016)
20. C. Wang, X. Zhang, C.-W. Shu, J. Ning, Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations. J. Comput. Phys. **231**, 653–665 (2012)
21. K. Wu, H. Tang, High-order accurate physical-constraints-preserving finite difference WENO schemes for special relativistic hydrodynamics. J. Comput. Phys. **298**, 539–564 (2015)
22. Y. Xing, C.-W. Shu, High-order finite volume WENO schemes for the shallow water equations with dry states. Adv. Water Resour. **34**, 1026–1038 (2011)
23. Y. Xing, X. Zhang, C.-W. Shu, Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations. Adv. Water Resour. **33**, 1476–1493 (2010)
24. T. Xiong, J.-M. Qiu, Z. Xu, A parametrized maximum principle preserving flux limiter for finite difference RK-WENO schemes with applications in incompressible flows. J. Comput. Phys. **252**, 310–331 (2013)
25. Z. Xu, Parametrized maximum principle preserving flux limiters for high order scheme solving hyperbolic conservation laws: one-dimensional scalar problem. Math. Comput. **83**, 2213–2238 (2014)
26. D. Yuan, J. Cheng, C.-W. Shu, High order positivity-preserving discontinuous Galerkin methods for radiative transfer equations. SIAM J. Sci. Comput. **38**, A2987–A3019 (2016)
27. R. Zhang, M. Zhang, C.-W. Shu, High order positivity-preserving finite volume WENO schemes for a hierarchical size-structured population model. J. Comput. Appl. Math. **236**, 937–949 (2011)
28. X. Zhang, C.-W. Shu, A genuinely high order total variation diminishing scheme for one-dimensional scalar conservation laws. SIAM J. Numer. Anal. **48**, 772–795 (2010)
29. X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws. J. Comput. Phys. **229**, 3091–3120 (2010)
30. X. Zhang, C.-W. Shu, On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. J. Comput. Phys. **229**, 8918–8934 (2010)
31. X. Zhang, C.-W. Shu, Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms. J. Comput. Phys. **230**, 1238–1248 (2011)
32. X. Zhang, C.-W. Shu, A minimum entropy principle of high order schemes for gas dynamics equations. Numer. Math. **121**, 545–563 (2012)
33. X. Zhang, C.-W. Shu, Positivity-preserving high order finite difference WENO schemes for compressible Euler equations. J. Comput. Phys. **231**, 2245–2258 (2012)

34. X. Zhang, C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. Proc. R. Soc. A **467**, 2752–2776 (2011)
35. X. Zhang, Y. Xia, C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes. J. Sci. Comput. **50**, 29–62 (2012)
36. W. Zhu, L.-L. Feng, Y. Xia, C.-W. Shu, Q. Gu, L.-Z. Fang, Turbulence in the intergalactic medium: solenoidal and dilatational motions and the impact of numerical viscosity. Astrophys. J. **777**, 48 (2013)

Comparison of Shallow Water Models for Rapid Channel Flows

Stefanie Elgeti, Markus Frings, Anne Küsters, Sebastian Noelle and Aleksey Sikstel

Abstract To model shallow free surface flows, the Saint-Venant Equations (SVE) are a convenient simplification of the incompressible Navier–Stokes Equations (NSE). In the present study, we compare the two models for one-dimensional channel flow over a hump (cf. Behr (XNS simulation program, 2016 [5]), Küsters (Comparison of a Navier–Stokes and a shallow water model using the example of flow over a semi-circular bump, 2013 [8]), Noelle et al. (J Comput Phys 226(1):29–58, 2007 [10]), Sikstel (Comparison of hydrostatic and non-hydrostatic shallow water models, 2016 [13])). Our numerical experiments show that the SVE fail for some rather standard transcritical flows, where the two models compute different water heights ahead of and different shock speeds behind the hump. Using numerical computations as well as a formal Cauchy–Kowalevski argument, we give a qualitative explanation of the shortcoming of the SVE. In addition, we examine a recently developed non-hydrostatic shallow water model Sainte-Marie et al. (Discrete and Cont Dyn Syst Ser B 20(4):361–388, 2014 [12]) which proposes to produce physically more realistic results.

S. Elgeti · M. Frings
CATS, RWTH Aachen University, Schinkelstraße 2, 52062 Aachen, Germany
e-mail: elgeti@cats.rwth-aachen.de

M. Frings
e-mail: frings@cats.rwth-aachen.de

A. Küsters
JSC, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, 52428 Jülich, Germany
e-mail: a.kuesters@fz-juelich.de

S. Noelle (✉) · A. Sikstel
IGPM, RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany
e-mail: noelle@igpm.rwth-aachen.de

A. Sikstel
e-mail: sikstel@igpm.rwth-aachen.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_45

Keywords Navier–Stokes equations · Shallow Water Equations
Flow over a Weir · Non-hydrostatic pressure · Inflow–Outflow boundary conditions

AMS Subject Classification: 76B15 · 35L65

1 Shallow Water Models

The modeling of shallow water flows plays an important role in geophysical research. The basic equations of motion are the incompressible NSE [6]

$$\rho(\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \mathbf{g}) + \nabla p - \mu \triangle \mathbf{u} = 0 \text{ in } \Omega(t), \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0 \text{ in } \Omega(t) \tag{2}$$

where $\mathbf{x} = (x, z)$, $\mathbf{u} = (u, w)^T$, and p are the space variables, velocities, and thermodynamic pressure, respectively. Furthermore, μ denotes the viscosity coefficient, $\mathbf{g} := (0, g)^T$ the gravitational acceleration and ρ the constant density. The solution of the NSE is defined in a domain $\Omega(t) := \{(x, z) : b(x) \leq z \leq \eta(x, t),\ x_A \leq x \leq x_B\}$ with a free surface $\eta(x, t)$ and a solid, time-independent bottom b (cf. Fig. 1). Moreover, we introduce the water depth $H(x, t) := \eta(x, t) - b(x)$. The formulation of the boundary conditions (BC) is postponed to the following section.

In case of shallow flows, the NSE are often simplified under certain conditions to a nonlinear system of hyperbolic equations. For this, we start with the NSE in their dimensionless form,

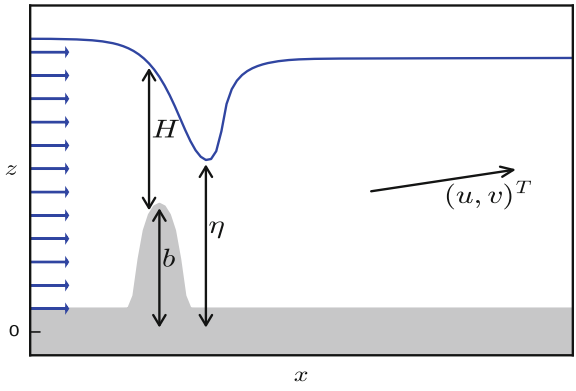


Fig. 1 Notation for the numerical experiments

Table 1 Dimensionless parameters. x_{ref}—reference length, z_{ref}—reference height, t_{ref}—reference time, and $u_{ref} = \frac{x_{ref}}{t_{ref}}$—reference velocity

1	$u_{ref}\frac{t_{ref}}{x_{ref}}$	$\frac{\text{flow speed}}{\text{grid speed}}$	
δ	$\frac{z_{ref}}{x_{ref}}$	$\frac{\text{vertical length scale}}{\text{horizontal length scale}}$	shallowness
G	$\frac{gt_{ref}}{u_{ref}}$	$\frac{\text{gravitational acceleration}}{\text{grid acceleration}}$	
Fr	$\frac{u_{ref}}{\sqrt{gz_{ref}}}$	$\frac{\text{speed}}{\sqrt{\text{length}}}$	reference Froude Number
Re	$\frac{u_{ref}x_{ref}\rho_{ref}}{\mu}$	$\frac{\text{inertial forces}}{\text{viscous forces}}$	reference Reynolds Number

$$
\begin{aligned}
0 = \rho(\partial_t u_i + \nabla\cdot(\mathbf{u}\,u_i) + \frac{G}{\delta}[i=2]) \\
+ \frac{1}{\mathrm{Fr}^2}\left\{\begin{array}{ll} 1, & \text{for } i=1 \\ 1/\delta^2 & \text{for } i=2 \end{array}\right\}\nabla p \\
- \frac{1}{\mathrm{Re}}\sum_{j=1}^{2}\partial_j^2 u_i \left\{\begin{array}{ll} 1, & \text{for } j=1 \\ 1/\delta & \text{for } j=2 \end{array}\right\} \text{ for } i\in\{1,2\}, x \text{ in } \Omega(t) \qquad (3)\\
0 = \nabla\cdot\mathbf{u} \text{ in } \Omega(t). \qquad (4)
\end{aligned}
$$

where $[i=2]$ denotes the Kronecker's delta symbol to avoid confusion with the dimensionless parameter δ. The definitions of the dimensionless parameters in Eqs. (3)–(4) are listed in Table 1. For the sake of simplicity, we omit any units when explicitly defining values of parameters, constants, or material properties in the next sections. Rewriting the vertical component of the Momentum Equation (3) yields an equation for the change of the pressure in the vertical direction

$$
\frac{\partial}{\partial z}p = -\delta^2\mathrm{Fr}^2\rho\left(\frac{G}{\delta} + \frac{\partial}{\partial t}w + \nabla\cdot(\mathbf{u}\,w) - \frac{1}{\rho\mathrm{Re}}\left(\frac{\partial^2}{\partial x^2}w + \frac{1}{\delta^2}\frac{\partial^2}{\partial z^2}w\right)\right). \qquad (5)
$$

For many shallow flows, the vertical acceleration is small, while the Reynolds number is large, so

$$
R := \frac{\partial}{\partial t}w + \nabla\cdot(\mathbf{u}\,w) - \frac{1}{\rho\mathrm{Re}}\left(\frac{\partial^2}{\partial x^2}w + \frac{1}{\delta^2}\frac{\partial^2}{\partial z^2}w\right) \ll \frac{G}{\delta} \qquad (6)
$$

This motivates the assumption that R vanishes and leads to.

Assumption 1 *(Hydrostatic Pressure)*

$$
\frac{\partial}{\partial z}p = -\delta^2\mathrm{Fr}^2 G\rho. \qquad (7)
$$

Since the dimensionless formulation was only used to motivate the hydrostatic pressure assumption, we now return to physical variables. Integrating (7) from $b(x)$ to

$\eta(x)$ along the vertical axis and assuming zero atmospheric pressure, we obtain

$$p^h := \rho g(\eta - b), \tag{8}$$

so the hydrostatic pressure p^h is due to the weight of the fluid on top of it.

Definition 1. We define the depth integral, the depth average, and the deviation of an integrable function $\varphi(x, z, t)$ by

$$\bar{\varphi}(x,t) := \int_b^\eta \varphi(x,z,t)\, dz, \tag{9}$$

$$\langle\varphi\rangle(x,t) := \frac{1}{H}\int_b^\eta \varphi(x,z,t)\, dz, \tag{10}$$

$$\tilde{\varphi}(x,z,t) := \varphi(x,z,t) - \langle\varphi\rangle(x,t). \tag{11}$$

It is crucial to note that

$$\langle\varphi^2\rangle = \langle\varphi\rangle^2 + \langle\tilde{\varphi}^2\rangle, \tag{12}$$

and hence depth averaging the advective flux in the momentum equation yields the term $\frac{\partial}{\partial x}\left(H(\langle u\rangle^2 + \langle\tilde{u}^2\rangle)\right)$. To avoid introducing $\langle\tilde{u}^2\rangle$ as an unknown in a depth-averaged shallow water equation, it is common to refer to the following assumption, which has been used by Levermore [9] in the context of kinetic equations:

Assumption 2 *(Zero Moment)* The horizontal velocity u is independent of the vertical direction z, i.e.

$$\tilde{u} \equiv 0. \tag{13}$$

In channel flows, this assumption is only valid approximately, if at all, and we attempt to assess its influence in Sect. 3.1 below.

Depth-averaging Eqs. (3)–(4), neglecting the viscous forces and applying Assumptions 1 and 2 yields the one-dimensional Saint-Venant Equations (or Shallow Water Equations) [4]

$$\frac{\partial}{\partial t}H + \frac{\partial}{\partial x}(H\langle u\rangle) = 0, \tag{14}$$

$$\frac{\partial}{\partial t}(H\langle u\rangle) + \frac{\partial}{\partial x}\left(H\langle u\rangle^2 + \frac{1}{2}gH^2\right) = -gH\frac{\partial}{\partial x}b. \tag{15}$$

We would like to mention two generalizations of the SVE: the multilayer model of Audusse [2] does not use Assumption 2 on the velocity profile. The non-hydrostatic depth average Euler equations (AVE) of Sainte-Marie et al. [12] does not assume hydrostatic pressure (Assumption 1), but introduces a vertical velocity component:

$$\frac{\partial}{\partial t}H + \frac{\partial}{\partial x}(H\langle u\rangle) = 0, \tag{16}$$

$$\frac{\partial}{\partial t}(H\langle u\rangle) + \frac{\partial}{\partial x}(H\langle u\rangle^2 + \frac{1}{2}gH^2 + H\langle p_{nh}\rangle), = -(gH + 2\langle p_{nh}\rangle)\frac{\partial}{\partial x}b, \tag{17}$$

$$\frac{\partial}{\partial t}(H\langle w\rangle) + \frac{\partial}{\partial x}(H\langle w\rangle\langle u\rangle) = 2\langle p_{nh}\rangle, \tag{18}$$

$$\frac{\partial}{\partial x}(H\langle u\rangle) - \langle u\rangle\frac{\partial}{\partial x}(H + 2b) + 2\langle w\rangle = 0. \tag{19}$$

We include the AVE in the numerical comparison below.

2 Numerical Experiments

In this section, we describe a numerical experiment comparing hydrostatic and non-hydrostatic models for a one-dimensional channel flow. For notation refer to Fig. 1.

2.1 Initial and Boundary Conditions

We consider one-dimensional channel flow over a bump. The inflow boundary is positioned at $x_L := -7.0$ and an outflow boundary at $x_R := 15.0$. We set $g = 1$, $\rho = 1$, and $\mu = 5 \cdot 10^{-7}$ (almost inviscid flow). The bump is a semi-circle

$$b(x) = \begin{cases} 0 & \text{if } |x| < 0.45 + \delta \\ \sqrt{0.45^2 - x^2} & \text{if } |x| < 0.45 - \delta \end{cases} \tag{20}$$

whose corners are connected twice continuously differentiable by a fifth-degree polynomial. The intention—alas not proved so far—is that this yields continuous dependence of the solution on the data.

The initial conditions for the NSE are constructed as a solution of the corresponding steady-state equations, which is defined by an inflow discharge $Hu = 0.5$ to the left, and and water height $H = 1$ to the right. The initial datum for the water level is set to $\eta = 1.0$ everywhere, the horizontal velocity $u = 0.5$ and the vertical velocity $w = 0.0$ except for the surroundings of the bump. These data are depth-averaged in order to obtain the initial conditions for the SVE and AVE.

To define boundary conditions for the NSE, we divide $\Gamma := \partial\Omega$ into four parts, $\Gamma = \Gamma_{in} \cup \Gamma_{surf} \cup \Gamma_{out} \cup \Gamma_{bot}$ which stand for inflow boundary, the free surface, outflow boundary, and the bottom topography, respectively. We impose a Dirichlet condition $\mathbf{u} = (0.5, 0)^{tr}$ on the inflow boundary Γ_{in}. At the outflow boundary, we define a free boundary condition in terms of the unknown pressure p and velocity field $\mathbf{u}$ following [11]. In a discrete setting, this gives a well-posed problem with

a Neumann BC for the highest derivative (cf. [7]). We apply kinematic boundary conditions at the free surface and the bottom and neglect wind forces and atmospheric pressure. Finally, we set the total tension tensor to zero at the surface and assume perfect slip at the bottom.

For the SVE, we have only two boundary points, x_L and x_R. At x_L, we impose $Hu(x_L, t) := (Hu)_0(t)$ (see Sect. 2.2). At the outflow, we consider two types of BCs. Either we prescribe a Neumann BC $(H, Hu)_x(x_R) = 0$ or a far-field BC by setting $H(x_R)$ to a constant value.

The AVE include additional unknowns, namely the vertical momentum Hv and the non-hydrostatic pressure p_{nh}. For the non-hydrostatic pressure, we refer to [1]. The vertical momentum at the inflow is set to zero, in accordance with the NSE. The other variables are treated exactly in the same way as for the SVE.

2.2 Numerical Solvers

For the NSE, we use the XNS finite element solver designed and developed at the CATS institute, [5]. The SVE are solved using the first-order hydrostatic reconstruction scheme [3] with an HLL numerical flux. For the AVE, this scheme is augmented by a projection–correction step for the non-hydrostatic pressure p_{nh}, see [1].

We turn to the numerical boundary conditions. For the NSE, a convenient way to realize the Neumann boundary at the outflow is to keep the values p and u from the previous time step (cf. [8]). Now, we discuss the inflow BC for the SVE. We denote the ghost cell to the left of x_L, the values therein by subscript 0 and the interior values by subscript 1. We assign $(Hu)_0$ as the depth-averaged horizontal velocity times the water height of the NSE solution at the beginning of each time step. To filter out small oscillations in the Navier–Stokes data near the inlet, we average the momentum of the NSE solution in the interval $x \in [-6.5, 5]$ at the left end of the domain. The interior values H_1 and $(Hu)_1$ are given by the numerical solution. It remains to calculate H_0. We require that it lies on an outgoing characteristic starting at $(H_1, (HU)_1)$, so

$$H_0 = \frac{(Hu)_0 + H_1\sqrt{H_1}}{u_1 + \sqrt{H_1}}. \tag{21}$$

At the outflow (with ghost cell $(N + 1)$ and interior cell N), we have two boundary conditions for the SVE. First, we have the Neumann BC extrapolated to the ghost cell, i.e., $H_{N+1} = H_N$ and $(Hu)_{N+1} = (Hu)_N$. Alternatively, the far-field BC sets the water height in the SVE to be equal to the average value of the water height of the NSE solution. Using the characteristic conditions again, one obtains

$$(Hu)_{N+1} = H_{N+1}\left(\frac{(Hu)_N}{H_N} - \sqrt{0.5\left(\frac{H_N}{H_{N+1}} - \frac{H_{N+1}}{H_N}\right)(H_N - H_{N+1})}\right). \tag{22}$$

The boundary conditions for the AVE are implemented exactly in the same way as for the SVE. For the term p_{nh}, we again refer to [1].

2.3 *Numerical Results*

Running the simulations described above for 190 s gives the results shown in Fig. 2. For the rest of this note, red circles in the graphics will denote the NSE solution, blue crosses the SVE solution, and green squares the AVE solutions.

3 Discussion

In this section, we discuss two major differences between the NSE computations on one hand, and the depth-averaged SVE and AVE computations on the other hand:

1. The water level of the SVE and AVE computations is about 17% higher than for the NSE computations ahead of the bump (see Fig. 2 upper left).
2. The AVE and in particular the SVE shocks move considerably faster than the NSE shock (see Fig. 2 upper right and lower left).

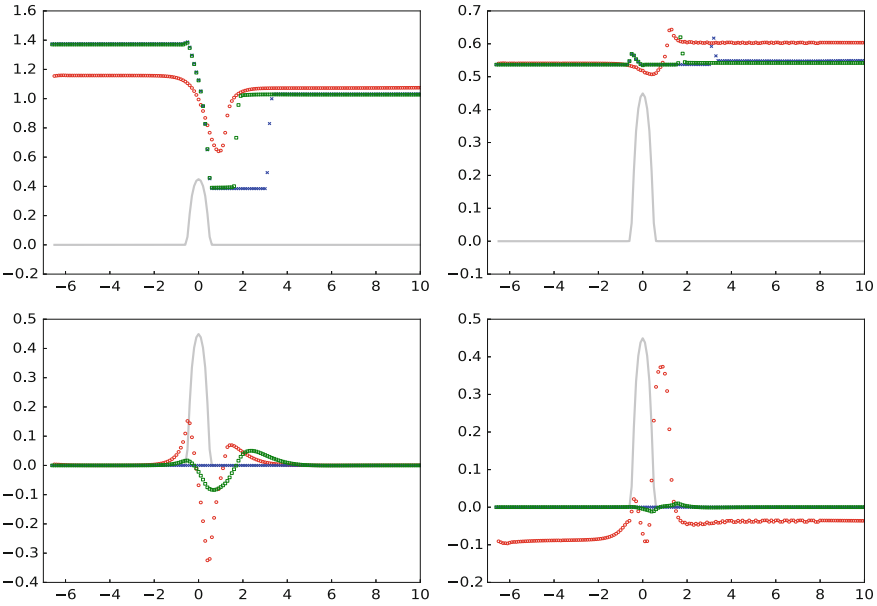


Fig. 2 Differences of NSE, SVE, and AVE regarding free surface (upper left), horizontal momentum (upper right), vertical momentum (lower left), and non-hydrostatic pressure (lower right)

We would like to relate these errors to the modeling Assumptions 1 and 2 as well as to the numerical boundary coditions.

3.1 Water Level

The water levels produced by the SVE and the AVE to the left of the bump are almost indistinguishable and are both about 17% larger than that of the NSE. In Fig. 3, we show the rise of the water height (left) and the slow down of the depth-averaged velocity (right) at the beginning of the experiment.

We begin by showing that this is due to a left-going wave which is reflected from the bump. For this, we linearize the SVE around the inflow state $\hat{H} = 1.0$ and $\hat{u} = 0.5$ (with eigenvalues $\hat{\lambda}_1 = -0.5$ and $\hat{\lambda}_2 = 1.5$) and decouple the system by projecting the variables onto the left eigenvectors. Thus, the projected variables are $v_1 := H(0.5 - u^S)$ and $v_2 = H(1.5 + u^S)$. Figure 4 clearly shows that there is a left-going wave starting at the foot of the bump.

Now, the z-independent profile of the horizontal velocity postulated in Assumption 2 initiates the unphysical reflection for the SVE. For this argument, we compare the evolution of the SVE solution $(H^S, \langle u^S \rangle)$ and the solution $(H^N, \langle u^N \rangle)$ of the depth-averaged NSE using a Cauchy–Kowalevski type argument. Neglecting the viscous forces, the depth-averaged NSE read

$$\frac{\partial}{\partial t} H + \frac{\partial}{\partial x} (H \langle u \rangle) = 0, \tag{23}$$

$$\frac{\partial}{\partial t} (H \langle u \rangle) + \frac{\partial}{\partial x} \left(H \left(\langle u \rangle^2 + \langle \tilde{u}^2 \rangle \right) + \frac{1}{2} g H^2 + H \langle \tilde{p}^{nh} \rangle \right) = -g H \frac{\partial}{\partial x} b. \tag{24}$$

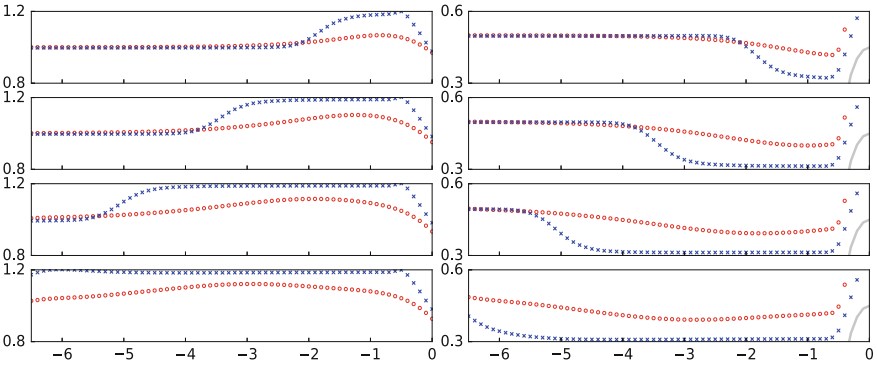


Fig. 3 Water height in the SVE rises at the beginning of the experiment—η (left) and $\langle u \rangle$ (right) at times $t = (2.5, 5, 7.5, 10)$ (from top to bottom)

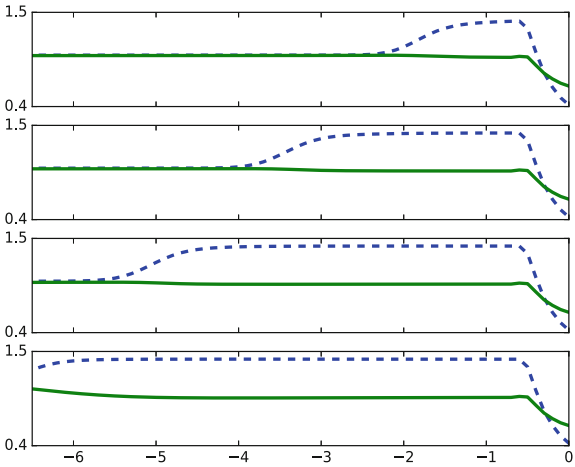


Fig. 4 v_1 (dashed line) and v_2 (solid line) for the SVE at times $t = (2.5, 5, 7.5, 10)$ from top to bottom

Due to the assignment of the SVE initial data, $\langle u^S \rangle = \langle u^N \rangle$ at time zero. Substracting (23)–(24) from (14)–(15), we obtain

$$\frac{\partial}{\partial t}\left(H^S - H^N\right) = 0 \tag{25}$$

$$\frac{\partial}{\partial t}(H^S\langle u^S\rangle - H^N\langle u^N\rangle = -\frac{\partial}{\partial x}\left(H\langle\tilde{u}^2\rangle + H\langle p^{nh}\rangle\right). \tag{26}$$

For the amplitude of the left-going wave, we get

$$\frac{\partial}{\partial t}\left(v_1^S - v_1^N\right) = \frac{\partial}{\partial x}\left(H\langle\tilde{u}^2\rangle + H\langle p^{nh}\rangle\right). \tag{27}$$

In Fig. 5, we display $\tau := H\langle\tilde{u}^2\rangle$. Clearly, $\frac{\partial}{\partial x}\tau$ is positive ahead of the bump, which corresponds to the onset of the left-going wave in Fig. 4. Similarly, the non-hydrostatic pressure in the NSE (see lower-right plot in Fig. 2) also increases to the left of the bump. This indicates that both simplifications made in Assumptions 1 and 2 contribute to the onset of the reflected wave to the left of the bump. As can be seen from Fig. 6, the change of the BC does not have an influence on the water height to the left of the bump.

Note also that in our implementation, the solutions of non-hydrostatic depth-averaged AVE model are closer to the SVE than the NSE.

3.2 Shock Position

Let us now discuss the flow at the rear side and behind the bump. At the top of the bump, the flow changes from sub-critical to supercritical, and subsequently

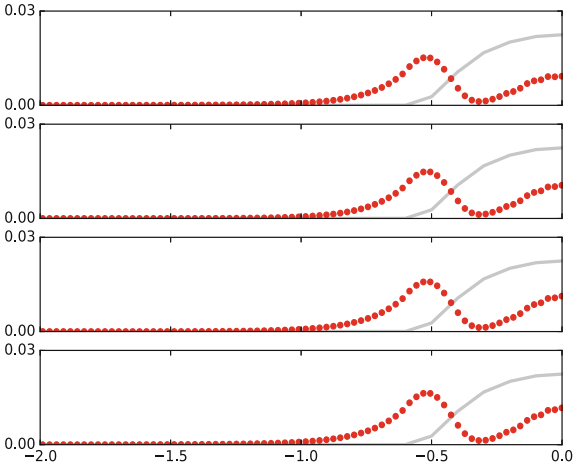


Fig. 5 Second spatial derivative of τ is almost constant in time; at times $t = (2.5, 5, 7.5, 10)$ from top to bottom

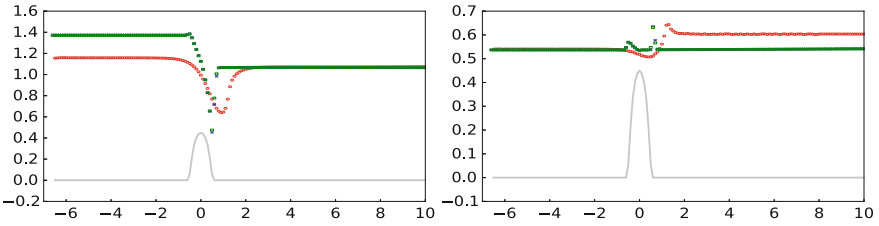


Fig. 6 Water height and horizontal momentum after applying the far-field BC $H(x_{N+1}) = 1.0$

accelerates as a waterfall, which ends in a hydraulic jump. This is shown at time $t = 190$ in the upper left plot in Fig. 2. For the NSE, the hydraulic jump is a stationary viscous profile, while the SVE and AVE yield shocks traveling to the right. All three models give distinctly different solutions with different wave speeds.

In [8, 13], it was argued that the pressure difference between the models due to Assumption 1 changes the left momentum flux in the Rankine–Hugoniot condition, and hence, the shock speeds. Here, we argue that also the boundary condition at the outflow plays an important role. For this, we replace the Neumann BC by the farfield BC $H(x_{N+1}) = 1.0$. This corresponds closely to an implementation detail of the moving grid NSE-solver, which fixes the top grid point at the outflow to height $z = 1$. The results are presented in Fig. 6. The SVE and AVE shocks now coincide, are stationary, and are located at the right corner of the bump. In particular, note that the non-hydrostatic pressure, which is different in the three models, does not seem to play a role in the shock position.

4 Conclusion

We have studied a supercritical channel flow over a bump where the incompressible, free-surface Navier–Stokes equations give distinctly different solutions from two depth-averaged shallow water models. An unphysical reflection off the bump for the depth-averaged models is caused equally by the simplified velocity profile, which is constant in vertical direction, and by the lack of non-hydrostatic pressure. Contrary to this, a wrong shock speed is mainly due to the outflow boundary conditions.

It would be interesting to see if a variant of the non-hydrostatic depth-averaged model of Sainte-Marie et al. [12], or another implementation of the model, would produce results which are closer to the NSE. It would also be interesting to see if Audusse multi-layer model [2] reduces the reflection at the bump.

Overall, our example shows that depth-averaged models, which are very common in the hyperbolic community, should be used with some caution.

Acknowledgements The authors would like to thank Emmanuel Audusse, Jacques Sainte-Marie and Marek Behr for sharing their insights, and Henning Sauerland for help with the NSE solver.

References

1. N. Aïssiouene, M.O. Bristeau, E. Godlewski, J. Sainte-Marie, A robust and stable numerical scheme for a depth-averaged Euler system. arXiv preprint arXiv:1506.03316 (2015)
2. E. Audusse, A multilayer Saint-Venant model: derivation and numerical validation. Discret. Cont. Dyn. Syst. B **5**(2), 189–214 (2005)
3. E. Audusse, F. Bouchut, M.O. Bristeau, R. Klein, B. Perthame, A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. Sci. Comput. **25**(6), 2050–2065 (2004)
4. A.J.C. Barré de Saint-Venant, Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leurs lits. Comptes Rendus des Séances de l'Académie des Sci. **73**, 237–240 (1871)
5. M. Behr, XNS Simulation Program. Chair for Computational Analysis of Technical Systems. RWTH Aachen University (2016). http://www.cats.rwth-aachen.de/about/software/simulation/xns
6. P. Constantin, C. Foias, *Navier-Stokes Equations* (University of Chicago Press, 1988)
7. D.F. Griffiths, The 'No Boundary Condition' outflow boundary condition. Int. J. Numer. Methods Fluids **24**(4), 393–411 (1997)
8. A. Küsters, Comparison of a Navier-Stokes and a shallow-water model using the example of flow over a semi-circular bump. Master's thesis, RWTH Aachen, 2013
9. C.D. Levermore, Entropy-based moment closures for kinetic equations. Trans. Theory Stat. Phys. **26**(4–5), 591–606 (1997)
10. S. Noelle, Y. Xing, C.W. Shu, High-order well-balanced finite volume WENO schemes for shallow water equation with moving water. J. Comput. Phys. **226**(1), 29–58 (2007)
11. T.C. Papanastasiou, N. Malamataris, K.R. Ellwood, A new outflow boundary condition. Int. J. Numer. Methods Fluids **14**, 587–608 (1992)

12. J. Sainte-Marie, M.O. Bristeau, A. Mangeney, N. Seguin et al., An energy-consistent depth-averaged Euler system: derivation and properties. Discrete and Cont. Dyn. Syst. Ser. B **20**(4), 361–388 (2014)
13. A. Sikstel, Comparison of hydrostatic and non-hydrostatic shallow water models. Master's thesis, RWTH Aachen, 2016
14. R. Temam, On the Euler equations of incompressible perfect fluids. Séminaire Équations aux dérivées partielles (Polytechnique) (1974), pp. 1–14

On Stability and Conservation Properties of (s)EPIRK Integrators in the Context of Discretized PDEs

Philipp Birken, Andreas Meister, Sigrun Ortleb and Veronika Straub

Abstract Exponential integrators are becoming increasingly popular for stiff problems of high dimension due to their attractive property of solving the linear part of the system exactly and hence being A-stable. In practice, however, exponential integrators are implemented using approximation techniques to matrix-vector products involving functions of the matrix exponential (the so-called φ-functions) to make them efficient and competitive to other state-of-the-art schemes. We will examine linear stability and provide a Courant–Friedrichs–Lewy (CFL) condition of special classes of exponential integrator schemes called EPIRK and sEPIRK and demonstrate their dependence on the parameters of the embedded approximation technique. Furthermore, a conservation property of the EPIRK schemes is proven.

Keywords Exponential integrators · CFL condition · A-stability · Conservation

2010 MSC: 65L20 · 65L07 · 35L65 · 65F60

1 Introduction

Many simulations require efficient time integration schemes for large and stiff systems of ODEs resulting from the method of lines for PDEs such as the compressible

P. Birken
Numerical Analysis, Lund University, Box 118, 22100 Lund, Sweden
e-mail: philipp.birken@na.lu.se

A. Meister · S. Ortleb · V. Straub (✉)
Department of Mathematics, University of Kassel, Heinrich-Plett-Str. 40, 34132Kassel, Germany
e-mail: meister@mathematik.uni-kassel.de

S. Ortleb
e-mail: ortleb@mathematik.uni-kassel.de

V. Straub
e-mail: vstraub@mathematik.uni-kassel.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_46

Navier–Stokes equations on complex grids. In this field, implicit methods often appear to be more efficient than explicit schemes due to the bounded stability region of explicit schemes. Exponential integrators also provide good stability properties like A-stability and recently gained increasing interest due to efficient approximation techniques for the evaluation of the matrix exponential and functions of it.

We will focus on the classes of exponential integrators called EPIRK [1] and sEPIRK [2]. They were shown to be efficient for sufficiently stiff problems through the application of φ-functions to vectors by sophisticated approximation strategies utilizing Krylov projections [3]. Furthermore, they offer high precision and good stability properties like A- and L-stability in theory, i.e., evaluating the φ-functions analytically. In practice, however, we will show in Sect. 2.1 that the chosen dimension m of the Krylov subspace as well as other parameters influence the size of the stability region and A-stability is not given anymore.

Given the framework of discretized PDEs, the Courant–Friedrichs–Lewy (CFL) condition has to be satisfied. We will see that this is always the case for exact φ-evaluations but not guaranteed in combination with the approximation techniques anymore. A CFL condition including the dependence on the parameters of the scheme and the approximation strategy will be presented and verified by a numerical experiment in Sect. 2.2.

Conservativity of the numerical method is another desirable property when dealing with conservation and balance laws such as the Navier–Stokes equations. It guarantees that the scheme does not produce or lose mass, momentum, and energy in agreement with the exact solution. The finite volume or discontinuous Galerkin approach typically apply conservative numerical flux functions for this purpose. This qualitative property should also be maintained by the time discretization method. In Sect. 2.3, we prove that the EPIRK schemes are conservative.

All in all, we will see that (s)EPIRK schemes are well suited for application to discretized PDEs, since the stability properties can be adaptively improved by adjusting the parameters within the approximation strategy. In [4], the interested reader may find the application of an EPIRK scheme to a discretized viscous Burgers' equation resulting in a geometry-induced stiff problem in an implicit–explicit time integration setting.

2 (s)EPIRK Schemes

For numerically solving an initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{U}(t) = \mathrm{F}(\mathrm{U}(t)),\ \mathrm{U}(0) = \mathrm{U}_0, \quad \mathrm{U} : \mathbb{R}_0^+ \to \mathbb{R}^N,\ \ \mathrm{F} : \mathbb{R}^N \to \mathbb{R}^N, \tag{1}$$

an explicit EPIRK scheme with s stages and a set of scheme parameters b_i, g_{ij}, a_{km} for $i, j \in \{1, 2, \ldots, s\}$, $k, m \in \{1, 2, \ldots, s-1\}$ can be written as

$$\mathrm{U}_{n+1} = \mathrm{U}_n + b_1\psi_{s1}(g_{s1}h\mathbf{A}_n)h\mathrm{F}_n + h\sum_{k=2}^{s} b_k\psi_{sk}(g_{sk}h\mathbf{A}_n)\Delta^{(k-1)}\mathrm{R}_n(\mathrm{U}_n)$$

$$\mathrm{Y}_i = \mathrm{y}_n + a_{i1}\psi_{i1}(g_{i1}h\mathbf{A}_n)h\mathrm{F}_n + h\sum_{j=2}^{i} a_{ij}\psi_{ij}(g_{ij}h\mathbf{A}_n)\Delta^{(j-1)}\mathrm{R}_n(\mathrm{U}_n),$$

$i = 1, \ldots, s-1$, see [1]. The time step size is denoted by h, $\mathrm{U}_n \approx \mathrm{U}(t_n)$, $\mathrm{F}_n = \mathrm{F}(\mathrm{U}_n)$, $\mathbf{A}_n = \mathrm{F}'(\mathrm{U}_n)$, $\Delta^j\mathrm{R}_n(\mathrm{U}_n)$ is the jth forward difference through the nodes $\mathrm{U}_n, \mathrm{Y}_1, \ldots, \mathrm{Y}_j$ and

$$\mathrm{R}_n(\mathrm{U}_n) = \mathrm{F}(\mathrm{U}(t)) - \mathrm{F}_n - \mathbf{A}_n(\mathrm{U}(t) - \mathrm{U}_n)$$

denotes the nonlinear remainder. The ψ-functions are defined by

$$\psi_{ik}(z) = \sum_{j=1}^{s} p_{ikj}\varphi_j(z), \quad i,k = 1,\ldots,s, \quad \text{with} \quad \varphi_j(z) = \sum_{\ell=0}^{\infty}\frac{1}{(\ell+j)!}z^\ell, \quad j \in \mathbb{N}_0, \tag{2}$$

using the additional parameters p_{ikj}, $i, j, k = 1, \ldots, s$. For the special case of a split right-hand side F

$$\mathrm{F}(\mathrm{U}) = \mathbf{L}\mathrm{U}(t) + \mathrm{N}(\mathrm{U}(t))$$

with a matrix $\mathbf{L} \in \mathbb{R}^{N\times N}$ and the nonlinear part $\mathrm{N}(\mathrm{U}(t))$, the sEPIRK schemes were proposed by [2]. They exploit the observation that the stiffness of a system is typically restricted to the linear part $\mathbf{L}\mathrm{U}(t)$. They are defined analogously to EPIRK schemes with the difference that $\mathbf{A}_n$ is replaced by $\mathbf{L}$ and the remainder term R_n by the nonlinear term N. Thus, it is not necessary to computate the complete Jacobian of F. However, for a linear problem

$$\frac{d}{dt}\mathrm{U}(t) = \mathrm{F}(\mathrm{U}(t)), \ \mathrm{F}(\mathrm{U}) = \mathbf{A}\mathrm{U}, \ \mathrm{U}(0) = \mathrm{U}_0, \quad \mathrm{U} : \mathbb{R}_0^+ \to \mathbb{R}^N, \mathbf{A} \in \mathbb{R}^{N\times N}, \tag{3}$$

it follows that $\mathbf{L} = \mathrm{F}'(\mathrm{U}) = \mathbf{A}$ and $\mathrm{N}(\mathrm{U}(t)) = 0$. Hence, an sEPIRK scheme reduces to

$$\mathrm{U}_{n+1} = \mathrm{U}_n + b_1\psi_{s1}(g_{s1}h\mathbf{A})\,h\mathrm{F}(\mathrm{U}_n),$$

which corresponds to an EPIRK scheme applied to a linear problem (3). Consequently, the properties analyzed in Sects. 2.1 and 2.2 likewise hold true for the sEPIRK schemes.

The evaluation of the exponential-like φ-functions for a matrix $\mathbf{A}$ of high dimension N and linear combinations of them in form of the ψ-functions applied to a vector v is a computationally expensive process and therefore has to be approximated in an efficient way. This is done by an algorithm well documented by Niesen and Wright [5] and is summarized as follows:

Algorithm 1 (**Adaptive algorithm documented by Niesen and Wright** [5]) *The task is to approximate the evaluation of*

$$\psi(gh\mathbf{A})v = \sum_{j=1}^{s} p_j \varphi_j(gh\mathbf{A})v \tag{4}$$

for arbitrary $g, h, p_1, \ldots, p_s \in \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{N\times N}$, $v \in \mathbb{R}^N$.

- $\psi(gh\mathbf{A})v$ *is the solution at* $\tau = 1$ *of the special IVP*

$$y'(\tau) = gh\boldsymbol{A}y(\tau) + \sum_{j=1}^{s} \frac{\tau^{j-1}}{(j-1)!} p_j v, \quad y(0) = y_0 = 0,$$

for which the exact solution can be expressed using just one φ*-evaluation as*

$$y(\tau_k + \eta_k) = \eta_k^s \varphi_s(gh\eta_k\boldsymbol{A})w_s + \sum_{j=0}^{s-1} \frac{\eta_k^j}{j!} w_j \quad \textit{for } k = 0, 1, \ldots$$

with

$$w_0 = y(\tau_k) \textit{ and } w_\ell = gh\boldsymbol{A}w_{\ell-1} + \sum_{j=0}^{s-\ell} \frac{\tau_k^j}{j!} p_{(\ell+j)} v \quad \textit{for } \ell = 1, \ldots, s. \tag{5}$$

- *Start at* $\tau_0 = 0$ *and reach* $\tau_K = \tau_{K-1} + \eta_{K-1} = 1$ *after* K *interior steps with step sizes* η_k, $k = 0, \ldots, K-1$.
- *In each step perform a Krylov projection of*

$$\varphi_s(gh\eta_k\mathbf{A})w_s = \sum_{i=0}^{\infty} \frac{(gh\eta_k)^i}{(i+s)!} \mathbf{A}^i w_s$$

into a Krylov subspace $K_m(\mathbf{A}, w_s) = span\{text w_s, \mathbf{A}w_s, \mathbf{A}^2 w_s, \ldots, \mathbf{A}^{m-1} w_s\}$*:*

$$\varphi_s(gh\eta_k\mathbf{A})w_s \approx P_m(\varphi_s(gh\eta_k\mathbf{A})w_s) := ||w_s|| \mathbf{V_m} \varphi_s(gh\eta_k \mathbf{H}_m) e_1 \; + \; corr.$$

with matrices $\mathbf{V_m}, \mathbf{H}_m \in \mathbb{R}^{m\times m}$ *provided by the Arnoldi algorithm,* $m \leq N$ *and a correction term corr., which will be neglected in the following.*

- *The resulting interior time stepping scheme is of the form*

$$y_{k+1} = \eta_k^s P_m(\varphi_s(gh\eta_k \mathbf{A})w_s) \; + \; \sum_{j=0}^{s-1} \frac{\eta_k^j}{j!} w_j$$

  for $k = 0, 1, \ldots, K-1$ *with* $w_0 = y_k$ *and* $w_1, \ldots, w_s$ *as in (5). The last iterated* y_K *is the desired approximation to the expression (4).*
- *An adaptivity strategy for* m *and* η_k *minimizing the computational cost at a prescribed tolerance is provided.*

In the following, we will examine the stability properties as well as the CFL condition in the case of applying the Algorithm 1. In the last Sect. 2.3, we will investigate a conservation property of the EPIRK schemes.

2.1 A-Stability

Let us consider a linear problem (3).

Theorem 2. *Each (s)EPIRK scheme satisfying*

$$b_1 p_{s11} = 1, \;\; g_{s1} = 1, \;\; p_{s1i} = 0 \;\; \textit{for} \;\; i = 2, 3 \ldots, s. \tag{6}$$

solves a linear problem (3) exactly if φ_1 *is computed exactly.*

Proof. The exact solution of (3) assuming $\mathrm{U}_n = \mathrm{U}(t_n)$ is given by

$$\mathrm{U}(t_{n+1}) = e^{h\mathbf{A}}\mathrm{U}_n = \mathrm{U}_n + h(e^{h\mathbf{A}} - \mathbf{I})(h\mathbf{A})^{-1}\mathbf{A}\mathrm{U}_n = \mathrm{U}_n + h\varphi_1(h\mathbf{A})\mathrm{F}(\mathrm{U}_n).$$

Due to $\mathrm{R}_n(\mathrm{U}_n) = \mathrm{F}(\mathrm{U}(t)) - \mathrm{F}_n - \mathbf{A}(\mathrm{U}(t) - \mathrm{U}_n) = \mathbf{A}\mathrm{U}(t) - \mathbf{A}\mathrm{U}_n - \mathbf{A}(\mathrm{U}(t) - \mathrm{U}_n) = 0$ and consequently, $\Delta^{(j)}\mathrm{R}_n(\mathrm{U}_n) = 0, \;\; j = 1, 2, \ldots, s-1$, an EPIRK scheme reduces to

$$\mathrm{U}_{n+1} = \mathrm{U}_n + \; b_1 \; h \sum_{i=1}^{s} p_{s1i}\varphi_i(g_{s1}h\mathbf{A})\mathrm{F}_n.$$

Postulating $\mathrm{U}_{n+1} \stackrel{!}{=} \mathrm{U}(t_{n+1})$ and thus,

$$\mathrm{U}_n + \; b_1 \; h \sum_{i=1}^{s} p_{s1i}\varphi_i(g_{s1}h\mathbf{A})\mathrm{F}_n \stackrel{!}{=} \mathrm{U}_n + h\varphi_1(h\mathbf{A})\mathrm{F}_n,$$

we arrive at the conditions (6).

Since a linear system is solved exactly by EPIRK schemes satisfying the parameter conditions (6), we trivially obtain the following corollary:

Corollary 1. *EPIRK schemes satisfying conditions (6) are A- and L-stable if φ_1 is evaluated exactly.*

The large benefit of an exponential integrator in general is the property shown in Theorem 2, i.e., supposing the conditions (6) yields the exact solution to linear problems. All EPIRK methods proposed by Tokman and co-authors [1, 3] satisfy those conditions. Hence, we will consider only the practically useful EPIRK schemes which fulfill these parameter conditions and therefore the considered EPIRK schemes have the form

$$\mathrm{U}_{n+1} = \mathrm{U}_n + h\varphi_1(h\mathbf{A})\mathrm{F}(\mathrm{U}_n) + h\sum_{k=2}^{s} b_k \psi_{sk}(g_{sk}h\mathbf{A}_n)\Delta^{(k-1)}\mathrm{R}_n(\mathrm{U}_n). \tag{7}$$

Now, let's have a look at the stability properties of EPIRK schemes when using the approximated φ-evaluations.

Theorem 3. *Let* $\deg(v)$ *denote the degree of the minimal polynomial of v with respect to* $\mathbf{A}$. *EPIRK schemes are not A- and L-stable if the Algorithm 1 is applied for Krylov subspace dimensions m with* $m < \deg(v) \leq N$ *and* $m > 1$.

Proof. The projection within Algorithm 1 can be written as

$$P_m(\varphi_j(\mathbf{A})\mathrm{v}) = \sum_{i=1}^{m} \widetilde{\beta}_i^{(j)}\mathrm{v}_i \quad \text{with} \quad \widetilde{\beta}_i^{(j)} = ||\mathrm{v}||\left(\varphi_j(\mathrm{H}_m)\mathrm{e}_1\right)_i.$$

using the Arnoldi basis vectors v_i. After a basis transformation to the standard Krylov basis vectors $\mathbf{A}^{i-1}\mathrm{v}$, $i = 1, \ldots, m$, the projection can be expressed as

$$P_m(\varphi_j(\mathbf{A})\mathrm{v}) = \sum_{i=1}^{m} \beta_i^{(j)}\mathbf{A}^{i-1}\mathrm{v} \tag{8}$$

with some implicit coefficients $\beta_i^{(j)}$ determined by the basis transformation matrix operating on the coefficients $\widetilde{\beta}_i^{(j)}$.

For the linear problem (3), it is $\mathrm{R}_n(\mathrm{Y}) = 0, \mathrm{Y} \in \mathbb{R}^N$, and therefore, $\Delta^{(k)}\mathrm{R}_n(\mathrm{U}_n) = 0, k = 0, 1 \ldots, s$. Consequently, an EPIRK scheme reduces to

$$\mathrm{U}_{n+1} = \mathrm{U}_n + h\varphi_1(h\mathbf{A})\mathrm{F}(\mathrm{U}_n).$$

Applying Algorithm 1 for evaluating $\varphi_1(h\mathbf{A})\mathrm{F}(\mathrm{U}_n)$ means having $p_1 = 1, p_2 = p_3 = \ldots = p_s = 0$, $g = 1$, and $\mathrm{v} = \mathrm{F}(\mathrm{U}_n)$ in the expression (4), leading to

$$\mathrm{w}_0 = \mathrm{y}_k, \quad \mathrm{w}_j = (h\mathbf{A})^j\mathrm{y}_k + (h\mathbf{A})^{j-1}\mathrm{F}(\mathrm{U}_n), \quad j = 1, \ldots, s.$$

The interior step y_{k+1}, $k \in \{0, 1, \ldots, K\}$ is then given by

$$\begin{aligned}
\mathrm{y}_{k+1} &= \eta_k^s P_m(\varphi_s(h\eta_k \mathbf{A})\mathrm{w}_s) \; + \; \sum_{j=0}^{s-1} \frac{\eta_k^j}{j!}\mathrm{w}_j \\
&= \eta_k^s P_m(\varphi_s(h\eta_k \mathbf{A})\mathrm{w}_s) \; + \mathrm{y}_k + \; \sum_{j=1}^{s-1} \frac{\eta_k^j}{j!}\Big((h\mathbf{A})^j \mathrm{y}_k + (h\mathbf{A})^{j-1}\mathrm{F}(\mathrm{U}_n)\Big) \\
&\overset{(8)}{=} \eta_k^s \sum_{i=1}^{m} \beta_i^{(1)} (h\eta_k \mathbf{A})^i \Big((h\mathbf{A})^s \mathrm{y}_k + (h\mathbf{A})^{s-1}\mathrm{F}(\mathrm{U}_n)\Big) \\
&\quad + \mathrm{y}_k + \; \sum_{j=1}^{s-1} \frac{\eta_k^j}{j!}\Big((h\mathbf{A})^j \mathrm{y}_k + (h\mathbf{A})^{j-1}\mathrm{F}(\mathrm{U}_n)\Big) \\
&= \sum_{i=1}^{m} \beta_i^{(1)} \eta_k^{s+i} \Big((h\mathbf{A})^{s+i} \mathrm{y}_k + (h\mathbf{A})^{s-1+i}\mathrm{F}(\mathrm{U}_n)\Big) \\
&\quad + \mathrm{y}_k + \; \sum_{j=1}^{s-1} \frac{\eta_k^j}{j!}\Big((h\mathbf{A})^j \mathrm{y}_k + (h\mathbf{A})^{j-1}\mathrm{F}(\mathrm{U}_n)\Big).
\end{aligned}$$

In case of one inner step ($K = 1$), we have $\mathrm{y}_0 = 0$ and

$$\mathrm{y}_{K=1} = \mathrm{y}_1 = \sum_{i=1}^{m} \beta_i^{(1)} \eta_0^{s+i} (h\mathbf{A})^{s-1+i}\mathrm{F}(\mathrm{U}_n) \; + \; \sum_{j=1}^{s-1} \frac{\eta_0^j}{j!}(h\mathbf{A})^{j-1}\mathrm{F}(\mathrm{U}_n) \tag{9}$$

with step size $\eta_0 = 1$, leading to the EPIRK scheme accounting the approximation:

$$\begin{aligned}
\mathrm{U}_{n+1} &= \mathrm{U}_n + h\mathrm{y}_K \\
&= \mathrm{U}_n + h\Big(\sum_{i=1}^{m} \beta_i^{(1)} \eta_0^{s+i} (h\mathbf{A})^{s-1+i}\mathbf{A}\mathrm{U}_n \; + \; \sum_{j=1}^{s-1} \frac{\eta_0^j}{j!}(h\mathbf{A})^{j-1}\mathbf{A}\mathrm{U}_n\Big) \\
&= \mathrm{U}_n + \sum_{i=1}^{m} \beta_i^{(1)} \eta_0^{s+i} (h\mathbf{A})^{s+i}\mathrm{U}_n \; + \; \sum_{j=1}^{s-1} \frac{\eta_0^j}{j!}(h\mathbf{A})^{j}\mathrm{U}_n \\
&= \underbrace{\Big(\mathbf{I} + \sum_{i=1}^{m} \beta_i^{(1)} \eta_0^{s+i} (h\mathbf{A})^{s+i} \; + \; \sum_{j=1}^{s-1} \frac{\eta_0^j}{j!}(h\mathbf{A})^{j}\Big)}_{:=S(h\mathbf{A})} \mathrm{U}_n.
\end{aligned}$$

The stability function S turns out to be a polynomial of degree $s + m$, since $\beta_m^{(1)} \neq 0$ due to the natural assumption $m < \deg(\mathrm{v})$ resulting from $\dim K_m(\mathbf{A}, \mathrm{v}) = \min\{m, \deg(\mathrm{v})\}$ and the typical case $m \ll N$, $\deg(\mathrm{v}) \lesssim N$.

In case of two inner steps ($K = 2$), notice that y_1 is of the form $\mathrm{y}_1 = \widetilde{S}(h\mathbf{A})\mathrm{F}(\mathrm{U}_n)$ with $\widetilde{S}$ being a polynomial of degree $m + s - 1$ (see (9)). This form is inserted into y_2

$$\mathrm{y}_{K=2} = \mathrm{y}_2 = \sum_{i=1}^{m} \beta_i^{(1)} \eta_1^{s+i} \Big((h\mathbf{A})^{i+s} \widetilde{S}(h\mathbf{A})\mathrm{F}(\mathrm{U}_n) + (h\mathbf{A})^{s-1+i} \mathrm{F}(\mathrm{U}_n) \Big)$$
$$+ \widetilde{S}(h\mathbf{A})\mathrm{F}(\mathrm{U}_n) + \sum_{j=1}^{s-1} \frac{\eta_1^j}{j!} \Big((h\mathbf{A})^j \widetilde{S}(h\mathbf{A})\mathrm{F}(\mathrm{U}_n) + (h\mathbf{A})^{j-1} \mathrm{F}(\mathrm{U}_n) \Big)$$

providing $\mathrm{U}_{n+1} = \mathrm{U}_n + h\mathrm{y}_2$ and hence leading to a stability polynomial of degree $2(s + m)$.

In an analogue manner, it can be shown for an arbitrary number of inner steps $K \geq 1$ that the stability function is a polynomial of degree $K(s + m)$. Consequently, linear stability depends on the Krylov dimension m, the number of stages s, the number of interior steps K as well as the inner steps η_k, and, naturally, on the time step size h. Hence, the stability region is bounded and A- and L-stability are not guaranteed anymore.

2.2 CFL Condition

In the context of discretized PDE's, the CFL condition has to be accounted for to guarantee stability and thus, convergence of the numerical method. The CFL condition can be viewed as a condition on the chosen time step. In its original form, it demands that the numerical domain of dependence given by the numerical stencil includes the physical domain of dependence given by characteristic speed with which information travels through the computational grid. Since we are using explicit EPIRK schemes, the computational stencil might not cover the whole physical domain of dependence for arbitrarily large time step sizes (see Fig. 1 for illustration).

Let us consider the linear advection equation

$$\partial_t u(t, x) + a\, \partial_x u(t, x) = 0, \quad a > 0, \quad x \in [x_0, x_{end}] \subset \mathbb{R} \tag{10}$$

with the characteristic speed a and appropriate initial and boundary conditions. A first order backward difference discretization on the nodes $x_0, x_1, \ldots, x_{N+1} = x_{end}$ with mesh width Δx leads to the system of ODE's

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{U}(t) = \mathrm{F}(\mathrm{U}(t)) \quad \text{with} \quad \mathrm{F}(\mathrm{U}(t)) = \mathbf{A}\mathrm{U}(t), \quad \mathrm{U} : \mathbb{R}_0^+ \to \mathbb{R}^N, \tag{11}$$

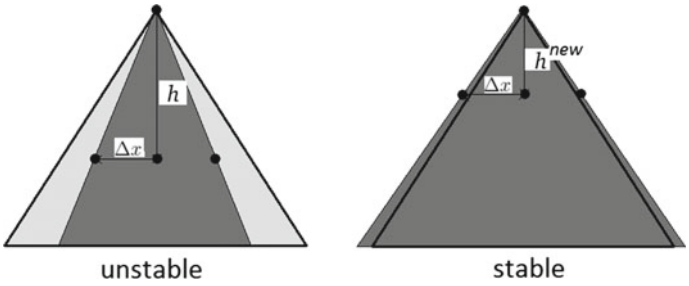


Fig. 1 Illustration of the CFL condition: The physical domain of dependence (light gray) given by the characteristics of the linear advection equation must be included in the numerical one given by the computational stencil (dark gray)

whereby $\mathbf{A}$ shows the structure

$$\mathbf{A} = -\frac{a}{\Delta x}\begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

The following result concerning a CFL condition for EPIRK schemes will be shown:

Theorem 4. *An EPIRK method with s stages applied to the discretization (11) using mesh width Δx of the linear advection equation (10) with characteristic speed a*

- *satisfies the CFL condition for any time step size h if the φ-functions are evaluated exactly, i.e. the time step size is not restricted by the CFL condition*
- *has to satisfy the CFL condition*

$$h \leq \frac{K(s+m)\Delta x}{|a|}$$

when the Algorithm 1 is utilized for the approximation of the φ-evaluations in a Krylov subspace of dimension m with K interior steps.

Proof. The physical domain of dependence of a component $(\mathrm{U}_{n+1})_p \approx u(\underbrace{t_n + h}_{=t_{n+1}}, x_p)$, $p \in \{0, \ldots, N\}$, is given by

$$D^{\text{phys}}((\mathrm{U}_{n+1})_p) = \{(\mathrm{U}_n)_q\}$$

with $q \in \{1, \ldots, p\}$ such that $x_p - |a|h \in [x_q, x_{q+1}]$.

First, let's consider the case of exact φ-evaluations. We notice that the kth power of $\mathbf{A}$ has lower bandwidth k and consequently, $\mathbf{A}^{N-1}$ is a lower triangular matrix without zeros beneath the diagonal. Due to the series definition (2) of φ_j, $\varphi_j(\mathbf{A})$ has the same structure as $\mathbf{A}^{N-1}$ and accordingly, is a lower triangular matrix without zeros beneath or on the diagonal. According to the form (7) of EPIRK schemes, the computation of U_{n+1} contains the expression $\varphi_1(\mathbf{A})\mathrm{F}_n$ leading to a lower triangular matrix operating on F_n. Therefore, the component $(\mathrm{U}_{n+1})_p$ depends on $(\mathrm{F}_n)_q$, $q \in \{1, \ldots, p\}$ and accordingly, on $(\mathrm{U}_n)_q$, $q \in \{1, \ldots, p\}$ providing the numerical domain of dependence

$$D^{\mathrm{num},*}((\mathrm{U}_{n+1})_p) = \{(\mathrm{U}_n)_1, (\mathrm{U}_n)_2, \ldots, (\mathrm{U}_n)_p\}.$$

Therefore, the CFL condition $D^{\mathrm{phys}}((\mathrm{U}_{n+1})_p) \subset D^{\mathrm{num},*}((\mathrm{U}_{n+1})_p)$ is always satisfied and the first statement of the theorem is shown.

Now let's consider approximate φ-evaluations. As we have seen in the proof of Theorem 3, an EPIRK scheme in combination with Algorithm 1 applied to a linear system can be expressed in the form

$$\mathrm{U}_{n+1} = S(h\mathbf{A})\mathrm{U}_n$$

with a stability polynomial S of degree $K(s+m)$. In the monomial expansion of S, the matrix $A^{K(s+m)}$ thus contributes to the largest number of nonzero entries leading to the numerical domain of dependence

$$D^{\mathrm{num}}((\mathrm{U}_{n+1})_p) = \{(\mathrm{U}_n)_{p-K(s+m)}, \ldots, (\mathrm{U}_n)_p\}.$$

Having a more precise look at the stability function, we notice that $(\mathrm{U}_n)_{p-K(s+m)}$ is always contained in the numerical domain of dependence since $\beta_m^{(1)} \neq 0$ and $\eta_k \neq 0$, $k = 0, 1, \ldots, K-1$. Unfortunately, we cannot exclude that other components may disappear for specific combinations of the projection coefficients $\beta_i^{(1)}$, $i = 1, 2, \ldots, m$.

In case of the full numerical domain of dependence, the CFL condition for advection is thereby given by $D^{\mathrm{phys}} \subset D^{\mathrm{num}}$ for any $q \in \{1, \ldots, p\}$ and accordingly,

$$x_{p-K(s+m)} \leq x_q \quad \Leftrightarrow \quad x_p - K(s+m)\Delta x \leq x_p - |a|h \quad \Leftrightarrow \quad h \leq \frac{K(s+m)\Delta x}{|a|}.$$

In the other case, that condition is an upper bound, so that in fact even smaller time step sizes may be required to guarantee stability.

To computationally verify the CFL condition, we numerically determined the maximal stable time step sizes h^{num} for some values of m and K and compared them to the upper bound h^{CFL} given by the CFL condition.

Table 1 Comparison of the theoretical and numerical maximal stable time step size for a linear advection discretization

K	1	1	1	1	1	1	1	1	2	4	8	16
m	1	2	3	4	5	10	50	100	5	5	5	5
h^{CFL}	4.0e-3	6.0e-3	8.0e-3	1.0e-2	1.2e-2	2.2e-2	1.02e-1	2.02e-1	2.4e-2	4.8e-2	9.6e-2	1.92e-1
h^{num}	4.0e-3	6.0e-3	8.0e-3	0.9e-2	1.1e-2	2.0e-2	1.01e-1	2.02e-1	2.2e-2	4.5e-2	9.0e-2	1.81e-1
Diff[α]	0	0	0	0.1e-2	0.1e-2	0.2e-2	0.01e-1	0	0.2e-2	0.3e-2	0.6e-2	0.11e-1

[α] Diff denotes the difference $h^{\text{CFL}} - h^{\text{num}}$

The setup of the test case is the same as in Eq. (11) with $x \in [0, 2]$, $\Delta x = 1\text{e-}3$, $N = 1999$, $a = 0.5$, $s = 1$. The initial condition is given by the exact solution $u(t, x) = cos(\pi(x - at))$ and we computed until $t_{end} = 4$.

The results in Table 1 approve that the proposed CFL condition is a necessary condition for stability but not a sufficient one. Nevertheless, it gives a reasonable guide and in combination with a safety factor < 1 it appears well suited for time step size adaption within an implementation.

2.3 Conservation

For conservation laws as well as for balance laws such as the Navier–Stokes equations, it is natural to apply conservative numerical flux functions in the spatial discretization, meaning that the flow between two cells is the same for both normal directions. Since the right-hand side F of an ODE system (1) resulting from the spatial discretization with conservative numerical flux functions is a spatial discretization operator, it follows for a computational domain with periodic boundary conditions that

$$\sum_{j=1}^{N} (\mathrm{F}(\mathrm{U}))_j = 0 \quad \forall\, \mathrm{U} \in \mathbb{R}^N. \tag{12}$$

We call such a right-hand side F globally conservative. With this property of F, the sum of the components of the solution U does not change in time:

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{j=1}^{N} (\mathrm{U})_j = \sum_{j=1}^{N} \left(\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{U}\right)_j \overset{(1)}{=} \sum_{j=1}^{N} (\mathrm{F})_j \overset{(12)}{=} 0.$$

Hence, the time discretization should also maintain this global conservation property, defined as follows:

Definition 1. Given an initial value problem (1) with a globally conservative right-hand side F, see (12), a time stepping method is called *globally conservative*, if

$$\sum_{j=1}^{N} (\mathrm{U}_{n+1})_j = \sum_{j=1}^{N} (\mathrm{U}_n)_j.$$

We will prove the following statement.

Theorem 5. *The EPIRK schemes are globally conservative.*

Proof. Let $\mathbf{1}_N \in \mathbb{R}^N$ denote the vector consisting of ones and $\mathbf{0}_N \in \mathbb{R}^N$ the zero vector. Due to the conservation property (12) of F, we have

$$\frac{\mathrm{d}}{\mathrm{dU}} \sum_{j=1}^{N} \big(\mathrm{F}(\mathrm{U})\big)_j = 0 \Leftrightarrow \sum_{j=1}^{N} \frac{\mathrm{d}}{\mathrm{dU}} \big(\mathrm{F}(\mathrm{U})\big)_j = 0 \quad \forall\, U \in \mathbb{R}^N$$

and accordingly,

$$\mathbf{1}_N^T \mathrm{F}'(\mathrm{U}) = \Big(\sum_{j=1}^{N} \frac{\partial}{\partial \mathrm{U}_1} \big(\mathrm{F}(\mathrm{U})\big)_j, \ \sum_{j=1}^{N} \frac{\partial}{\partial \mathrm{U}_2} \big(\mathrm{F}(\mathrm{U})\big)_j, \ \ldots, \ \sum_{j=1}^{N} \frac{\partial}{\partial \mathrm{U}_N} \big(\mathrm{F}(\mathrm{U})\big)_j \Big) = \mathbf{0}_N^T. \tag{13}$$

It follows with the series definition (2) of φ that

$$\begin{aligned} \mathbf{1}_N^T \psi_{sk}(c\mathrm{F}'(\mathrm{U})) &\overset{(2)}{=} \sum_{\ell=1}^{s} p_{sk\ell} \mathbf{1}_N^T \left(\frac{1}{\ell!}\mathbf{I} + c\mathrm{F}'(\mathrm{U}) \Big(\frac{1}{(\ell+1)!}\mathbf{I} + \sum_{j=2}^{\infty} \frac{1}{(\ell+j)!} (c\mathrm{F}'(\mathrm{U}))^{j-1} \Big) \right) \\ &\overset{(13)}{=} \sum_{\ell=1}^{s} p_{sk\ell} \frac{1}{\ell!} \mathbf{1}_N^T \end{aligned} \tag{14}$$

holds for arbitrary $c \in \mathbb{R}$ and all $\mathrm{U} \in \mathbb{R}^N$. Furthermore, we get

$$\mathbf{1}_N^T \mathrm{R}_n(\mathrm{U}(t)) = \sum_{j=1}^{N} \Big(\mathrm{F}(\mathrm{U}(t)) \Big)_j - \sum_{j=1}^{N} \Big(\mathrm{F}(\mathrm{U}_n) \Big)_j - \mathbf{1}_N^T \mathrm{F}'(\mathrm{U}_n)(\mathrm{U}(t) - \mathrm{U}_n) \overset{(12),(13)}{=} 0$$

and accordingly,

$$\mathbf{1}_N^T \Delta^k \mathrm{R}_n(\mathrm{U}_n) = \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} \mathbf{1}_N^T \mathrm{R}_n(\mathrm{Y}_{k-i}) = 0 \tag{15}$$

for all $k = 1, 2, \ldots, s$. In conclusion, we arrive at

$$\begin{aligned}\sum_{j=1}^{N}(\mathrm{U}_{n+1})_j &= \sum_{j=1}^{N}(\mathrm{U}_n)_j + b_1\mathbf{1}_N^T\psi_{s1}(g_{s1}\mathbf{A}_n h)h\mathrm{F}_n \\ &\quad + h\sum_{k=2}^{s} b_k\mathbf{1}_N^T\psi_{sk}(g_{sk}\mathbf{A}_n h)\Delta^{(k-1)}\mathrm{R}_n(\mathrm{Y}_{k-1}) \\ &\overset{(14)}{=} \sum_{j=1}^{N}(\mathrm{U}_n)_j + b_1\sum_{\ell=1}^{s} p_{s1\ell}\frac{1}{\ell!}\mathbf{1}_N^T h\mathrm{F}_n \\ &\quad + h\sum_{k=2}^{s} b_k\sum_{\ell=1}^{s} p_{sk\ell}\frac{1}{\ell!}\mathbf{1}_N^T\Delta^{(k-1)}\mathrm{R}_n(\mathrm{Y}_{k-1}) \\ &\overset{(12),(15)}{=} \sum_{j=1}^{N}(\mathrm{U}_n)_j.\end{aligned}$$

The conservation property of the sEPIRK schemes depends on the specific splitting of the right-hand side F into **L** and N. In our current work, we adopt the sEPRIK schemes into a domain-based implicit–explicit setting, which is also done for example in [6]. We are even able to show a conservation property of these schemes in that framework, which will be published next.

Acknowledgements We thank the German Research Foundation DFG for its financial support within the project GZ: ME 1889/7-1.

References

1. M. Tokman, A new class of exponential propagation iterative methods of Runge-Kutta type (EPIRK). J. Comput. Phys. **230**, 8762–8778 (2011)
2. G. Rainwater, M. Tokman, A new class of split exponential propagation iterative methods of Runge-Kutta type (sEPIRK) for semilinear systems of ODEs. J. Comput. Phys. **269**, 40–60 (2014)
3. J. Loffeld, M. Tokman, Comparative performance of exponential, implicit, and explicit integrators for stiff systems of ODEs. J. Comput. Appl. Math. **241**(1), 45–67 (2013)
4. V. Straub, S. Ortleb, P. Birken, A. Meister, Efficient time integration of IMEX type using exponential integrators for compressible, viscous flow simulation. PAMM **16**, 867–868 (2016)
5. J. Niesen, W.M. Wright, Algorithm 919: a Krylov subspace algorithm for evaluating the phi-functions appearing in exponential integrators. ACM TOMS **38**(3), 1–19 (2012)
6. A. Kanevsky, M.H. Carpenter, D. Gottlieb, J.S. Hesthaven, Application of implicit-explicit high order Runge-Kutta methods to discontinuous-Galerkin schemes. J. Comput. Phys. **225**, 1753–1781 (2007)

Compactness on Multidimensional Steady Euler Equations

Tian-Yi Wang

Abstract Recently, the new compactness frameworks on the multidimensional steady Euler equations are established. At the beginning, we will start from a motivative example on the steady Euler equation. Then, the formal compactness framework for approximate solutions to subsonic-sonic flows governed by the steady compressible Euler equations in arbitrary dimension is introduced. Later, we will present the compactness framework of incompressible limit to the steady compressible Euler flow. At the end, as the direct applications of the compactness framework are mentioned.

Keywords Multidimensional · Steady Euler flow · Compactness framework Subsonic-Sonic limit · Incompressible limit

1 Introduction

The full Euler equations for steady compressible flows in $\mathbf{R}^d$ read

$$\begin{cases} \operatorname{div}(\rho u) = 0, \\ \operatorname{div}(\rho u \otimes u) + \nabla p = 0, \\ \operatorname{div}(\rho u E + u p) = 0, \end{cases} \tag{1}$$

where $x = (x_1, \cdots, x_d) \in \mathbf{R}^d$, $d \geq 2$, $u = (u_1, \cdots, u_d) \in \mathbf{R}^d$ is the fluid velocity, and

$$q := |u| = \Big(\sum_{i=1}^{d} u_i^2 \Big)^{1/2}$$

T.-Y. Wang (✉)
Department of Mathematics, School of Science, Wuhan University of Technology, Wuhan 430070, Hubei, China
e-mail: tianyiwang@whut.edu.cn; tian-yi.wang@gssi.infn.it; wangtianyi@amss.ac.cn

T.-Y. Wang
Gran Sasso Science Institute, viale Francesco Crispi, 7, 67100 L'Aquila, Italy

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_47

is the speed, while ρ, p, and E represent the density, pressure, and total energy, respectively. The non-negative quantities ρ, q, p, and E are not independent. For ideal polytropic gas,

$$E = \frac{q^2}{2} + \frac{p}{(\gamma - 1)\rho}$$

with adiabatic exponent $\gamma > 1$. In this case, the Bernoulli law is written as

$$\frac{q^2}{2} + h(\rho, p) = B, \tag{2}$$

where $h(\rho, p) = \frac{\gamma p}{(\gamma-1)\rho}$ is the enthalpy, and B is a Bernoulli function determined by appropriate additional conditions (such as boundary conditions and/or asymptotic conditions at infinity). The sound speed of the flow is

$$c = \sqrt{\frac{\gamma p}{\rho}}, \tag{3}$$

and the Mach number is defined as

$$M = \frac{q}{c}. \tag{4}$$

Then, for a fixed Bernoulli function B, there is a critical speed $q_{\rm cr} = \sqrt{2\frac{\gamma-1}{\gamma+1}B}$ such that, when $q \le q_{\rm cr}$, the flow is subsonic-sonic (i.e., $M \le 1$); otherwise, it is supersonic (i.e., $M > 1$) [1, 2].

It is well known that the steady Euler equations for compressible fluids are of composite-mixed type, which is determined by the Mach number M. That is, the system can be reduced to a system such that two of the equations are elliptic-hyperbolic mixed: elliptic when $M < 1$ and hyperbolic when $M > 1$, while the other n equations are hyperbolic.

Also, the entropy is a typical quantity:

$$S = \frac{\gamma p}{(\gamma - 1)\rho^\gamma}, \tag{5}$$

which is conserved along the streamline. When the S be a positive constant in the field, the flow is called as homentropic, in which the pressure becomes $p = \rho^\gamma$ by picking the proper constant.

During the 1950s, the effort on steady Euler equations was focused mainly on the irrotational case, namely when u is constrained to satisfy the additional equation $\operatorname{curl} u = 0$, when the flow is homentropic. Since the equations of uniform subsonic flow possess ellipticity, solutions have better regularity than those corresponding to transonic or supersonic flow. The airfoil problem for two-dimensional subsonic flow was solved; *cf.* Shiffman [3], Bers [4], and Finn-Gilbarg [5]. The first result for three-

dimensional subsonic flow past an obstacle was given by Finn and Gilbarg [6] under some restrictions on the Mach number. Dong [7] and Dong-Ou [8] extended the results to the maximum Mach number $M < 1$ for the arbitrarily dimensional case. And, Du-Xin-Yan [9] constructed the smooth uniform subsonic flow in an infinitely long nozzle in $\mathbf{R}^d$, $d \geq 2$, while the largely open nozzle case was solved by Liu-Yuan [10]. Recently, the existence and uniqueness of the subsonic flows with conservative forces were considered in [11, 12] for the airfoil problem and infinitely long nozzle problem, respectively.

For the rotational case, the global existence of homentropic subsonic flow through two-dimensional infinitely long nozzles was proved in Xie-Xin [13]. The result was also extended to the two-dimensional periodic nozzles in Chen-Xie [14] and to axisymmetric nozzles in Du-Duan [15]. For the full Euler flow, the first result was given by Chen-Deng-Xiang [16] for two-dimensional infinitely long nozzles. Bae [17] showed the stability of contact discontinuities for subsonic full Euler flow in the two-dimensional, infinitely long nozzles. Duan-Luo [18] considered the axisymmetric nozzle problem for the smooth subsonic flow. The compressible flow in the infinitely long nozzle with the stagnation points was considered by Du-Xie [19]. For the recent series of works on large vorticity with the increasing condition, please see [20–23]. The progress on the airfoil problem of the rotational flow is still restricted in the symmetric body case, which is also called as half plane problem[1], the existence and uniqueness of the subsonic flow are shown in [24, 25].

On the other hand, few results are currently known for the cases of subsonic-sonic flow and transonic flow, since the uniform ellipticity is lost and shocks may be present. That is, smooth solutions may not exist. Instead, one must consider weak solutions. Morawetz [26, 27] introduced an approach via compensated compactness to analyze irrotational steady flow of the Euler equations. Indeed, Morawetz established a compactness framework under the assumption that the solutions are free of stagnation points and cavitation points with the finite flow angle. Morawetz's result has been improved by Chen-Slemrod-Wang [28] in which the approximate solutions away from cavitation are constructed by a viscous perturbation.

The compactness framework for subsonic-sonic irrotational flow allowing for stagnation in two dimensions was due to Chen-Dafermos-Slemrod-Wang [29] by combining the mass conservation, momentum, and irrotational equations. The key observation in [29] is that the two-dimensional steady flow can be regarded as a one-dimensional unsteady system of conservation laws, that is, one of the spatial variables can be regarded as the time variable, so that the div-curl lemma can be applied to the two momentum equations. In fact, the momentum equations are first employed in [29] to reduce the support of the corresponding Young measure to two points, and then the irrotational equation and the mass equation are used to reduce the Young measure to a Dirac measure. Using a similar idea, Xie-Xin [30] investigated the subsonic-sonic limit of the two-dimensional irrotational, infinitely long nozzle problem. Later, in [31], they extended the result to the three-dimensional axisymmetric flow through an axisymmetric nozzle. The compactness framework in the multidimensional irrotational case was established in Huang-Wang-Wang [32].

Recently, the compactness framework on the steady Euler equation without irrotational condition was completed in Chen-Huang-Wang [33].

The compactness framework established for irrotational flow no longer applies directly to the steady full Euler equations in $\mathbf{R}^d$ with $d \geq 2$. When $d \geq 3$, the equations cannot be reduced to a one-dimensional system of conservation laws. More importantly, the div-curl lemma is no longer valid for the momentum equations, due to the presence of linear characteristics. In [33], the main observations are that it is still possible to achieve the same compactness result, i.e., to reduce the Young measure to a Dirac measure, by using only natural weak estimates for the mass balance and the vorticity, along with the Bernoulli law and entropy relation, through a more delicate analysis on the phase space.

On the other hand, the incompressible limit is one of the fundamental fluid dynamic limits in fluid mechanics. Formally, the steady compressible full Euler equations (1) converge to the steady inhomogeneous incompressible Euler equations:

$$\begin{cases} \operatorname{div} u = 0, \\ \operatorname{div} (\rho u) = 0, \\ \operatorname{div} (\rho u \otimes u) + \nabla p = 0. \end{cases} \tag{6}$$

However, the rigorous justification of this limit for weak solutions has been a challenging mathematical problem, since it is a singular limit for which singular phenomena usually occur in the limit process. In particular, both the uniform estimates and the convergence of the nonlinear terms in the incompressible models are usually difficult to obtain.

Generally speaking, there are two processes for the incompressible limit: the adiabatic exponent γ tending to infinity, and the compressible parameter tending to zero. The latter is also called the low Mach number limit, and the previous research please check [34, 35] and references within. For the limit $\gamma \to \infty$, it was shown in [36] that the compressible homentropic Navier–Stokes flow would converge to the homogeneous incompressible Navier–Stokes flow. Later, the similar limit from the Korteweg barotropic Navier–Stokes model to the homogeneous incompressible Navier–Stokes model was also considered in [37].

For the steady flow, the uniqueness of weak solutions of the steady incompressible Euler equations is still an open issue. Thus, the incompressible limit of the steady Euler equations becomes more fundamental mathematically; it may serve as a selection principle of physical relevant solutions for the steady incompressible Euler equations since a weak solution should not be regarded as the compressible perturbation of the steady incompressible Euler flow in general. Furthermore, for the general domain, it is quite challenging to obtain directly a uniform estimate for the Leray projection of the velocity in the compressible fluids.

In [38], we formulate a suitable compactness framework for weak solutions with weak uniform bounds with respect to the adiabatic exponent γ by employing the weak convergence argument. One of the main observations is that the compactness can be achieved by using only natural weak estimates for the mass conservation and the vorticity, which was introduced in [32, 33]. Another observation is that the

incompressibility of the limit for the full Euler flow is from a combination of all the Euler equations.

The proofs of both subsonic-sonic limit and incompressible limit are based on the compensated compactness method and relative method. The suggested references are [39–44].

The rest of this paper is organized as follows. In Sect. 2, we give a heuristic analysis to show the idea on the compactness of velocity for both subsonic-sonic limit and incompressible limit. In Sect. 3, we present the compactness framework for subsonic-sonic approximate solutions to subsonic-sonic flows governed by the steady full Euler equations for compressible fluids in $\mathbf{R}^d$ with $d \geq 2$, In Sects. 4, the compactness framework on the incompressible limit to the multidimensional steady Euler equation is stated. At last, Sect. 5 shows the applications on the above frameworks.

2 A Heuristic Example

In this section, we will consider a simplified model to show the general idea of the coming compactness frameworks. For taking the limits of the steady Euler equations, one of the major difficulties is on the convention term, which needs the strong convergence of the flow speed u. To present the motivation this part, we could propose the homentropic irrotional conditions on the full Euler equations. And the equations come to be:

$$\begin{cases} \mathrm{curl} u = 0, \\ \mathrm{div}(\rho u) = 0, \end{cases} \tag{7}$$

with the Bernoulli law:

$$\frac{q^2}{2} + h(\rho) = \bar{B}, \tag{8}$$

while $\bar{B}$ is a positive constant. In this case, the critical speed $q_{cr}(\bar{B})$ is a fixed positive constant q_{cr}. Then, the density ρ can be regarded as the function of speed q, which will be written as $\rho(q)$. On the other hand, by the irrotation condition $(7)_1$, one can introduce the potential φ, which satisfies $u = \nabla\varphi$. From the above transfer, $(7)_2$ becomes:

$$\mathrm{div}(\rho(|\nabla\varphi|)\nabla\varphi) = 0. \tag{9}$$

Here, for $v \in \mathbf{R}^d$, we introduce the operator $E(v) : \mathbf{R}^d \to \mathbf{R}^d$ as $E(v) := \rho(|v|)v$. Due to the existence of the maximum density, it is easy to see $|E(v)| \leq C(1 + |v|)$, [7]. Next, we will show that for the subsonic-sonic flow, $E(v)$ is monotone, which means for $v^{(1)}, v^{(2)} \in \mathbf{R}^d$, and $|v^{(1)}|, |v^{(2)}| \leq q_{\mathrm{cr}}$,

$$\left(E(v^{(1)}) - E(v^{(2)})\right) \cdot (v^{(1)} - v^{(2)}) \geq 0. \tag{10}$$

Notice that

$$\begin{aligned}
&\big(E(v^{(1)}) - E(v^{(2)})\big) \cdot (v^{(1)} - v^{(2)}) \\
&= \sum_{i=1}^{d} \big(\rho(|v^{(1)}|)v_i^{(1)} - \rho(|v^{(2)}|)v_i^{(2)}\big)\big(v_i^{(1)} - v_i^{(2)}\big) \\
&= \sum_{i=1}^{d} \Big(\rho(|v^{(1)}|)|v_i^{(1)}|^2 - \rho(|v^{(1)}|)v_i^{(1)}v_i^{(2)} - \rho(|v^{(2)}|)v_i^{(2)}v_i^{(1)} + \rho(|v^{(2)}|)|v_i^{(2)}|^2\Big) \\
&= \rho(|v^{(1)}|)\big(|v^{(1)}|^2 - \sum_{i=1}^{d} v_i^{(1)}v_i^{(2)}\big) + \rho(|v^{(2)}|)\big(|v^{(2)}|^2 - \sum_{i=1}^{d} v_i^{(1)}v_i^{(2)}\big).
\end{aligned}$$

The Cauchy inequality implies

$$\begin{aligned}
&\big(E(v^{(1)}) - E(v^{(2)})\big) \cdot (v^{(1)} - v^{(2)}) \\
&\geq \rho(|v^{(1)}|)\big(|v^{(1)}|^2 - |v^{(1)}||v^{(2)}|\big) + \rho(|v^{(2)}|)\big(|v^{(2)}|^2 - |v^{(1)}||v^{(2)}|\big) \\
&= \big(|v^{(1)}| - |v^{(2)}|\big)\big(\rho(|v^{(1)}|)|v^{(1)}| - \rho(|v^{(2)}|)|v^{(2)}|\big) \\
&= \big(|v^{(1)}| - |v^{(2)}|\big)^2 \frac{d(\rho q)}{dq}(\tilde{q}),
\end{aligned}$$

where $\tilde{q}$ lies between $|v^{(1)}|$ and $|v^{(2)}|$ by the mean value theorem. Taking derivative with respect to q on (8), we obtain

$$\frac{d\rho}{dq} = -\frac{\rho q}{c^2}.$$

Then

$$\frac{d(\rho q)}{dq} = \rho(1 - M^2).$$

For subsonic-sonic flows, i.e., $|v^{(1)}|, |v^{(2)}| \leq q_{\mathrm{cr}}$, we have

$$M^2(\tilde{q}) \leq 1.$$

Then

$$\big(E(v^{(1)}) - E(v^{(2)})\big) \cdot (v^{(1)} - v^{(2)}) \geq \big(|v^{(1)}| - |v^{(2)}|\big)^2 \rho(\tilde{q})\big(1 - M^2(\tilde{q})\big) \geq 0. \tag{11}$$

Generally speaking, the monotonicity of $E(v)$ implies the strong compactness of $u = \nabla\varphi$. Please see Chap. 5 of [45], Chap. 2 of [46], and [47] for the further details on the monotonicity method. In the Sect. 3, the general compactness framework of the subsonic-sonic limits will be presented. Furthermore, the idea has been

developed to consider the limit $\gamma \to \infty$, which comes out to be the incompressible limit compactness framework, in which case the operator is $E(v) = v$.

3 Subsonic-Sonic Limits

In this section, we present the compensated compactness framework for approximate solutions of the steady full Euler equations in $\mathbf{R}^d$ with $d \geq 2$ with the form:

$$\begin{cases} \operatorname{div}(\rho^\varepsilon u^\varepsilon) = e_1(\varepsilon), \\ \operatorname{div}(\rho^\varepsilon u^\varepsilon \otimes u^\varepsilon) + \nabla p^\varepsilon = e_2(\varepsilon), \\ \operatorname{div}(\rho^\varepsilon u^\varepsilon E^\varepsilon + u^\varepsilon p^\varepsilon) = e_3(\varepsilon), \end{cases} \tag{12}$$

where $e_1(\varepsilon)$, $e_2(\varepsilon) = (e_{21}(\varepsilon), \cdots, e_{2d}(\varepsilon))^\top$, and $e_3(\varepsilon)$ are sequences of functions depending on the parameter ε.

Let a sequence of functions $\rho^\varepsilon(x), u^\varepsilon(x) = (u_1^\varepsilon, \cdots, u_d^\varepsilon)(x)$, and $p^\varepsilon(x)$ be defined on an open subset $\Omega \subset \mathbf{R}^n$ such that the following qualities:

$$q^\varepsilon := |u^\varepsilon| = \sqrt{\sum_{i=1}^{n} (u_i^\varepsilon)^2}, \quad c^\varepsilon := \sqrt{\frac{\gamma p^\varepsilon}{\rho^\varepsilon}}, \quad M^\varepsilon := \frac{q^\varepsilon}{c^\varepsilon}, \tag{13}$$

$$B^\varepsilon := \frac{(q^\varepsilon)^2}{2} + \frac{\gamma p^\varepsilon}{(\gamma - 1)\rho^\varepsilon}, \qquad S^\varepsilon := \frac{\gamma p^\varepsilon}{(\gamma - 1)(\rho^\varepsilon)^\gamma}$$

can be well defined and satisfy the following conditions:

(A.1). $M^\varepsilon \leq 1$ a.e. in Ω;

(A.2). S^ε and B^ε are uniformly bounded and, for any compact set K, there exists a uniform constant $c(K)$ such that $\inf_{x \in K} S^\varepsilon(x) \geq c(K) > 0$. Moreover, $(S^\varepsilon, B^\varepsilon) \to (\overline{S}, \overline{B})$ a.e. in Ω;

(A.3). curl u^ε and $e_1(\varepsilon)$ are in a compact set in $W_{loc}^{-1,p}$ for some $1 < p \leq 2$. Then we have

Theorem 1 (Compensated compactness framework for the full Euler case [33]). *Let a sequence of functions $\rho^\varepsilon(x)$, $u^\varepsilon(x) = (u_1^\varepsilon, \cdots, u_d^\varepsilon)(x)$, and $p^\varepsilon(x)$ satisfy conditions* (A.1)–(A.3). *Then there exists a subsequence (still labeled) $(\rho^\varepsilon, u^\varepsilon, p^\varepsilon)(x)$ such that*

$$\rho^\varepsilon(x) \to \rho(x), \quad u^\varepsilon(x) \to (u_1, \cdots, u_d)(x), \quad p^\varepsilon(x) \to p(x) \quad \text{a.e. in } x \in \Omega \text{ as } \varepsilon \to 0,$$

and

$$M(x) := \frac{q(x)}{c(x)} \leq 1 \quad \text{a.e. } x \in \Omega.$$

Remark 1. Consider any function $Q(\rho, u, p) = (Q_1, \cdots, Q_d)(\rho, u, p)$ satisfying

$$\operatorname{div}\left(Q(\rho^\varepsilon, u^\varepsilon, p^\varepsilon)\right) = o_Q(\varepsilon), \tag{14}$$

where $o_Q(\varepsilon) \to 0$ in the distributional sense as $\varepsilon \to 0$. We can see from the strong convergence of $(\rho^\varepsilon, u^\varepsilon, p^\varepsilon)$ ensured by Theorem 3 that $\operatorname{div}(Q(\rho, u, p)) = 0$ holds in the distributional sense. Thus, if

$$\operatorname{div}\left(\rho^\varepsilon u^\varepsilon \otimes u^\varepsilon + p^\varepsilon I\right) = e_2(\varepsilon) \to 0 \qquad \text{in the sense of distributions,} \tag{15}$$

the weak solution also satisfies the momentum equations in $(1)_2$ and the energy equation $(1)_3$ in the distributional sense.

Then, as corollaries, we conclude the following theorems.

Theorem 2 (Convergence of approximate solutions for the full Euler flow [33]). *Let $\rho^\varepsilon(x)$, $u^\varepsilon(x) = (u_1^\varepsilon, \cdots, u_d^\varepsilon)(x)$, and $p^\varepsilon(x)$ be a sequence of approximate solutions satisfying* (A.1)–(A.3) *and $e_j(\varepsilon) \to 0$, $j = 1, 2, 3$, in the distributional sense as $\varepsilon \to 0$. Then there exists a subsequence (still labeled) $(\rho^\varepsilon, u^\varepsilon, p^\varepsilon)(x)$ that converges a.e. as $\varepsilon \to 0$ to a weak solution (ρ, u, p) to the Euler equations of (1), which satisfies $M(x) \le 1$, a.e. $x \in \Omega$.*

For the homentropic case, [33] also consider the general pressure–density case, and condition $(A.2)$ and $(A.3)$ are modified.

4 Incompressible Limits

In this section, compensated compactness framework for approximate solutions of the steady Euler equations is presented in $\mathbf{R}^d$ with $d \ge 2$.

Here, we assume that the approximate solutions $(\rho^{(\gamma)}, u^{(\gamma)}, p^{(\gamma)})$ satisfy

$$\begin{cases} \operatorname{div}\left(\rho^{(\gamma)} u^{(\gamma)}\right) = e_1(\gamma), \\ \operatorname{div}\left(\rho^{(\gamma)} u^{(\gamma)} \otimes u^{(\gamma)}\right) + \nabla p^{(\gamma)} = e_2(\gamma), \\ \operatorname{div}\left(\rho^{(\gamma)} u^{(\gamma)} E^{(\gamma)} + u^{(\gamma)} p^{(\gamma)}\right) = e_3(\gamma), \end{cases} \tag{16}$$

where $e_1(\gamma)$, $e_2(\gamma) = (e_{21}(\gamma), \cdots, e_{2d}(\gamma))^\top$, and $e_3(\gamma)$ are sequences of distributional functions depending on the parameter γ.

Let the sequences of functions $u^{(\gamma)}(x) := (u_1^{(\gamma)}, \cdots, u_d^{(\gamma)})(x)$ and $p^{(\gamma)}(x)$ be defined on an open bounded subset $\Omega \subset \mathbf{R}^n$ such that the following qualities:

$$\rho^{(\gamma)} := (p^{(\gamma)})^{\frac{1}{\gamma}}, \quad |u^{(\gamma)}| := \sqrt{\sum_{i=1}^{d}(u_i^{(\gamma)})^2}, \quad c^{(\gamma)} := \sqrt{\gamma}\big(p^{(\gamma)}\big)^{\frac{\gamma-1}{2\gamma}}, \tag{17}$$
$$M^{(\gamma)} := \frac{|u^{(\gamma)}|}{c^{(\gamma)}}, \quad E^{(\gamma)} := \frac{|u^{(\gamma)}|^2}{2} + \frac{p^{(\gamma)}}{(\gamma-1)\rho^{(\gamma)}}, \quad G^{(\gamma)} := \frac{\rho^{(\gamma)}}{(p^{(\gamma)})^{\frac{1}{\gamma}}} \geq 0,$$

can be well defined. Moreover, the following conditions hold:

(B.1). $M^{(\gamma)}$ are uniformly bounded by $\bar{M}$;

(B.2). $|u^{(\gamma)}|^2$ and $p^{(\gamma)} \geq 0$ are uniformly bounded in $L^1_{loc}(\Omega)$;

(B.3). $e_1(\gamma)$ and curl $u^{(\gamma)}$ are in a compact set in $H^{-1}_{loc}(\Omega)$;

(B.4). As $\gamma \to \infty$,

$$\int_{\Omega} \ln(p^{(\gamma)})dx = o(\gamma) \qquad \text{as } \gamma \to \infty;$$

(B.5). $G^{(\gamma)}$ converges to a bounded function $\overline{G}$ a.e. in Ω as $\gamma \to \infty$.

Remark 2. For the Euler equations, G is streamline conservative quantity, which is equivalent to entropy S with the relation: $S = \frac{\gamma}{\gamma-1}G^{-\frac{1}{\gamma}}$.

Theorem 3 (Compensated compactness framework for the full Euler case [38]). *Let a sequence of functions $\rho^{(\gamma)}(x)$, $u^{(\gamma)}(x) = (u_1^{(\gamma)}, \cdots, u_d^{(\gamma)})(x)$, and $p^{(\gamma)}(x)$ satisfy conditions* (B.1)–(B.5). *Then there exists a subsequence (still denoted by) $(\rho^{(\gamma)}, u^{(\gamma)}, p^{(\gamma)})(x)$ such that, as $\gamma \to \infty$,*

$$\begin{aligned} &p^{(\gamma)}(x) \rightharpoonup \bar{p} && \text{in bounded measure,} \\ &\rho^{(\gamma)}(x) \to \bar{\rho}(x) && \text{a.e. in } x \in \Omega, \\ &u^{(\gamma)}(x) \to (\bar{u}_1, \cdots, \bar{u}_d)(x) && \text{a.e. in } x \in \{x : \bar{\rho}(x) > 0, x \in \Omega\}. \end{aligned} \tag{18}$$

Remark 3. Similar to Remark 1, consider any function $Q(\rho, u, p) := (Q_1, \cdots, Q_d)(\rho, u, p)$ satisfying

$$\operatorname{div}(Q(\rho^{(\gamma)}, u^{(\gamma)}, p^{(\gamma)})) = e_Q(\gamma), \tag{19}$$

where $e_Q(\gamma) \to 0$ in the distributional sense as $\gamma \to \infty$.

Then, we come to:

Theorem 4 (Convergence of approximate solutions for the full Euler flow[38]). *Let $\rho^{(\gamma)}(x)$, $u^{(\gamma)}(x) = (u_1^{(\gamma)}, \cdots, u_d^{(\gamma)})(x)$, and $p^{(\gamma)}(x)$ be a sequence of approximate solutions satisfying conditions* (B.1)–(B.5), *and*

$$\begin{aligned} &e_i(\gamma) \to 0 \quad \text{for } i = 1, 2, \\ &(p^{(\gamma)})^{-1}\Big(e_3(\gamma) - u^{(\gamma)} \cdot e_2(\gamma) + \frac{|u^{(\gamma)}|^2}{2}e_1(\gamma)\Big) \to 0 \end{aligned}$$

in the distributional sense as $\gamma \to \infty$. *Then there exists a subsequence (still denoted by)* $(\rho^{(\gamma)}, u^{(\gamma)}, p^{(\gamma)})(x)$ *that converges a.e. to a weak solution* $(\bar{\rho}, \bar{u}, \bar{p})$ *of the inhomogeneous incompressible Euler equations (6) as* $\gamma \to \infty$.

For the homentropic case, [38] also consider the γ-law pressure–density case, and condition $(B.4)$ turns to the condition on the total energy, while $(B.5)$ is released.

5 Application

The conditions $(A.1)$–$(A.3)$ and $(B.1)$–$(B.3)$ are naturally satisfied by the solutions constructed in [9, 11–16, 19–24, 30, 31]. By Theorems 2 and 4, the subsonic-sonic limits and the incompressible limits can be proven from the above results on the subsonic flow. For more detail, please check [33, 38]. There are some further applications on the both limits are included in the coming paper [48, 49].

Acknowledgements The research of author was supported in part by the NSFC Grant No. 11601401, and the Fundamental Research Funds for the Central Universities (WUT: 2017 IVA 072 & 2017 IVB 066).

References

1. L. Bers, *Mathematical Aspects of Subsonic and Transonic Gas Dynamics* (Wiley, Chapman & Hall, New York, London, 1958)
2. R. Courant, K.O. Friedrichs, *Supersonic Flow and Shock Waves* (Interscience Publishers Inc., New York, 1948)
3. M. Shiffman, On the existence of subsonic flows of a compressible fluid. J. Ration. Mech. Anal. **1**, 605–652 (1952)
4. L. Bers, Existence and uniqueness of a subsonic flow past a given profile. Commun. Pure Appl. Math. **7**, 441–504 (1954)
5. R. Finn, D. Gilbarg, Asymptotic behavior and uniqueness of plane subsonic flows. Commun. Pure Appl. Math. **10**, 23–63 (1957)
6. R. Finn, D. Gilbarg, Three-dimensional subsonic flows and asymptotic estimates for elliptic partial differential equations. Acta Math. **98**, 265–296 (1957)
7. G.-C. Dong, *Nonlinear Partial Differential Equations of Second Order* (AMS, Providence, 1991)
8. G.-C. Dong, B. Ou, Subsonic flows around a body in space. Commun. Partial Differ. Equ. **18**, 355–379 (1993)
9. L.-L. Du, Z.-P. Xin, W. Yan, Subsonic flows in a multidimensional nozzle. Arch. Ration. Mech. Anal. **201**, 965–1012 (2011)
10. L. Liu, H.-R. Yuan, Steady subsonic potential flows through infinite multi-dimensional largely-open nozzles. Calc. Var. **49**, 1–36 (2014)
11. X. Gu, T.-Y. Wang, On subsonic and subsonic-sonic flows with general conservatives force in exterior domains. Acta Math. App Sinica (To appear)
12. X. Gu, T.-Y. Wang, On subsonic and subsonic-sonic flows in the infinity long Nozzle with general conservatives force. Acta Math. Sci. Ser. B **37**, 752–767 (2017)
13. C.-J. Xie, Z.-P. Xin, Existence of global steady subsonic Euler flows through infinitely long nozzles. SIAM J. Math. Anal. **42**, 751–784 (2010)

14. C. Chen, C.-J. Xie, Existence of steady subsonic Euler flows through infinitely long periodic nozzles. J. Differ. Equ. **252**, 4315–4331 (2012)
15. L.-L. Du, B. Duan, Global subsonic Euler flows in an infinitely long axisymmetric nozzle. J. Differ. Equ. **250**, 813–847 (2011)
16. G.-Q. Chen, X. Deng, W. Xiang, Global steady subsonic flows through infinitely long nozzles for the full Euler equations. SIAM J. Math. Anal. **44**, 2888–2919 (2012)
17. M. Bae, Stability of contact discontinuity for steady Euler system in the infinite duct. Z. Angew Math. Phys. **64**, 917–936 (2013)
18. B. Duan, Z. Luo, Three-dimensional full Euler flows in axisymmetric nozzles. J. Differ. Equ. **254**, 2705–2731 (2013)
19. L.-L. Du, C.-J. Xie, On subsonic Euler flows with stagnation points in two-dimensional nozzles. Indiana Univ. Math. J. **63**(5), 1499–1523 (2014)
20. C. Chen, Subsonic non-isentropic ideal gas with large vorticity in nozzles. Math. Method Appl. Sci. (2015)
21. L.-L. Du, B. Duan, Subsonic Euler flows with large vorticity through an infinitely long axisymmetric nozzle. J. Math. Fluid Mech. 1–20 (2016)
22. L.-L. Du, C.-J. Xie, Z.-P. Xin, Steady subsonic ideal flows through an infinitely long nozzle with large vorticity. Commun. Math. Phys. **328**, 327–354 (2014)
23. B. Duan, Z. Luo, Subsonic non-isentropic Euler flows with large vorticity in axisymmetric nozzles. J. Math. Anal. Appl. **430**, 1037–1057 (2015)
24. C. Chen, L.-L. Du, C. Xie, Z.-P. Xin, Two dimensional subsonic Euler flows past a wall or a symmetric body. Arch. Ration. Mech. Anal. **221**(2), 559–602 (2016)
25. J. Chen, Subsonic flows for the full Euler equations in half plane. J. Hyperbolic Differ. Equ. **6**(02), 207–228 (2009)
26. C.S. Morawetz, On a weak solution for a transonic flow problem. Commun. Pure Appl. Math. **38**, 797–818 (1985)
27. C.S. Morawetz, On steady transonic flow by compensated compactness. Methods Appl. Anal. **2**, 257–268 (1995)
28. G.-Q. Chen, M. Slemrod, D.-H. Wang, Vanishing viscosity method for transonic flow. Arch. Ration. Mech. Anal. **189**, 159–188 (2008)
29. G.-Q. Chen, C.M. Dafermos, M. Slemrod, D.-H. Wang, On two-dimensional sonic-subsonic flow. Commun. Math. Phys. **271**, 635–647 (2007)
30. C.-J. Xie, Z.-P. Xin, Global subsonic and subsonic-sonic flows through infinitely long nozzles. Indiana Univ. Math. J. **56**, 2991–3023 (2007)
31. C.-J. Xie, Z.-P. Xin, Global subsonic and subsonic-sonic flows through infinitely long axially symmetric nozzles. J. Differ. Equ. **248**, 2657–2683 (2010)
32. F.-M. Huang, T.-Y. Wang, Y. Wang, On multidimensional sonic-subsonic flow. Acta Math. Sci. Ser. B **31**, 2131–2140 (2011)
33. G.-Q. Chen, F.-M. Huang, T.-Y. Wang, Sonic-subsonic limit of approximate solutions to multidimensional steady Euler equations. Arch. Ration. Mech. Anal. **219**, 719–740 (2016)
34. N. Masmoudi, Asymptotic problems and compressible-incompressible limit, *Advances in Mathematical Fluid Mechanics* (Springer, Berlin, 2000), pp. 119–158
35. N. Masmoudi, Examples of singular limits in hydrodynamics, *Handbook of Differential Equations: Evolutionary Equations*, vol. 3 (Amsterdam, London, 2007), pp. 195–275
36. P.-L. Lions, N. Masmoudi, On a free boundary barotropic model. Ann. Inst. H. Poincaré Anal. Non Linéaire **16**, 373–410 (1999)
37. S. Labbé, E. Maitre, A free boundary model for Korteweg fluids as a limit of barotropic compressible Navier-Stokes equations. Method Appl. Anal. **20**, 165–178 (2013)
38. G.-Q. Chen, F.-M. Huang, T.-Y. Wang, W. Xiang, Incompressible limit of solutions of multidimensional steady compressible Euler equations. Z. Angew. Math. Phys. **67**(3), 1–18 (2016)
39. J. Ball, A version of the fundamental theorem of Young measures, *PDEs and Continuum Models of Phase Transitions*. Lecture Notes in Physics (Springer, Berlin, 1989), pp. 207–215
40. G.-Q. Chen, H. Frid, Divergence-measure fields and hyperbolic conservation laws. Arch. Ration. Mech. Anal. **147**, 89–118 (1999)

41. R.J. DiPerna, Compensated compactness and general systems of conservation laws. Trans. Am. Math. Soc. **292**, 383–420 (1985)
42. F. Murat, Compacite par compensation. Ann. Suola Norm. Pisa **4**(5), 489–507 (1978)
43. D. Serre, *Systems of Conservation Laws*, vol. 1–2 (Cambridge University Press, Cambridge, 1999, 2000)
44. L. Tartar, Compensated compactness and applications to partial differential equations, *Nonlinear Analysis and Mechanics: Herriot-Watt Symposium*, vol. 4, ed. by R.J. Knops (Pitman Press, New Jersey, 1979)
45. L.C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations* (American Mathematical Society, Providence, 1990)
46. J.-L. Lions, *Quelques Méthodes De Résolution Des Problémes Aux Limites Non Linéaires* (Dunod, Paris, 1969)
47. R.I. Kachurovskii, Non-linear monotone operators in Banach spaces J. Russ. Math. Surv. **23**(2), 117–165 (1968)
48. G.-Q. Chen, F.-M. Huang, T.-Y. Wang, W. Xiang, Steady Euler flows with large vorticity and characteristic discontinuities in arbitrary infinitely long nozzles. arXiv:1712.08605
49. X.-M. Deng, T.-Y. Wang, W. Xiang, Three-dimensional full Euler flows with nontrivial swirl in axisymmetric nozzles. Accepted by SIAM J. Math. Anal

A Constraint-Preserving Finite Difference Method for the Damped Wave Map Equation to the Sphere

Franziska Weber

Abstract We present and analyze a constraint-preserving finite difference method for approximating the damped wave map equation

$$\varepsilon u_{tt} + \alpha u_t - \Delta u = \gamma u, \quad |u| = 1, \quad \text{in } (0,\infty) \times \Omega,$$

into the sphere. The numerical method preserves a discrete version of the energy balance associated with the equation and the unit length constraint of the solution at every grid point. We show that the approximations converge to a weak solution as the discretization parameters go to zero and present some numerical experiments investigating the limit $\varepsilon \to 0$ for $\alpha = 1$.

Keywords Wave map equation · Finite difference method · Convergence Constraint preserving

1 Introduction

The Ericksen–Leslie equations

$$\operatorname{div} v = 0, \tag{1a}$$

$$\partial_t v + (v \cdot \nabla) v = F + \operatorname{div} \sigma, \tag{1b}$$

$$\rho_1 \left(\partial_t \omega + (v \cdot \nabla)\omega\right) = \rho_1 G + g + \operatorname{div} \pi, \tag{1c}$$

$$\sigma = -p\mathbb{1} - \frac{\partial W}{\partial \nabla u} \nabla u + \widehat{\sigma}, \tag{1d}$$

$$\pi = \beta \otimes u + \frac{\partial W}{\partial \nabla u}, \tag{1e}$$

$$g = \gamma u - \nabla u \beta - \frac{\partial W}{\partial u} + \widehat{g}, \tag{1f}$$

F. Weber (✉)
ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
e-mail: fraenschii@gmail.com

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_48

where $\mathbb{1}$ is the identity matrix, v is the fluid velocity, and $\omega = \partial_t u + (v \cdot \nabla)u$ is the material derivative of the director field u with $|u| = 1$, are a well-known model for simulating the dynamics of nematic liquid crystals [1]. Here, β, γ, ρ_1 are physical constants, p the pressure, F an external body force, G an external director body force, g the intrinsic force associated with the director, $\widehat{g}$ the kinematic transport of the director, and W the Oseen–Frank energy functional. From [1, 2], we have the following expressions for $\widehat{g}, \widehat{\sigma}$ and W:

$$\begin{aligned}
\widehat{g} &= \lambda_1 N + \lambda_2 A u, \\
\widehat{\sigma} &= \mu_1 (u^\top A u) u \otimes u + \mu_2 N \otimes u + \mu_3 u \otimes N + \mu_4 A + \mu_5 A u \otimes u + \mu_6 u \otimes A u, \\
W(u, \nabla u) &= \frac{k_1}{2} (\operatorname{div} u)^2 + \frac{k_2}{2} |u \cdot \operatorname{curl} u|^2 + \frac{k_3}{2} |u \times \operatorname{curl} u|^2 \\
&\quad + \frac{k_2 + k_4}{2} \operatorname{div}((\nabla u) u - (\operatorname{div} u) u), \\
A &= \frac{\nabla v + (\nabla v)^\top}{2}, \quad N = \omega - \frac{\nabla v - (\nabla v)^\top}{2} u.
\end{aligned}$$

$\mu_1, \ldots, \mu_6$ are called the Leslie coefficients and satisfy certain relations [1, 2]. k_1, k_2, and k_3 in the Oseen–Frank energy correspond to different orientations of the liquid crystal director u [1].

In many applications, the inertial constant ρ_1 is rather small in comparison with the other parameters, and therefore, the terms involving it are neglected for practical applications. In this short note, we would like to investigate the effect of this inertial term, as $\rho_1 \to 0$ for the simplified model of the damped wave map equation

$$\varepsilon u_{tt} + \alpha u_t - \Delta u = \gamma u, \quad |u| = 1, \quad \text{in } (0, \infty) \times \Omega. \tag{2}$$

Here, γ is the Lagrange multiplier enforcing the constraint $|u| = 1$ and the fluid velocity v and external force terms have been set to zero. Moreover, we have set $k_1 = k_2 = k_3$, the so-called *one constant approximation*. $\Omega \subset \mathbb{R}^n$, $n = 2, 3$ will either be the unit box $[0, 1]^n$ or the torus $\mathbb{T}^n$. In the first case, we use Neumann boundary conditions and in the second case periodic boundary conditions. We reformulate this equation in terms of the angular momentum $w = u_t \times u$,

$$u_t = u \times w, \tag{3a}$$

$$\varepsilon w_t = \Delta u \times u - \alpha w. \tag{3b}$$

and then construct a finite difference method that is stable for any choice of parameters $\varepsilon \geq 0$ and $\alpha > 0$ for this system. The numerical method is based on ideas from [3], and the approximations satisfy a discrete version of the energy balance

$$\frac{d}{dt} \int_\Omega \varepsilon |u_t|^2 + |\nabla u|^2 \, dx = -\alpha \int_\Omega |u_t|^2 \, dx,$$

of the system (2) and can be shown to converge for any $\varepsilon \geq 0$ and $\alpha > 0$. We then do some numerical tests for some choices of initial data and various values of ε to investigate the limit $\varepsilon \to 0$. We find that for the smooth data, the approximations u_h^ε to the damped wave map equation converge at a rate of about 4 to the approximation of the solution of the heat map flow. For example, with initial data that exhibits blowup for the wave map equation, the convergence seems very low and is only observable for very small ε. This indicates that there might not be a convergence rate for convergence in the L^2- or energy norm.

2 The Numerical Method

The numerical method we present here is based on the reformulation (3a)–(3b) of (2) which can be derived as follows: To obtain the first equation, we take the cross product of u with w, insert the definition of w, and use vector identities:

$$u \times w = u \times (u_t \times u) = (u \cdot u)u_t - (u_t \cdot u)u = u_t,$$

using the constraint $|u| = 1$ and that u_t is orthogonal to u. To derive the w-equation, take the time derivative of w and insert the definition of w and then the equation for u:

$$\varepsilon w_t = \varepsilon(u_{tt} \times u + u_t \times u_t) = \varepsilon u_{tt} \times u = \Delta u \times u - \alpha u_t \times u + \gamma u \times u = \Delta u \times u - \alpha w.$$

2.1 Discretization of the Domain and the Differential Operators

Let $M \in \mathbb{N}$ be the number of grid points in each dimension and $N := M^n$ the total number of grid cells, where $n = 2, 3$ is the spatial dimension. We outline the definition of the numerical method for $n = 3$, and the modifications needed for $n = 2$ are straightforward. Then, we set $h = 1/M$ the mesh width and $\Delta t > 0$ the time step size. The conditions on Δt that are needed to obtain convergence of the method will be determined later.

We define grid points and grid cells

$$\begin{aligned} \mathscr{C}_{i_1,i_2,\ldots,i_n} &:= ((i_1 - 1)h, i_1 h] \times \cdots \times ((i_n - 1)h, i_n h], \\ x_{i_1,\ldots,i_n} &:= ((i_1 - 1/2)h, \ldots, (i_n - 1/2)h), \end{aligned}$$

and time steps $t^m := m\Delta t, m = 0, \ldots, N_t$. To simplify notation, we introduce the multi-index $\underline{i} \in \mathscr{I}_N := \{0, \ldots, M\}^n$, such that $\underline{i} = (i_1, \ldots, i_n)$, and we can write

$$\mathscr{C}_{\underline{i}} = \mathscr{C}_{i_1,\ldots,i_n}, \quad x_{\underline{i}} = x_{i_1,\ldots,i_n}.$$

We will approximate u and w at the cell midpoints $x_{\underline{i}}$. Specifically,

$$u^m_{\underline{i}} \approx u(m\Delta t, x_{\underline{i}}), \quad w^m_{\underline{i}} \approx w(m\Delta t, x_{\underline{i}}).$$

Next, let $\mathbf{e}_1 := (1, 0, 0)$, $\mathbf{e}_2 := (0, 1, 0)$, and $\mathbf{e}_3 := (0, 0, 1)$. Using these vectors, we then define the forward and backward difference operators

$$D^+_j u_{\underline{i}} = \frac{u_{\underline{i}+\mathbf{e}_j} - u_{\underline{i}}}{h}, \qquad D^-_j u_{\underline{i}} = D^+_j u_{\underline{i}-\mathbf{e}_j},$$

respectively, for $j = 1, 2, 3$, and $\underline{i} \in \mathscr{I}_N$. The discrete Laplacian Δ_h is then defined as

$$\Delta_h u_{\underline{i}} = \sum_{j=1}^{3} D^+_j D^-_j u_{\underline{i}}.$$

If we furthermore introduce the backward gradient $\nabla_h = [D^-_1, D^-_2, D^-_3]^T$ and forward divergence $\operatorname{div}_h v = D^+_1 v^{(1)} + D^+_2 v^{(2)} + D^+_3 v^{(3)}$, we have the identity

$$\operatorname{div}_h \nabla_h = \Delta_h,$$

which will be convenient in the upcoming analysis.

For the time discretization, we will use the notation

$$u^{m+1/2} := \frac{u^m + u^{m+1}}{2}, \quad D^+_t u^m = \frac{u^{m+1} - u^m}{\Delta t}, \quad D^-_t u^m = \frac{u^m - u^{m-1}}{\Delta t}.$$

We approximate the initial conditions u^0 and w^0 as follows:

$$(u^0_{\underline{i}},\ w^0_{\underline{i}}) = \left(\Pi[u^0]_{\underline{i}}, \Pi[u^0_t]_{\underline{i}} \times u^0_{\underline{i}}\right), \quad \forall \underline{i},$$

where the projection operator Π is defined by

$$\Pi[f]_{\underline{i}} = \frac{1}{h^n} \int_{(i_1-1/2)h}^{(i_1+1/2)h} \cdots \int_{(i_n-1/2)h}^{(i_n+1/2)h} f(y)\, dy.$$

2.2 Definition of the Finite Difference Scheme

We are now ready to state the new method.

Definition 1. Given initial data $u^0 \in H^1(\Omega)$, $u_t^0 \in L^2(\Omega)$, let

$$(u_{\underline{i}}^0,\ w_{\underline{i}}^0) = \left(\Pi[u^0]_{\underline{i}},\ \Pi[u_t^0]_{\underline{i}} \times u_{\underline{i}}^0\right), \quad \forall \underline{i}.$$

Determine $(u_{\underline{i}}^m, w_{\underline{i}}^m)$, $\forall \underline{i} \in \mathscr{I}_N$, $m = 1, \ldots,$ sequentially, by solving the nonlinear system

$$D_t^+ u_{\underline{i}}^m = u_{\underline{i}}^{m+1/2} \times w_{\underline{i}}^{m+1/2}, \tag{4a}$$

$$\varepsilon D_t^+ w_{\underline{i}}^m = \Delta_h u_{\underline{i}}^{m+1/2} \times u_{\underline{i}}^{m+1/2} - \alpha w_{\underline{i}}^{m+1/2}. \tag{4b}$$

In the following, we will prove that the approximations computed using this method inherit discrete versions of some fundamental properties that the continuous system (3a)–(3b) satisfies. To this end, it will be convenient to extend the numerical solution to all of Ω. For this purpose, we shall use the piecewise constant extensions:

$$\begin{aligned} &u_h^m(x) = u_{\underline{i}}^m, \quad x \in \mathscr{C}_{\underline{i}}; \quad u_h(t,x) = u_h^m(x), \quad t \in (t^{m-1}, t^m]; \\ &w_h^m(x) = w_{\underline{i}}^m, \quad x \in \mathscr{C}_{\underline{i}}; \quad w_h(t,x) = w_h^m(x), \quad t \in (t^{m-1}, t^m]; \\ &\overline{u}_h(t,x) = u_h^{m-1/2}(x); \quad \overline{w}_h(t,x) = w_h^{m-1/2}(x), \quad t \in (t^{m-1}, t^m]. \end{aligned} \tag{5}$$

Observe that the numerical method can then be written

$$D_t^+ u_h^m = u_h^{m+1/2} \times w_h^{m+1/2}, \tag{6a}$$

$$\varepsilon D_t^+ w_h^m = \Delta_h u_h^{m+1/2} \times u_h^{m+1/2} - \alpha w_h^{m+1/2}, \tag{6b}$$

where Δ_h is derived in the obvious way.

Lemma 1. *There exists a unique numerical solution to the method posed in Definition 1. Moreover, the length is preserved*

$$|u_{\underline{i}}^m| = |u_{\underline{i}}^0| = 1, \quad \forall \underline{i}, \quad m = 0, \ldots, \tag{7}$$

and we have the discrete energy law for all $m = 0, 1, \ldots$:

$$\int_\Omega \varepsilon |w_h^{m+1}|^2 + |\nabla_h u_h^{m+1}|^2 \, dx + 2\Delta t \alpha \int_\Omega |w_h^{m+1/2}|^2 \, dx = \int_\Omega \varepsilon |w_h^m|^2 + |\nabla_h u_h^m|^2 \, dx. \tag{8}$$

Proof. The existence of a unique solution is proved using a convergent fixed point iteration. In particular, one shows that the following iterative scheme

Definition 2. Given $h > 0$, $\Delta t > 0$ satisfying

$$\frac{\Delta t}{\sqrt{2\varepsilon + \alpha \Delta t}} \leq \kappa h, \tag{9}$$

and functions (u_h^m, w_h^m) satisfying (4a)–(4b), we approximate the next time step (u_h^{m+1}, w_h^{m+1}) to a given tolerance $\tau > 0$ by the following procedure: Set

$$(u_h^{m,0}, w_h^{m,0}) = (u_h^m, w_h^m),$$

and iteratively define $(u_h^{m,s+1}, w_h^{m,s+1})$, $s = 0, 1, \ldots$ by

$$\begin{aligned}
\frac{u_h^{m,s+1} - u_h^m}{\Delta t} &= \frac{1}{2}\left(u_h^m + u_h^{m,s+1}\right) \times \frac{1}{2}\left(w_h^m + w_h^{m,s}\right), \\
\varepsilon \frac{w_h^{m,s+1} - w_h^m}{\Delta t} &= \frac{1}{2}\left(\Delta_h u_h^m + \Delta_h u_h^{m,s+1}\right) \times \frac{1}{2}\left(u_h^m + u_h^{m,s+1}\right) - \frac{\alpha}{2}\left(w_h^m + w_h^{m,s+1}\right),
\end{aligned}$$

until the following stopping criteria are met:

$$\sqrt{\alpha \Delta t + 2\varepsilon}\left\| w_h^{m,s+1} - w_h^{m,s} \right\|_{L^2(\Omega)} + \left\| \nabla u_h^{m,s+1} - \nabla u_h^{m,s} \right\|_{L^2(\Omega)} < \tau.$$

terminates in $\mathcal{O}(N(|\log(\tau)| + |\log(N)|))$ operations for arbitrary tolerances $\tau > 0$. A fixed point of this scheme is a solution of (6a)–(6b). The proof uses ideas from [3, Theorem 4.2] and might be detailed in a future work.

That the length is conserved, (7), follows immediately from (4a). Indeed, taking the dot product of the first Eq. (6a) with $u_h^{m+1/2}$ yields

$$D_t^+ u_h^m \cdot u_h^{m+1/2} = 0,$$

thanks to the orthogonality properties of the cross product. To prove (8), we denote the energy

$$E_m := \frac{1}{2}\int_\Omega |\nabla_h u_h^m|^2 + \varepsilon |w_h^m|^2 \, dx.$$

and calculate

$$\begin{aligned}
D_t^+ E_m &= \int_\Omega \varepsilon w_h^{m+1/2} \cdot D_t^+ w_h^m - \Delta_h d_h^{m+1/2} \cdot D_t^+ d_h^m \, dx \\
&= \int_\Omega \left(\Delta_h u_h^{m+1/2} \times u_h^{m+1/2}\right) \cdot w_h^{m+1/2} \, dx - \alpha \int_\Omega |w_h^{m+1/2}|^2 \, dx \\
&\quad - \int_\Omega \left(u_h^{m+1/2} \times w_h^{m+1/2}\right) \cdot \Delta_h u_h^{m+1/2} \, dx \\
&= -\alpha \int_\Omega |w_h^{m+1/2}|^2 \, dx.
\end{aligned}$$

This concludes the proof. □

Remark 1. The condition (9) is only needed to obtain convergence of the fixed point iteration. The numerical scheme (4a)–(4b) is stable and converges for any choice $\varepsilon \geq 0, \alpha > 0$ as $h, \Delta t \to 0$, independently of the relation between h and Δt as we will see in Sect. 3.

3 Convergence

Next, we prove that the numerical approximations computed by the scheme (4a)–(4b) converge to a weak solution of (2) as the discretization parameters go to zero. But first, we need to define a suitable notion of weak solution. Due to the presence of the Lagrange multiplier γ, weak solutions are defined using the angular momentum w. Specifically, since $\gamma = |\nabla u|^2 - \varepsilon |u_t|^2$, the energy only provides an L^1 bound on γ. For this reason, the weak formulation of (2) is often posed using (3b) and the following integration by parts formula:

Lemma 2. *For all sufficiently smooth functions* (u, ϕ)*, there holds*

$$\int_\Omega (u \times \Delta u)\phi \, dx = \int_\Omega (\nabla u \times u) : \nabla \phi \, dx. \tag{10}$$

This simple lemma has been proved in [3, Lemma 3.1].

Remark 2. One can readily derive a discrete version of (10) for the operators Δ_h, ∇_h, and div_h since the proof only relies on the identity $\mathrm{div}\,(\nabla a\ b) = b\Delta a + \nabla a \cdot \nabla b$, which is satisfied by the numerical operators.

The weak formulation of (2) is given by the following definition. We refer to [4] for more on this formulation and the corresponding existence theory.

Definition 3. Given initial data $u^0 \in H^1(\Omega)$, $u_t^0 \in L^2(\Omega)$, with finite energy

$$E(u^0) := \frac{1}{2}\left(\varepsilon \|u_t^0\|_{L^2(\Omega)}^2 + \|\nabla u^0\|_{L^2(\Omega)}^2\right) \leq C,$$

and $|u^0| = 1$ a.e., we call u a weak solution of (2) provided:

1. The energy satisfies

$$E(u(t)) + \alpha \int_0^t \|u_s(s)\|_{L^2(\Omega)}^2 \, ds \leq E(u^0). \tag{11}$$

2. The following weak formulation holds for all $\phi \in C_c^\infty([0, \infty) \times \Omega)$,

$$\int_0^\infty \int_\Omega -\varepsilon(u_t \times u)\phi_t + (\nabla u \times u) : \nabla\phi + \alpha(u_t \times u)\phi \, dxdt = \varepsilon \int_\Omega (u_t^0 \times u^0)\phi(0,\cdot)\, dx. \tag{12}$$

3. The initial condition is satisfied, i.e., as $t \to 0$,

$$u(t) \to u^0 \text{ in } H^1(\Omega), \qquad u_t(t) \rightharpoonup u_t^0 \text{ in } L^2(\Omega)$$

Having defined a notion of weak solution, we proceed to prove that our scheme (4a)–(4b) converges to one. Our main result in this section is the following convergence result:

Theorem 1. *Let $\alpha > 0$ and $\varepsilon \geq 0$, and let $\{(u_h, w_h)\}_{h>0}$ be a sequence of numerical approximations obtained using Definition 1 and (5). Then, as $h \to 0$, $u_h \to u$ a.e. and in $L^p((0,\infty)\times\Omega)$ for any $p < \infty$, $\sqrt{\varepsilon} w_h \overset{\star}{\rightharpoonup} \sqrt{\varepsilon} w$ in $L^\infty(0,T;L^2(\Omega))$ and moreover $\overline{w}_h \rightharpoonup w$ in $L^2([0,T]\times\Omega)$ for any $\varepsilon \geq 0$ and $\alpha > 0$, where*

$$\begin{aligned} |u| &= 1, \text{ a.e. in } [0,\infty)\times\Omega, \\ w &= u_t \times u, \text{ a.e. in } [0,\infty)\times\Omega, \end{aligned}$$

Furthermore, u is a weak solution of the damped wave map equation (2) in the sense of Definition 3.

Proof. To prove this theorem, our starting point is Lemma 1 yielding the h-uniform bounds for any $T > 0$:

$$\begin{aligned} D_t^+ u_h &\subset L^2(0,T;L^2(\Omega)), \\ \nabla_h u_h &\subset L^\infty(0,T;L^2(\Omega)), \\ \sqrt{\varepsilon} w_h &\subset L^\infty(0,T;L^2(\Omega)), \\ \overline{w}_h &\subset L^2((0,T)\times\Omega). \end{aligned}$$

From these bounds, we can assert the existence of functions u and w, and a subsequence h_j, such that

$$\begin{aligned} \sqrt{\varepsilon} w_{h_j} &\overset{\star}{\rightharpoonup} \sqrt{\varepsilon} w \text{ in } L^\infty(0,T;L^2(\Omega)), \\ \overline{w}_{h_j} &\rightharpoonup w \text{ in } L^2(0,T;L^2(\Omega)), \\ D_t^+ u_{h_j} &\rightharpoonup u_t \text{ in } L^2(0,T;L^2(\Omega)), \\ \nabla_{h_j} u_{h_j} &\overset{\star}{\rightharpoonup} \nabla u \text{ in } L^\infty(0,T;L^2(\Omega)), \\ u_{h_j} &\to u \text{ a.e. and in } L^p((0,T)\times\Omega) \text{ for } p < \infty, \end{aligned} \tag{13}$$

where the limit u also satisfies the constraint $|u(t,x)| = 1$ a.e. in $[0,T)\times\Omega$, thanks to the strong convergence. The energy inequality (11) follows from the discrete energy balance (8) when passing to the limit $h \to 0$ and using the weak lower semi-continuity of the L^2-norm.

Next, we show that the limit (u, w) satisfies the weak formulation (12). For test functions $\varphi, \psi \in C_0^1([0,T)\times\Omega;\mathbb{R}^n)$, we denote $\varphi^m(x) := \varphi(t^m, x)$, $\psi^m(x) := \psi(t^m, x)$. Then, we take the dot product of (6a) and (6b) with φ^m, ψ^m, integrate over Ω, and sum over m, to discover

$$\Delta t \sum_{m=0}^{\infty} \int_\Omega \left(D_t^+ u_h^m - u_h^{m+1/2} \times w_h^{m+1/2} \right) \cdot \varphi^m \, dx = 0,$$

$$\Delta t \sum_{m=0}^{\infty} \int_\Omega \left(\varepsilon D_t^+ w_h^m + u_h^{m+1/2} \times \Delta_h u_h^{m+1/2} + \alpha w_h^{m+1/2} \right) \cdot \psi^m \, dx = 0.$$

Using Lemma 2 (see Remark 2) and summation by parts, we deduce that

$$\Delta t \sum_{m=0}^{\infty} \int_\Omega \left(-u_h^{m+1} \cdot D_t^+ \varphi^m - \left(u_h^{m+1/2} \times w_h^{m+1/2} \right) \cdot \varphi^m \right) dx = \int_\Omega u_h^0 \cdot \varphi^0 \, dx,$$

$$\Delta t \sum_{m=0}^{\infty} \int_\Omega \Big(-\varepsilon w_h^{m+1} \cdot D_t^+ \psi^m - \left(\nabla_h u_h^{m+1/2} \times u_h^{m+1/2} \right) : \nabla_h \psi^m + \alpha w_h^{m+1/2} \cdot \psi^m \Big) dx = \varepsilon \int_\Omega w_h^0 \cdot \psi^0 \, dx.$$

Using the notation (5), this becomes

$$-\int_0^\infty \int_\Omega \left(u_h \cdot D_t^+ \varphi + (\overline{u}_h \times \overline{w}_h) \cdot \varphi \right) dx\, dt - \int_\Omega u_h^0 \cdot \varphi(0,\cdot)\, dx = 0,$$

$$-\int_0^\infty \int_\Omega \left(\varepsilon w_h \cdot D_t^+ \psi + (\nabla_h \overline{u}_h \times \overline{u}_h) : \nabla_h \psi - \alpha \overline{w}_h \cdot \psi \right) dx - \varepsilon \int_\Omega w_h^0 \cdot \psi(0,\cdot)\, dx = 0.$$

Now, since $\nabla_h \psi \to \nabla \psi$ a.e. and $(D_t^+ \varphi, D_t^+ \psi) \to (\varphi_t, \psi_t)$ a.e., one may apply the convergence statements (13) to discover that the limit (u, w) satisfies

$$\begin{aligned} -\int_0^\infty \int_\Omega (u \cdot \varphi_t + (u \times w) \cdot \varphi)\, dx\, dt - \int_\Omega u^0 \cdot \varphi(0,\cdot)\, dx = 0, \\ -\int_0^\infty \int_\Omega (\varepsilon w \cdot \psi_t + (\nabla u \times u) : \nabla \psi - \alpha w \cdot \psi)\, dx - \int_\Omega (u_t^0 \times u^0) \cdot \psi(0,\cdot)\, dx = 0. \end{aligned} \tag{14}$$

It only remains to prove that this formulation is equivalent to (12) in Definition 3. In practice, this means proving that $w = u_t \times u$ since then the second equation in (14) becomes (12). We first note that by approximation, the weak formulation (14) also holds for test functions $\varphi, \psi \in H^1((0,T)\times\Omega)$ with compact support in $[0,T)$. Moreover, since $u_t \in L^2(0,T;L^2(\Omega))$ by the energy balance, we can move the time derivative in the first equation in (14) back onto u. Then, testing with $\varphi = w\eta$ and $\psi = u\eta$ for some scalar smooth function $\eta \in C_c^1([0,T)\times\Omega)$, we obtain after some algebra

$$\int_0^\infty \int_\Omega (u_t \cdot w\eta - (u \times w) \cdot w\eta)\, dx\, dt - \int_\Omega u^0 \cdot w^0 \eta(0,\cdot)\, dx$$
$$= \int_0^\infty \int_\Omega (u_t \cdot w\eta)\, dx\, dt = 0,$$
$$- \int_0^\infty \int_\Omega (\varepsilon w \cdot (u\eta)_t + (\nabla u \times u) : \nabla(u\eta) - \alpha w \cdot u\eta)\, dx - \int_\Omega (u_t^0 \times u^0) \cdot u^0 \eta(0,\cdot)\, dx$$
$$= - \int_0^\infty \int_\Omega (\varepsilon w \cdot (u\eta)_t - \alpha w \cdot u\eta)\, dx = 0.$$

since $u_0 \cdot w_0 = 0$ by definition of w_0 and using some vector algebra identities. If $\varepsilon = 0$, the second equation yields $u \cdot w = 0$ almost everywhere. If $\varepsilon \neq 0$, we divide the second equation by ε and add the two:

$$\int_0^\infty \int_\Omega \left(u_t \cdot w\eta - w \cdot (u\eta)_t + \frac{\alpha}{\varepsilon} w \cdot u\eta\right) dx\, dt$$
$$= \int_0^\infty \int_\Omega \left(-w \cdot u\eta_t + \frac{\alpha}{\varepsilon} w \cdot u\eta\right) dx\, dt$$
$$= 0.$$

Since $(u \cdot w) \in L^\infty(0, T; X) \subset L^1(0, T; X)$ for the Banach space $X = L^2(\Omega)$, we can apply Lemma 1.1 in [5, p. 250] to obtain that $(u \cdot w)(t)$ is absolutely continuous in $L^2(\Omega)$ and $(u \cdot w)(t) = \xi - {}^{\alpha}/_{\varepsilon} \int_0^t (u \cdot w)(s) ds$ for some $\xi \in L^2(\Omega)$. Moreover, we obtain from the energy inequality (11) and the weak lower semi-continuity of the L^2-norm that $\sqrt{\varepsilon} w(t) \to \sqrt{\varepsilon} w^0$ in $L^2(\Omega)$. Hence, $\xi = (u_0 \cdot w_0) = 0$. Now, Grönwall inequality for $(u \cdot w)$ and $-(u \cdot w)$ allows us to conclude that $(u \cdot w)(t) = 0$ almost everywhere.

By the definition of weak derivatives, the first equation in (14) tells us that

$$u_t = u \times w \text{ a.e in } (0, T) \times \Omega.$$

Since $|u| = 1$, this means that

$$w = u_t \times u + (u \cdot w)u = u_t \times u,$$

using the just established identity. Hence, the weak formulation (12) holds. The strong continuity of $u(t)$ in $H^1(\Omega)$ at zero (Point 3 in Definition 3) follows from the weak lower semi-continuity of the L^2-norm and the energy inequality. □

4 Numerical Examples

One can show that as $\varepsilon \to 0$, the sequence of solutions $(u^\varepsilon, w^\varepsilon)$ of the damped wave map equation (3a)–(3b) converges weakly to a solution of the heat map flow. It seems, however, hard to prove a convergence rate in ε. Moreover, if Ω is a domain in $\mathbb{R}^2$,

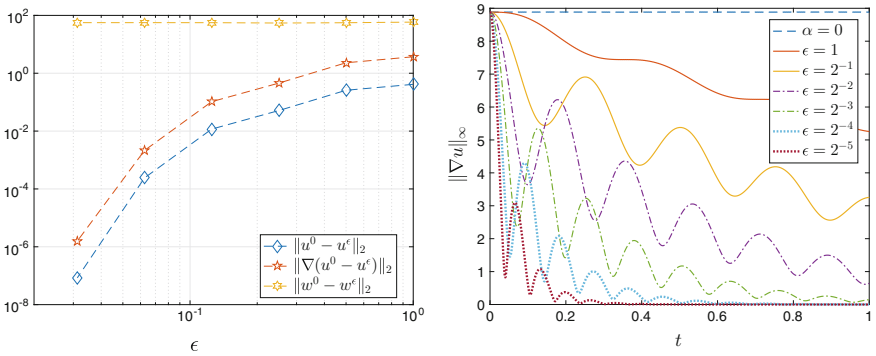


Fig. 1 Left: Convergence of u_h^ε and w_h^ε toward the solution of the heat map flow for smooth data. Right: Maximum of gradient $\nabla_h u_h^\varepsilon$ over time

there is evidence that solutions of the wave map equation may blow up [6] whereas solutions of the heat map flow are locally H^2 [7]. We investigate this convergence rate numerically in two examples, one with smooth data and one with initial data that develops singularities for the wave map equation.

Smooth data We compute approximations to the damped heat map flow for the initial data

$$\begin{aligned}\phi(x, y, t) &= \sin(2\pi(\sqrt{2})t + (x + y)),\\ u^0(x, y) &= (\cos(\phi(x, y, 0)), \sin(\phi(x, y, 0)), 0),\\ w^0(x, y) &= (0, 0, -\phi_t(x, y, 0)),\end{aligned}$$

on $\Omega = [-1/2, 1/2]^2$ for $h = 2^{-7}$ for $\varepsilon = 1, 2^{-1}, \ldots, 2^{-5}$ and $\alpha = 1$ at time $T = 1$. In Fig. 1, left-hand side, the convergence rates of u_h^ε, $\nabla_h u_h^\varepsilon$, and w_h^ε in $L^2(\Omega)$ are shown. We observe that the variables u_h^ε and $\nabla_h u_h^\varepsilon$ converge at a rate of approximately 4 whereas the variable w_h^ε does not converge, which is expected since the bounds on the $L^\infty(0, T; L^2(\Omega))$-norm of w^ε are not uniform in ε. In the same figure on the right-hand side, the evolution of $\|\nabla_h u_h^\varepsilon(t)\|_\infty$ over time is shown and we see that as $\varepsilon \to 0$, the maximum of the gradient decreases which could be attributed to the damping term $-\alpha w$ in the equation.

Singular data Next, we compute approximations for the initial data

$$\begin{aligned}r(x, y) &= \sqrt{x^2 + y^2},\\ a(r) &= (1 - 2r)^4,\\ u^0(x, y) &= \begin{cases}(0, 0, -1), & r \geq 1/2,\\ (2xa, 2ya, a^2 - r^2)/(a^2 + r^2), & r < 1/2,\end{cases}\\ w^0(x, y) &= (0, 0, 0),\end{aligned}$$

which has been used in [3, 8] to show blowup of solutions of the wave map equation.

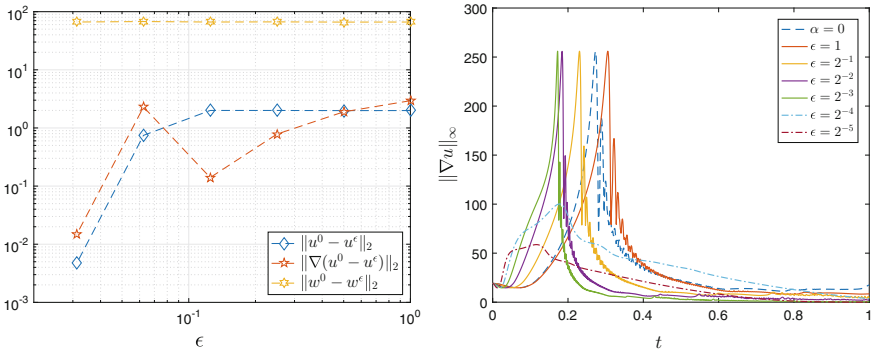


Fig. 2 Left: Convergence of u_h^ε and w_h^ε toward the solution of the heat map flow for initial data that develops singularities for the wave map equation. Right: Maximum of gradient $\nabla_h u_h^\varepsilon$ over time

Again, $\Omega = [-1/2, 1/2]^2$ and we use $h = 2^{-7}$ for $\varepsilon = 1, 2^{-1}, \ldots, 2^{-5}$ and $\alpha = 1$.

In this example, the distance between u^ε and the heat map reference solution does not seem to decrease except when ε is getting quite small. So, the convergence rate is either very small or there is no convergence in the L^2-norm (see Fig. 2). On the right-hand side of Fig. 2, we see that the gradient of u_h^ε appears to blow up for larger ε, while for smaller ε it does not. This might be the reason why the convergence is so bad. At smaller ε, the damping term appears to prevent the blowup of ∇u^ε.

Acknowledgements The author would like to thank Trygve K. Karper for insightful discussions on the subject.

References

1. I.W. Stewart, *The Static and Dynamic Continuum Theory of Liquid Crystals: A Mathematical Introduction* (CRC Press, Boca Raton, 2004)
2. H. Holden, Models for nematic liquid crystals. Unpublished manuscript, NTNU (2012)
3. T.K. Karper, F. Weber, A new angular momentum method for computing wave maps into spheres. SIAM J. Numer. Anal. **52**(4), 2073–2091 (2014)
4. J. Shatah, M. Struwe, *Geometric Wave Equations, Courant Lecture Notes in Mathematics*, vol. 2 (Courant Institute of Mathematical Sciences, American Mathematical Society, New York University, New York, Providence, 1998)
5. R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*, vol. 343 (American Mathematical Society, Providence, 2001)
6. P. Bizoń, T. Chmaj, Z. Tabor, Formation of singularities for equivariant $(2 + 1)$-dimensional wave maps into the 2-sphere. Nonlinearity **14**(5), 1041–1053 (2001)
7. Michael Struwe, On the evolution of harmonic mappings of Riemannian surfaces. Comment. Math. Helv. **60**(4), 558–581 (1985)
8. S. Bartels, X. Feng, A. Prohl, Finite element approximations of wave maps into spheres. SIAM J. Numer. Anal. **46**(1), 61–87 (2007/2008)

Integral Transform Approach to Solving Klein–Gordon Equation with Variable Coefficients

Karen Yagdjian

Abstract In this review, we present an integral transform that maps solutions of some class of the partial differential equations with time-independent coefficients to solutions of more complicated equations, which have time-dependent coefficients. We illustrate this transform by applications to model equations. In particular, we give applications to the Klein–Gordon and wave equations in the curved spacetimes such as the de Sitter universe.

Keywords Klein-Gordon equation · Curved spacetime · Black holes

1 Introduction

In this review, we present an integral transform that maps solutions of some class of the partial differential equations with time independent coefficients to solutions of more complicated equations, which have coefficients depending on time in some specific way. Consider for the smooth function $f = f(x,t)$ the solution $w = w(x,t;b)$ to the problem

$$w_{tt} - A(x,\partial_x)w = 0,\ \ w(x,0;b) = f(x,b),\ \ w_t(x,0;b) = 0,\ \ t \in [0,T_1] \subseteq \mathbb{R},\ \ x \in \Omega \subseteq \mathbb{R}^n, \tag{1}$$

with the parameter $b \in I = [t_0, T] \subseteq \mathbb{R}$, $t_0 < T \leq \infty$, and with $0 < T_1 \leq \infty$. Here Ω is a domain in $\mathbb{R}^n$, while $A(x,\partial_x)$ is the partial differential operator $A(x,\partial_x) = \sum_{|\alpha|\leq m} a_\alpha(x) D_x^\alpha$. For $M \in \mathbb{C}$, we are going to present the integral operator

$$\mathscr{K}[w](x,t) = 2\int_{t_0}^{t} db \int_0^{|\phi(t)-\phi(b)|} K(t;r,b;M)w(x,r;b)dr, \quad x \in \Omega,\ \ t \in I, \tag{2}$$

which maps the function $w = w(x,r;b)$ into solution $u = u(x,t)$ of the equation

K. Yagdjian (✉)
School of Mathematical and Statistical Sciences, University of Texas RGV,
1201 W. University Drive, Edinburg, TX 78539, USA
e-mail: karen.yagdjian@utrgv.edu

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_49

$$u_{tt} - a^2(t)A(x,\partial_x)u - M^2u = f, \qquad x \in \Omega\,,\ t \in I. \tag{3}$$

In fact, the function $u = u(x,t)$ takes initial values as follows

$$u(x,t_0) = 0,\ \ u_t(x,t_0) = 0, \quad x \in \Omega\,.$$

Here $\phi = \phi(t)$ is a distance function produced by $a = a(t)$, that is $\phi(t) = \int_{t_0}^{t} a(\tau)\,d\tau$, while $M \in \mathbb{C}$ is a constant. Moreover, we also give the corresponding operators, which generate solutions of the source-free equation and takes non-vanishing initial values. In the present review, we restrict ourselves to the smooth functions, but it is evident that similar formulas, with the corresponding interpretations, are applicable to the distributions as well. (For details see, e.g., [21].) In order to motivate our approach, we consider the solution $w = w(x,t;b)$ to the Cauchy problem

$$w_{tt} - \Delta w = 0,\ \ (t,x) \in \mathbb{R}^{1+n},\ \ w(x,0;b) = \varphi(x,b),\ \ w_t(x,0;b) = 0,\ \ x \in \mathbb{R}^n, \tag{4}$$

with the parameter $b \in I \subseteq \mathbb{R}$. We denote that solution by $w_\varphi = w_\varphi(x,t;b)$; if φ is independent of the second time variable b, then we write simply $w_\varphi(x,t)$. There are well-known explicit representation formulas for the solution of the problem (4).

The starting point of the integral transform approach suggested in [20] is the Duhamel's principle (see, e.g., [18]), which has been revised in order to prepare the ground for generalization. Our *first observation* is that the function

$$u(x,t) = \int_{t_0}^{t} db \int_{0}^{t-b} w_f(x,r;b)\,dr\,, \tag{5}$$

is the solution of the Cauchy problem $u_{tt} - \Delta u = f(x,t)$ in $\mathbb{R}^{n+1}$, and $u(x,t_0) = 0$, $u_t(x,t_0) = 0$ in $\mathbb{R}^n$, if the function $w_f = w_f(x;t;b)$ is a solution of the problem (4), where $\varphi = f$. The *second observation* is that in (5) the upper limit $t-b$ of the inner integral is generated by the propagation phenomena with the speed which equals to one. In fact, that is a distance function. Our *third observation* is that the solution operator $\mathscr{G} : f \longmapsto u$ can be regarded as a composition of two operators. The first one

$$\mathscr{WE} : f \longmapsto w$$

is a Fourier Integral Operator, which is a solution operator of the Cauchy problem for wave equation. The second operator

$$\mathscr{K} : w \longmapsto u$$

is the integral operator given by (5). We regard the variable b in (5) as a "subsidiary time." Thus, $\mathscr{G} = \mathscr{K} \circ \mathscr{WE}$. If we take into account the propagation cone by introducing the distance function $\phi(t)$, and if we provide the integral operator (5) with the kernel $K(t;r,b;M)$, as in (2), then we actually generate new representa-

tions for the solutions of different well-known equations with x-independent coefficients. Our *fourth observation* is that if we plug into (5) the solution $w = w(x; t; b)$ of the Dirichlet problem for the elliptic equation $w_{tt} + \Delta w = 0,\ (t, x) \in \mathbb{R}^{1+n}$, $w(x, 0; b) = f(x, b),\ x \in \mathbb{R}^n$, then the integral (5) defines the solution u of the equation $u_{tt} + \Delta u = f(x, t) + \int_{t_{in}}^{t} w_t(x; 0; b)\, db$, such that $u(x, t_{in}) = 0, u_t(x, t_{in}) = 0$.

In [26, 27], we extended the class of the equations for which we can obtain explicit representation formulas for the solutions, by varying the first mapping. More precisely, consider a solution $w = w_{A,\varphi}(x, t; b)$ to the problem (1) with the parameter $b \in I \subseteq \mathbb{R}$. If we have a resolving operator of the problem (1), then by applying (2), we can generate solutions of another equation. Thus, $\mathscr{G}_A = \mathscr{K} \circ \mathscr{E}\mathscr{E}_A$. The new class of equations contains operators with x-depending coefficients, and those equations are not necessarily hyperbolic.

That transform was used in a series of papers [6, 7, 9, 10, 20–27] to investigate in a unified way several equations such as the linear and semilinear Tricomi equations, Gellerstedt equation, the wave equation in Einstein-de Sitter spacetime, the wave and the Klein–Gordon equations in the de Sitter and anti-de Sitter spacetimes.

2 Linear Equations in the De Sitter Spacetime

Consider the Klein–Gordon equation in the de Sitter spacetime, that is $a(t) = e^{-t}$ in (3). Recently, the equations in the de Sitter and anti-de Sitter spacetimes became the focus of interest for an increasing number of authors (see, e.g., [1–4, 8, 14, 16, 17, 19, 25] and the bibliography therein).

We need the following notations. We define a *chronological future* $D_+(x_0, t_0)$ and a *chronological past* $D_-(x_0, t_0)$ of the point (x_0, t_0), $x_0 \in \mathbb{R}^n$, $t_0 \in \mathbb{R}$, as follows: $D_\pm(x_0, t_0) := \{(x, t) \in \mathbb{R}^{n+1}\,;\ |x - x_0| \leq \pm(e^{-t_0} - e^{-t})\}$. Then, for $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}$, $M \in \mathbb{C}$, we define the function

$$E(x, t; x_0, t_0; M) := 4^{-M} e^{M(t_0+t)} \Big((e^{-t_0} + e^{-t})^2 - (x - x_0)^2 \Big)^{M-\frac{1}{2}} \times F\Big(\frac{1}{2} - M, \frac{1}{2} - M; 1; \frac{(e^{-t_0} - e^{-t})^2 - (x - x_0)^2}{(e^{-t_0} + e^{-t})^2 - (x - x_0)^2} \Big), \quad (6)$$

where $(x, t) \in D_+(x_0, t_0) \cup D_-(x_0, t_0)$ and $F\big(a, b; c; \zeta\big)$ is the hypergeometric function. We use the notation $x^2 := |x|^2$ for $x \in \mathbb{R}^n$; the function E depends on $r^2 = (x - x_0)^2$, that is $E(x, t; x_0, t_0; M) = E(r, t; 0, t_0; M)$. According to Theorem 2.12 [26], the function $E(r, t; 0, t_0; M)$ solves the Klein–Gordon equation in the de Sitter spacetime:

$$E_{tt}(r, t; 0, t_0; M) - e^{-2t} E_{rr}(r, t; 0, t_0; M) - M^2 E(r, t; 0, t_0; M) = 0.$$

The kernels $K_0(z, t; M)$ and $K_1(z, t; M)$ are defined by

$$K_0(z,t;M) := -\left[\frac{\partial}{\partial b}E(z,t;0,b;M)\right]_{b=0}, \tag{7}$$

$$K_1(z,t;M) := E(z,t;0,0,M)\,. \tag{8}$$

Here $M \in \mathbb{C}$. From now on, we assume that $a_\alpha \in C^\infty(\Omega)$.

Theorem 1. [26] *For $f \in C(\Omega \times I)$, $I = [0,T]$, $0 < T \le \infty$, and φ_0, $\varphi_1 \in C(\Omega)$, let the function $v_f(x,t;b) \in C^{m,2,0}_{x,t,b}(\Omega \times [0, 1-e^{-T}] \times I)$ be a solution to the problem*

$$\begin{cases} v_{tt} - A(x,\partial_x)v = 0\,, \quad x \in \Omega\,, \quad t \in [0, 1-e^{-T}]\,, \\ v(x,0;b) = f(x,b)\,, \quad v_t(x,0;b) = 0\,, \quad b \in I, \quad x \in \Omega\,, \end{cases} \tag{9}$$

and the function $v_\varphi(x,t) \in C^{m,2}_{x,t}(\Omega \times [0, 1-e^{-T}])$ be a solution of the problem

$$\begin{cases} v_{tt} - A(x,\partial_x)v = 0, \quad x \in \Omega\,, \quad t \in [0, 1-e^{-T}]\,, \\ v(x,0) = \varphi(x), \quad v_t(x,0) = 0\,, \quad x \in \Omega\,. \end{cases} \tag{10}$$

Then the function $u = u(x,t)$ defined by

$$\begin{aligned} u(x,t) = 2\int_0^t db \int_0^{\phi(t)-\phi(b)} v_f(x,r;b)E(r,t;0,b;M)\,dr + e^{\frac{t}{2}}v_{\varphi_0}(x,\phi(t)) \\ + 2\int_0^{\phi(t)} v_{\varphi_0}(x,s)K_0(s,t;M)ds + 2\int_0^{\phi(t)} v_{\varphi_1}(x,s)K_1(s,t;M)ds, \end{aligned}$$

where $x \in \Omega$, $t \in I$, and $\phi(t) := 1 - e^{-t}$, solves the problem

$$\begin{cases} u_{tt} - e^{-2t}A(x,\partial_x)u - M^2u = f, \quad x \in \Omega\,, \quad t \in I, \\ u(x,0) = \varphi_0(x)\,, \quad u_t(x,0) = \varphi_1(x), \quad x \in \Omega\,. \end{cases} \tag{11}$$

Here the kernels E, K_0 and K_1 have been defined in (6), (7) and (8), respectively.

We note that the operator $A(x,\partial_x)$ is of arbitrary order; that is, the equation of (11) can be an evolution equation, not necessarily hyperbolic. Then, the problems in (9) and (11) can be a mixed initial-boundary value problem. The interval $[0, 1-e^{-T}] \subseteq [0,1]$ reflects the fact that de Sitter model possesses the horizon [11].

Among possible applications of the integral transform method are the $L^p - L^q$ estimates, Strichartz estimates, Huygens' principle, global and local existence theorem for semilinear and quasilinear equations.

Example 1. The metric g in the de Sitter-type spacetime, that is, $g_{00} = g^{00} = -1$, $g_{0j} = g^{0j} = 0$, $g_{ij}(x,t) = e^{2t}\sigma_{ij}(x)$, $|g(x,t)| = e^{2nt}|\det\sigma(x)|$, $g^{ij}(x,t) = e^{-2t}\sigma^{ij}(x)$, $i,j = 1,2,\ldots,n$, where $\sum_{j=1}^n \sigma^{ij}(x)\sigma_{jk}(x) = \delta_{ik}$, and δ_{ij} is Kronecker's delta. The linear covariant Klein–Gordon equation in the coordinates is

$$\psi_{tt} - \frac{e^{-2t}}{\sqrt{|\det \sigma(x)|}} \sum_{i,j=1}^{n} \frac{\partial}{\partial x^i} \left(\sqrt{|\det \sigma(x)|} \sigma^{ij}(x) \frac{\partial}{\partial x^j} \psi \right) + n\psi_t + m^2 \psi = f \,.$$

Here m is a physical mass of the particle. If we introduce the new unknown function $u = e^{nt/2}\psi$, then the equation takes the form of the Klein–Gordon equation (3) where $M^2 = \frac{n^2}{4} - m^2$ and

$$A(x, \partial_x)u = \frac{1}{\sqrt{|\det \sigma(x)|}} \sum_{i,j=1}^{n} \frac{\partial}{\partial x^i} \left(\sqrt{|\det \sigma(x)|} \sigma^{ij}(x) \frac{\partial}{\partial x^j} u \right) .$$

If Ω is a non-Euclidean space of constant negative curvature and the equation of the problems (9) and (10) is a non-Euclidean wave equation, then the explicit representation formulas are known (see, e.g., [12, 15]) and the Huygens' principle is a consequence of those formulas. Thus, for a non-Euclidean wave equation, due to Theorem 1, the functions $v_f(x, t; b)$ and $v_\varphi(x, t)$ have explicit representations, and the arguments of [21, 24] allow us to derive for the solution $u(x, t)$ of the problem (11) in the de Sitter-type metric with hyperbolic spatial geometry the explicit representation, the $L^p - L^q$ estimates, and to examine the Huygens' principle.

Example 2. We introduce a toy model that helps to understand the properties of *the black hole formally embedded in the de Sitter universe*. The metric tensor $g_{\mu\nu}$ is generated by line element

$$ds^2 = -\Big(1 - \frac{2GM_{bh}}{c^2 r}\Big)c^2 dt^2 + e^{\frac{2ct}{R}} \Big(1 - \frac{2GM_{bh}}{c^2 r}\Big)^{-1} dr^2 + e^{\frac{2ct}{R}} r^2 (d\theta^2 + \sin^2\theta \, d\phi^2).$$

The metric g is *an asymptotically Einstein metric*, in the sense that for the Ricci tensor $\mathscr{R}_{\mu\nu} = (k + O(r^{-1}))g_{\mu\nu} + O(r^{-2})$. At the same time, the metric g is *an asymptotically hyperbolic (de Sitter) metric*, in the sense that $\mathscr{R}_{\mu\nu} = k(r)g_{\mu\nu} + O(r^{-2})$, as $r \to \infty$. The stress energy tensor T is of Type II (see, [11, p.89]). It is easy to see that the *weak energy condition*, that is $T_{\mu\nu}u^\mu u^\nu \geq 0$ for all time-like vectors u, is not satisfied unless $u^0 u^1 \leq 0$. But it can be proved that it is satisfied on some conic set consisting of the time-like vectors with $u^0 u^1 \geq 0$. The weak energy condition (see, e.g., [11, p.89]) also can be addressed. The covariant wave equation in the black hole embedded in de Sitter universe background is

$$\begin{aligned} &-\left(1 - \frac{2GM_{bh}}{c^2 r}\right)^{-1} \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} - \frac{3}{cR} \left(1 - \frac{2GM_{bh}}{c^2 r}\right)^{-1} \frac{\partial \psi}{\partial t} \\ &+e^{-\frac{2ct}{R}} \left\{ \left(1 - \frac{2GM_{bh}}{c^2 r}\right) \frac{\partial^2 \psi}{\partial r^2} + \frac{2}{r} \left(1 - \frac{GM_{bh}}{c^2 r}\right) \frac{\partial \psi}{\partial r} \right. \\ &\left. +\frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial \psi}{\partial \theta} \right) + \frac{1}{r^2 \sin^2\theta} \frac{\partial}{\partial \phi} \left(\frac{\partial \psi}{\partial \phi} \right) \right\} = 0 \,. \end{aligned}$$

For the large r (the far field) the equation is the wave equation in FLRW spacetime, while the near field limit for small time is Schwarzschild. We make change $u = e^{\frac{3c}{2R}t}\psi$ ($\psi = e^{-\frac{3c}{2R}t}u$) in the wave equation, then it became non-covariant Klein–Gordon equation $u_{tt} - e^{-\frac{2ct}{R}}A(x,\partial_x)u - M^2u = 0$, were $M = \frac{3c}{2R}$ and

$$A(x,\partial_x)u := c^2\left\{\left(1-\frac{2GM_{bh}}{c^2r}\right)^2\frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\left(1-\frac{GM_{bh}}{c^2r}\right)\left(1-\frac{2GM_{bh}}{c^2r}\right)\frac{\partial u}{\partial r} + \left(1-\frac{2GM_{bh}}{c^2r}\right)\frac{1}{r^2}\Delta_{\mathbb{S}^2}u\right\}.$$

Theorem 1 allows us to reveal the properties of the waves propagating in the spacetime of black hole embedded in the de Sitter background.

3 The Semilinear Equations in the De Sitter Spacetime

In this section, we present some results obtained in [10] on the existence of a global in time solutions of the semilinear Klein–Gordon equation in the de Sitter spacetime with the time slices being Riemannian manifolds. In the spatially flat de Sitter model, this can be $\mathbb{R}^3$ and in the spatially closed and spatially open cases it can be the three-sphere $\mathbb{S}^3$ and the three-hyperboloid $\mathbb{H}^3$, respectively. The metric g in the de Sitter spacetime is defined as follows: $g_{00} = g^{00} = -1$, $g_{0j} = g^{0j} = 0$, $g_{ij}(x,t) = e^{2t}\sigma_{ij}(x)$, $i,j = 1,2,\dots,n$, where $\sum_{j=1}^n \sigma^{ij}(x)\sigma_{jk}(x) = \delta_{ik}$, and δ_{ij} is Kronecker's delta.

In quantum field theory, the matter fields are described by a function ψ that must satisfy equations of motion. In the case of a massive scalar field, the equation of motion is the semilinear Klein–Gordon equation generated by the metric g. In physical terms, this equation describes a local self-interaction for a scalar particle. The covariant Klein–Gordon equation in the de Sitter spacetime in the coordinates is

$$\psi_{tt} - \frac{e^{-2t}}{\sqrt{|\det\sigma(x)|}}\sum_{i,j=1}^n \frac{\partial}{\partial x^i}\left(\sqrt{|\det\sigma(x)|}\sigma^{ij}(x)\frac{\partial}{\partial x^j}\psi\right) + n\psi_t + m^2\psi = F(\psi).$$

Here m is a physical mass of the particle. This is a special case of the equation $\psi_{tt} + n\psi_t - e^{-2t}A(x,\partial_x)\psi + m^2\psi = F(\psi)$, where $A(x,\partial_x) = \sum_{|\alpha|\le 2} a_\alpha(x)\partial_x^\alpha$ is a second order partial differential operator. We assume that $a_\alpha(x)$, $|\alpha| = 2$, is positive definite. To formulate the theorem, we need the following description of the nonlinear term.

Condition ($\mathscr{L}$). *The function F is said to be Lipschitz continuous with exponent $\alpha \ge 0$ in the Sobolev space $H_{(s)}(\mathbb{R}^n)$ if there is a constant $C \ge 0$ such that*

$$\|F(x,\psi_1)-F(x,\psi_2)\|_{H_{(s)}(\mathbb{R}^n)} \le C\|\psi_1-\psi_2\|_{H_{(s)}(\mathbb{R}^n)}\left(\|\psi_1\|^{\alpha}_{H_{(s)}(\mathbb{R}^n)}+\|\psi_2\|^{\alpha}_{H_{(s)}(\mathbb{R}^n)}\right)$$

for all $\psi_1, \psi_2 \in H_{(s)}(\mathbb{R}^n)$.

Next, we define the complete metric space

$$X(R,s,\gamma) := \{\psi \in C([0,\infty); H_{(s)}(\mathbb{R}^n)) \mid \|\psi\|_X := \sup_{t\in[0,\infty)} e^{\gamma t}\|\psi(x,t)\|_{H_{(s)}(\mathbb{R}^n)} \le R\}$$

with the metric

$$d(\psi_1,\psi_2) := \sup_{t\in[0,\infty)} e^{\gamma t}\|\psi_1(x,t)-\psi_2(x,t)\|_{H_{(s)}(\mathbb{R}^n)}.$$

Let $\mathscr{B}^{\infty}$ be the space of all $C^{\infty}(\mathbb{R}^n)$ functions with uniformly bounded derivatives of all orders.

Theorem 2. [10] *Let* $A(x,\partial_x)=\sum_{|\alpha|\le 2} a_\alpha(x)\partial_x^\alpha$ *be a second-order negative elliptic differential operator with real coefficients* $a_\alpha \in \mathscr{B}^{\infty}$. *Assume that the nonlinear term* $F(u)$ *is a Lipschitz continuous with exponent* $\alpha > 0$ *in the space* $H_{(s)}(\mathbb{R}^n)$, $s > n/2 \ge 1$, *and* $F(0)=0$. *Assume also that* $m \in (0,\sqrt{n^2-1}/2] \cup [n/2,\infty)$. *Then, there exists* $\varepsilon_0 > 0$ *such that, for every given functions* $\psi_0, \psi_1 \in H_{(s)}(\mathbb{R}^n)$, *such that*

$$\|\psi_0\|_{H_{(s)}(\mathbb{R}^n)} + \|\psi_1\|_{H_{(s)}(\mathbb{R}^n)} \le \varepsilon, \qquad \varepsilon < \varepsilon_0\,,$$

there exists a global solution $\psi \in C^1([0,\infty); H_{(s)}(\mathbb{R}^n))$ *of the Cauchy problem*

$$\psi_{tt} + n\psi_t - e^{-2t}A(x,\partial_x)\psi + m^2\psi = F(\psi)\,, \tag{12}$$

$$\psi(x,0)=\psi_0(x)\,, \quad \psi_t(x,0)=\psi_1(x)\,. \tag{13}$$

That solution $\psi(x,t)$ *belongs to the space* $X(2\varepsilon,s,\gamma)$; *that is,*

$$\sup_{t\in[0,\infty)} e^{\gamma t}\|\psi(\cdot,t)\|_{H_{(s)}(\mathbb{R}^n)} < 2\varepsilon\,,$$

with γ *such that either* $0 < \gamma \le \frac{1}{\alpha+1}\left(\frac{n}{2}-\sqrt{\frac{n^2}{4}-m^2}\right)$ *if* $\sqrt{n^2-1}/2 \ge m > 0$, *or we choose* $0 \le \gamma_0 < \frac{n-1}{2}$ *if* $m = n/2$ *and* $0 \le \gamma_0 \le \frac{n-1}{2}$ *if* $m > n/2$, *then* $\gamma \le \min\left\{\gamma_0\,, \frac{n}{2(\alpha+1)}\right\}$.

If $m \in (\sqrt{n^2-1}/2, n/2)$, *then for the problem with* $\psi_0 = 0$ *the global solution exists and belongs to* $X(2\varepsilon,s,\gamma)$, *where* $\gamma \in (0, \frac{1}{\alpha+1}(\frac{n}{2}-\sqrt{\frac{n^2}{4}-m^2}))$.

For $n=3$, the mass m interval $(0,\sqrt{2})$ is called the Higuchi bound in quantum field theory [13]. The proof of the global existence [10] is based on the integral transform and L^p-L^q estimates. The range $m \in (\sqrt{n^2-1}/2, n/2)$, which seems to be a forbidden mass interval for the problem with general initial data, can be

allowed if we change the setting of the problem (see, also, [14]). Indeed, we have the following result for all $m > 0$.

Theorem 3. [10] *Let* $A(x,\partial_x)=\sum_{|\alpha|\le 2} a_\alpha(x)\partial_x^\alpha$ *be a second-order negative elliptic differential operator with real coefficients* $a_\alpha \in \mathscr{B}^\infty$. *Assume that the nonlinear term* $F(u)$ *is a Lipschitz continuous with exponent* $\alpha > 0$ *in the space* $H_{(s)}(\mathbb{R}^n)$, $s > n/2 \ge 1$, *and* $F(0) = 0$. *Assume also that* $m > 0$. *Then, there exists* $\varepsilon_0 > 0$ *such that for every given function* $f \in X(\varepsilon, s, \gamma_{rhs})$, *such that*

$$\sup_{t\in[0,\infty)} e^{\gamma_{rhs}t}\|f(x,t)\|_{H_{(s)}(\mathbb{R}^n)} \le \varepsilon < \varepsilon_0\,,$$

there exists a global solution $\psi \in C^1([0,\infty); H_{(s)}(\mathbb{R}^n))$ *of the Cauchy problem*

$$\begin{aligned}&\psi_{tt} + n\psi_t - e^{-2t}A(x,\partial_x)\psi + m^2\psi - F(\psi) = f\,,\\ &\psi(x,0) = 0\,,\quad \psi_t(x,0) = 0\,.\end{aligned}$$

That solution $\psi(x,t)$ *belongs to the space* $X(2\varepsilon, s, \gamma)$, *with* γ *such that*

$$\begin{cases}
\gamma < \dfrac{1}{\alpha+1}\gamma_{rhs} \quad if \quad m < \dfrac{n}{2} \quad and \quad \gamma_{rhs} \le \dfrac{n}{2} - \sqrt{\dfrac{n^2}{4} - m^2}\,,\\
\gamma < \dfrac{1}{\alpha+1}\left(\dfrac{n}{2} - \sqrt{\dfrac{n^2}{4} - m^2}\right) \quad if \quad m < \dfrac{n}{2} \quad and \quad \gamma_{rhs} > \dfrac{n}{2} - \sqrt{\dfrac{n^2}{4} - m^2}\,,\\
\gamma \le \min\left\{\gamma_{rhs}, \dfrac{n}{2(\alpha+1)}\right\} \quad if \quad m \ge \dfrac{n}{2} \quad and \quad \dfrac{n}{2} > \gamma_{rhs}\,,\\
\gamma \le \min\left\{\gamma_0, \dfrac{n}{2(\alpha+1)}\right\} \quad where \quad \gamma_0 < \gamma_{rhs} \quad if \quad m = \dfrac{n}{2} \quad and \quad \dfrac{n}{2} = \gamma_{rhs}\,,\\
\gamma \le \dfrac{n}{2(\alpha+1)} \quad if \quad m > \dfrac{n}{2} \quad and \quad \dfrac{n}{2} \le \gamma_{rhs}\,,\\
\gamma < \dfrac{n}{2(\alpha+1)} \quad if \quad m = \dfrac{n}{2} \quad and \quad \dfrac{n}{2} < \gamma_{rhs}\,.
\end{cases}$$

The mass $m = \sqrt{n^2-1}/2$ represents the only field that obeys the Huygens principle [24].

The Klein–Gordon quantum fields on the de Sitter manifold with imaginary mass, which take an infinite set of discrete values as follows

$$m^2 = -k(k+n)\,,\quad k = 0, 1, 2, \ldots, \tag{14}$$

present a family of scalar tachyonic quantum fields. Epstein and Moschella [5] give a complete study of a family of scalar tachyonic quantum fields which are linear Klein–Gordon quantum fields on the de Sitter manifold whose squared masses are negative and take an infinite set of discrete values (14). The corresponding linear equation is

$$\psi_{tt} + n\psi_t - e^{-2t}\Delta\psi + m^2\psi = 0\,,$$

for which the kernel is $E(x,t;x_0,t_0;M)$, where $M = \sqrt{\frac{n^2}{4} + k(k+n)} = k + \frac{n}{2}$, $k = 0, 1, 2, \ldots$. If n is an odd number, then m takes value at knot points set. The nonexistence of a global in time solution of the semilinear Klein–Gordon tachyonic quantum field equation in the de Sitter spacetime is proved in [22]. The conclusion is that the self-interacting tachyonic quantum fields in the de Sitter spacetime have finite lifespan. More precisely, consider the semilinear equation

$$\psi_{tt} + n\psi_t - e^{-2t}\Delta\psi - m^2\psi = c|\psi|^{1+\alpha}\,,$$

which is commonly used model for general nonlinear problems. Then, according to Theorem 1.1 [22], if $c \neq 0, \alpha > 0$, and $m \neq 0$, then for every positive numbers ε and s there exist functions $\psi_0,\ \psi_1 \in C_0^\infty(\mathbb{R}^n)$ such that $\|\psi_0\|_{H_{(s)}(\mathbb{R}^n)} + \|\psi_1\|_{H_{(s)}(\mathbb{R}^n)} \leq \varepsilon$ but the solution $\psi = \psi(x,t)$ to (12) with the initial values (13) blows up in finite time.

Acknowledgements This paper was completed during my visit at the Technical University Bergakademie Freiberg in the summer of 2016. I am grateful to Michael Reissig for the invitation to Freiberg and for the warm hospitality. I express my gratitude to the Deutsche Forschungsgemeinschaft for the financial support under the grant GZ: RE 961/21-1.

References

1. A. Bachelot, Waves in the Witten bubble of nothing and the Hawking wormhole. Commun. Math. Phys. **351**(2), 599–651 (2017). https://doi.org/10.1007/s00220-016-2792-7
2. D. Baskin, Strichartz estimates on asymptotically de sitter spaces. Ann. Henri Poincaré **14**(2), 221–252 (2013)
3. J. Bros, H. Epstein, U. Moschella, Particle decays and stability on the de Sitter universe. Ann. Henri Poincaré **11**(4), 611–658 (2010)
4. J.L. Costa, A. Alho, J. Natário, Spherical linear waves in de Sitter spacetime. J. Math. Phys. **53**(5), 052501, 9 (2012)
5. H. Epstein, U. Moschella, De Sitter tachyons and related topics. Commun. Math. Phys. **336**(1), 381–430 (2015)
6. A. Galstian, T. Kinoshita, K. Yagdjian, A note on wave equation in Einstein and de Sitter space-time. J. Math. Phys. **51**(5), 052501 (2010)
7. A. Galstian, K. Yagdjian, Microlocal analysis for waves propagating in Einstein & de Sitter spacetime. Math. Phys. Anal. Geom. **17**(1–2), 223–246 (2014)
8. A. Galstian, K. Yagdjian, Global solutions for semilinear Klein-Gordon equations in FLRW spacetimes. Nonlinear Anal. **113**, 339–356 (2015)
9. A. Galstian, T. Kinoshita, Representation of solutions for 2nd order one-dimensional model Hyperbolic equations. J. d'Analyse Mathematique. **130**, 355–374 (2016)
10. A. Galstian, K. Yagdjian, Global in time existence of the self-interacting scalar field in de Sitter spacetimes. Nonlinear Anal. Real World Appl. **34**, 110–139 (2017)
11. S.W. Hawking, G.F.R. Ellis, *The Large Scale Structure of Space-Time*. Cambridge Monographs on Mathematical Physics, vol. 1 (Cambridge University Press, New York, 1973)

12. S. Helgason, Wave equations on homogeneous spaces, *Lie Group Representations*, vol. III (University of Maryland, College Park, 1982/1983); Lecture Notes in Mathematics, vol. 1077 (Springer, Berlin, 1984), pp. 254–287
13. A. Higuchi, Forbidden mass range for spin-2 field theory in de Sitter spacetime. Nucl. Phys. B **282**(2), 397–436 (1987)
14. P. Hintz, A. Vasy, Semilinear wave equations on asymptotically de Sitter, Kerr-de Sitter and Minkowski spacetimes. Anal. PDE **8**(8), 1807–1890 (2015)
15. P.D. Lax, R.S. Phillips, Translation representations for the solution of the non-Euclidean wave equation. Commun. Pure Appl. Math. **32**(5), 617–667 (1979)
16. J. Näf, P. Jetzer, M. Sereno, On gravitational waves in spacetimes with a nonvanishing cosmological constant. Phys. Rev. D **79**, 024014 (2009)
17. M. Nakamura, The Cauchy problem for semi-linear Klein-Gordon equations in de Sitter spacetime. J. Math. Anal. Appl. **410**(1), 445–454 (2014)
18. W.A. Strauss, *Partial Differential Equations: An Introduction*, 2nd edn. (Wiley, Chichester, 2008)
19. A. Vasy, Microlocal analysis of asymptotically hyperbolic and Kerr-de Sitter spaces (with an appendix by Semyon Dyatlov). Invent. Math. **194**(2), 381–513 (2013)
20. K. Yagdjian, A note on the fundamental solution for the Tricomi-type equation in the hyperbolic domain. J. Differ. Equ. **206**, 227–252 (2004)
21. K. Yagdjian, A. Galstian, Fundamental solutions for the klein-Gordon equation in de Sitter spacetime. Commun. Math. Phys. **285**, 293–344 (2009)
22. K. Yagdjian, The semilinear Klein-Gordon equation in de Sitter spacetime. Discret. Contin. Dyn. Syst. Ser. S **2**(3), 679–696 (2009)
23. K. Yagdjian, On the global solutions of the Higgs boson equation. Commun. Partial Differ. Equ. **37**(3), 447–478 (2012)
24. K. Yagdjian, Huygens' principle for the Klein-Gordon equation in the de Sitter spacetime. J. Math. Phys. **54**(9), 091503 (2013)
25. K. Yagdjian, Semilinear hyperbolic equations in curved spacetime, *Fourier Analysis, Pseudo-differential Operators, Time-Frequency Analysis and Partial Differential Equations*. Trends in Mathematics (Birkhäuser Mathematics, Basel, 2014), pp. 391–415
26. K. Yagdjian, Integral transform approach to solving Klein-Gordon equation with variable coefficients. Math. Nachr. **288**(17–18), 2129–2152 (2015)
27. K. Yagdjian, Integral transform approach to generalized Tricomi equations. J. Differ. Equ. **259**, 5927–5981 (2015)

Asymptotic Consistency of the RS-IMEX Scheme for the Low-Froude Shallow Water Equations: Analysis and Numerics

Hamed Zakerzadeh

Abstract In the present work, we *formally* prove the asymptotic consistency of the recently presented Reference Solution IMplicit–EXplicit (RS-IMEX) scheme for the two-dimensional shallow water equations. The scheme has been analyzed extensively for the low-Froude one-dimensional shallow water equations in (Zakerzadeh IGPM report 455 (2016) [18]), and the present paper is going to discuss the asymptotic consistency analysis for the two-dimensional case, with the aid of some numerical experiments.

Keywords IMEX scheme · Asymptotic preserving · Shallow water equations

1 RS-IMEX Schemes: An Introduction

In the singular limits of conservation laws, characterized by the singular parameter $\varepsilon \in (0, 1]$ approaching zero, the type of the equations changes, e.g., when the Mach number, denoted by ε, approaches zero for the Euler equations (the incompressible limit), the sound speed goes to the infinity and the system changes to be hyperbolic-elliptic. Such a singularity not only hinders the analysis (see [16]), but also gives rise to lots of issues for numerical schemes, e.g., schemes may lose their accuracy for under-resolved mesh sizes (see [6]) for weakly compressible flows or the time step gets very restrictive for explicit schemes, in virtue of the Courant–Friedrichs–Lewy (CFL) condition, i.e., $\Delta t \lesssim \varepsilon \, \Delta x$, which leads to a huge computational cost.

Assuming that the "*solution*" of the singularly perturbed problem converges to the "*solution*" of the limit problem, we aim to discuss the counterpart of such a convergence in the discrete level. This is the idea of Asymptotic Preserving (AP) schemes [13] for an ε-dependent system converging to a limit for $\varepsilon \to 0$. The numerical scheme is AP if it provides a stable, consistent, and efficient scheme for the contin-

H. Zakerzadeh (✉)
Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany
e-mail: h.zakerzadeh@igpm.rwth-aachen.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_50

uous limit system. For the sake of simplicity, we only consider well-prepared initial data to eliminate spurious initial layers.

The AP property has been studied extensively for conservation laws (as well as kinetic equations, cf. [14]), and several AP schemes have been developed and analyzed; see [2, 5, 12, 17] among others. Although most of these works present a *formal* analysis, there are few results regarding the rigorous asymptotic consistency or stability, e.g., [1, 7, 8, 10, 19] for hyperbolic balance laws.

The bottom line of these AP schemes is a mixed implicit–explicit (IMEX) approach to split the flux (or its Jacobian) into stiff and non-stiff parts and treat them explicitly and implicitly in time. Such an approach is necessary for an ε-uniform CFL condition, but is not sufficient for asymptotic stability; see [17] for instance, where a *CFL-stable* IMEX scheme requires an ε-dependent time step for stability. This, in fact, gave the motivation for the RS-IMEX scheme, as we will review here. The *penalization method* [9] for the kinetic equations, as well as [2] for the shallow water equations are close to the RS-IMEX scheme, in essence.

The goal of this section is to provide a very brief introduction to the RS-IMEX scheme; see also [15, 18]. Then in the next section, we prove the asymptotic consistency of the scheme followed by some numerical experiments in Sect. 4. The reader is referred to [18] for a rigorous asymptotic analysis for the one-dimensional shallow water system, which is the backbone of the analysis in the present work.

Consider the general hyperbolic system of balance laws in $\Omega \subset \mathbb{R}^d$

$$\partial_t \boldsymbol{U}(\boldsymbol{x}, t; \varepsilon) + \operatorname{div}_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{U}, \boldsymbol{x}, t; \varepsilon) = \boldsymbol{S}(\boldsymbol{U}, \boldsymbol{x}, t; \varepsilon), \tag{1}$$

where $\Omega := \mathbb{T}^d$ is a d-dimensional torus, $\boldsymbol{U} \in \mathbb{R}^q$ is the vector of unknowns, $\boldsymbol{F} \in \mathbb{R}^{q\times d}$ is the flux matrix (in d space dimensions), $\varepsilon \in (0, 1]$ is the singular parameter, and $\boldsymbol{S} \in \mathbb{R}^q$ is the source term. Note that we often suppress the dependence of $\boldsymbol{U}$, $\boldsymbol{F}$ and $\boldsymbol{S}$ on ε. To have a hyperbolic system, we also assume that $\boldsymbol{F}$ has a real diagonalizable Jacobian $\boldsymbol{F}' := \partial_{\boldsymbol{U}} \boldsymbol{F}$.

The main idea of the RS-IMEX scheme is to split the solution $\boldsymbol{U}$ of the balance laws (1) into the (given) reference solution $\overline{\boldsymbol{U}}$ and a perturbation $\boldsymbol{U}_{pert}$, i.e., $\boldsymbol{U} = \overline{\boldsymbol{U}} + \boldsymbol{U}_{pert}$. The reference solution can be a steady-state solution of (1), or the solution of the asymptotic limit of (1) as $\varepsilon \to 0$. Then, as in [18], we use a Taylor expansion around $\overline{\boldsymbol{U}}$ to split the flux and source terms into reference ($\overline{\boldsymbol{F}}$, $\overline{\boldsymbol{S}}$), linear stiff ($\widetilde{\boldsymbol{F}}$, $\widetilde{\boldsymbol{S}}$) and nonlinear non-stiff parts ($\widehat{\boldsymbol{F}}$, $\widehat{\boldsymbol{S}}$):

$$\begin{aligned}
\boldsymbol{F}(\boldsymbol{U}) &= \boldsymbol{F}(\overline{\boldsymbol{U}}) + \boldsymbol{F}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert} + \big(\boldsymbol{F}(\boldsymbol{U}) - \boldsymbol{F}(\overline{\boldsymbol{U}}) - \boldsymbol{F}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert}\big) &&=: \overline{\boldsymbol{F}} + \widetilde{\boldsymbol{F}} + \widehat{\boldsymbol{F}},\\
\boldsymbol{S}(\boldsymbol{U}) &= \boldsymbol{S}(\overline{\boldsymbol{U}}) + \boldsymbol{S}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert} + \big(\boldsymbol{S}(\boldsymbol{U}) - \boldsymbol{S}(\overline{\boldsymbol{U}}) - \boldsymbol{S}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert}\big) &&=: \overline{\boldsymbol{S}} + \widetilde{\boldsymbol{S}} + \widehat{\boldsymbol{S}}.
\end{aligned}$$

We, then, scale the components of the perturbation (see [18] for a discussion) by the scaling matrix $D := \operatorname{diag}(\varepsilon^{d_1}, \ldots, \varepsilon^{d_q})$ and define the scaled perturbation as $\boldsymbol{V} := D^{-1}\boldsymbol{U}_{pert}$ to obtain the corresponding scaled splitting:

$$\boldsymbol{G} = \overline{\boldsymbol{G}} + \widetilde{\boldsymbol{G}} + \widehat{\boldsymbol{G}}, \qquad \boldsymbol{Z} = \overline{\boldsymbol{Z}} + \widetilde{\boldsymbol{Z}} + \widehat{\boldsymbol{Z}},$$

with similar definitions as for the splittings of $\boldsymbol{F}$ and $\boldsymbol{S}$. Defining $\boldsymbol{R} := -\mathrm{div}_{\boldsymbol{x}}\boldsymbol{G} + \boldsymbol{Z}$ (with analogous definitions for $\overline{\boldsymbol{R}}, \widetilde{\boldsymbol{R}}$ and $\widehat{\boldsymbol{R}}$), and also $\overline{\boldsymbol{T}}$ as the (a priori-known) scaled residual of the reference solution

$$\overline{\boldsymbol{T}} := D^{-1}\partial_t \overline{\boldsymbol{U}} - \overline{\boldsymbol{R}}, \tag{2}$$

one can reformulate the balance laws (1) as

$$\partial_t \boldsymbol{V} = -\overline{\boldsymbol{T}} + \widetilde{\boldsymbol{R}} + \widehat{\boldsymbol{R}}, \tag{3}$$

which is a system for the scaled perturbation $\boldsymbol{V} := (v_1, \ldots, v_q)^T$.

Solving this reformulated problem (3), numerically, defines the RS-IMEX scheme. We solve stiff $\widetilde{\boldsymbol{R}}$ implicitly in time to avoid restrictive time steps in the limit (by using the implicit Euler method) while the (expected to be) non-stiff part $\widehat{\boldsymbol{R}}$ is treated by the explicit Euler method. Moreover, $\overline{\boldsymbol{T}}$ is computed independently, e.g., by an incompressible solver if $\overline{\boldsymbol{U}}$ is the solution of the incompressible Euler equations. We use a Rusanov-type numerical flux, with numerical diffusion coefficients $\widetilde{\alpha}$ and $\widehat{\alpha}$ and an appropriate spatial discretization for the source term (to avoid well-balancing issues). Note that $\widetilde{\alpha}$ and $\widehat{\alpha}$ originally should be chosen as the maximum over the domain and all characteristic fields (of stiff or non-stiff parts). But here, not to add an excessive diffusion to the implicit step, we pick $\widetilde{\alpha} = 0$.

Definition 1. Given the reference solution $\overline{\boldsymbol{U}}$, the RS-IMEX scheme for (3) is given by

$$D_t \boldsymbol{V}_{\Delta}^{n} = -\overline{\boldsymbol{T}}_{\Delta}^{n+1} + \widetilde{\boldsymbol{R}}_{\Delta}^{n+1} + \widehat{\boldsymbol{R}}_{\Delta}^{n}, \tag{4}$$

with the Euler time integration D_t when Δ stands for spatial discretization.

The advantages of the scheme are twofold. Firstly, the implicit part of the scheme is linear by construction, which is very advantageous in terms of computational cost.[1] Secondly, as we will see in Remark 1, it makes the asymptotic consistency analysis easier as the scheme deals with the perturbations $\boldsymbol{V}$ directly.

To summarize, in the RS-IMEX algorithm two coupled systems should be solved separately: With a given reference state at step n, one finds the scaled perturbation $\boldsymbol{V}_{\Delta}^{n+1}$, while the reference state may evolve over time and should be computed independently. This procedure is repeated in each step.

[1] The idea of such a linearization goes back to the so-called linearly implicit methods for ODEs and has been used later in [2, 11].

2 RS-IMEX Scheme for the Shallow Water Equations

In this section, we apply the RS-IMEX scheme to the two-dimensional shallow water equations with bottom topography. Rather than the classical form of this system, we consider its reformulation as [2] in the periodic domain $\Omega = \mathbb{T}^2$:

$$\begin{cases} \partial_t z + \operatorname{div}_{\boldsymbol{x}} \boldsymbol{m} = 0, \\ \partial_t \boldsymbol{m} + \operatorname{div}_{\boldsymbol{x}} \left(\dfrac{\boldsymbol{m} \otimes \boldsymbol{m}}{z-b} + \dfrac{z^2 - 2bz}{2\varepsilon^2} \mathbb{I}_2 \right) = -\dfrac{z}{\varepsilon^2} \nabla_{\boldsymbol{x}} b, \end{cases} \tag{5}$$

where z is the surface elevation from the mean surface level H_{mean}, $\boldsymbol{m} := (z-b)\boldsymbol{u}$ is the momentum with the velocity $\boldsymbol{u} = (u_1, u_2)$, b is the water depth measured from H_{mean} with a negative sign, and the singular parameter $\varepsilon \in (0, 1]$ is called the Froude number, cf. [18]. Using (5), one can identify $\boldsymbol{U}$, $\boldsymbol{F}$, and $\boldsymbol{S}$ as

$$\boldsymbol{U} = \begin{bmatrix} z \\ m_1 \\ m_2 \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} m_1 & m_2 \\ \dfrac{m_1^2}{z-b} + \dfrac{z^2 - 2zb}{2\varepsilon^2} & \dfrac{m_1 m_2}{z-b} \\ \dfrac{m_1 m_2}{z-b} & \dfrac{m_2^2}{z-b} + \dfrac{z^2 - 2zb}{2\varepsilon^2} \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} 0 \\ -z\, b_x/\varepsilon^2 \\ -z\, b_y/\varepsilon^2 \end{bmatrix}. \tag{6}$$

Given the scaling matrix $D = \operatorname{diag}(\varepsilon^2, 1, 1)$, $\overline{\boldsymbol{U}} = (\overline{z}, \overline{m_1}, \overline{m_2})^T$, and the scaled perturbation $\boldsymbol{V} := D^{-1}(\boldsymbol{U} - \overline{\boldsymbol{U}})$, the RS-IMEX splitting for (5) gives the reference and stiff parts as

$$\overline{\boldsymbol{G}} = \begin{bmatrix} \overline{m_1}/\varepsilon^2 & \overline{m_2}/\varepsilon^2 \\ \dfrac{\overline{m_1}^2}{\overline{z}-b} + \dfrac{\overline{z}^2 - 2\overline{z}b}{2\varepsilon^2} & \dfrac{\overline{m_1 m_2}}{\overline{z}-b} \\ \dfrac{\overline{m_1 m_2}}{\overline{z}-b} & \dfrac{\overline{m_2}^2}{\overline{z}-b} + \dfrac{\overline{z}^2 - 2\overline{z}b}{2\varepsilon^2} \end{bmatrix}, \tag{7a}$$

$$\widetilde{\boldsymbol{G}} = \begin{bmatrix} v_2/\varepsilon^2 & v_3/\varepsilon^2 \\ -\dfrac{\overline{m_1}^2 v_1 \varepsilon^2}{(\overline{z}-b)^2} + \dfrac{2\overline{m_1} v_2}{\overline{z}-b} + (\overline{z}-b)v_1 & -\dfrac{\overline{m_1 m_2} v_1 \varepsilon^2}{(\overline{z}-b)^2} + \dfrac{\overline{m_1} v_3}{\overline{z}-b} + \dfrac{\overline{m_2} v_2}{\overline{z}-b} \\ -\dfrac{\overline{m_1 m_2} v_1 \varepsilon^2}{(\overline{z}-b)^2} + \dfrac{\overline{m_1} v_3}{\overline{z}-b} + \dfrac{\overline{m_2} v_2}{\overline{z}-b} & -\dfrac{\overline{m_2}^2 v_1 \varepsilon^2}{(\overline{z}-b)^2} + \dfrac{2\overline{m_2} v_3}{\overline{z}-b} + (\overline{z}-b)v_1 \end{bmatrix}, \tag{7b}$$

$$\overline{\boldsymbol{Z}} = \begin{bmatrix} 0 \\ -\overline{z}\, b_x/\varepsilon^2 \\ -\overline{z}\, b_y/\varepsilon^2 \end{bmatrix}, \quad \widetilde{\boldsymbol{Z}} = \begin{bmatrix} 0 \\ -v_1 b_x \\ -v_1 b_y \end{bmatrix}. \tag{7c}$$

while $\widehat{\boldsymbol{Z}} = \boldsymbol{0}$ and $\widehat{\boldsymbol{G}}(\overline{\boldsymbol{U}}, \boldsymbol{V}) = \boldsymbol{G}(\overline{\boldsymbol{U}} + \boldsymbol{V}) - \overline{\boldsymbol{G}}(\overline{\boldsymbol{U}}) - \widetilde{\boldsymbol{G}}(\overline{\boldsymbol{U}}, \boldsymbol{V})$. One can verify that the Jacobian matrices $\widehat{\boldsymbol{G}}'$ and $\widetilde{\boldsymbol{G}}'$ have complete sets of eigenvectors and that the eigenvalues of $\widehat{\boldsymbol{G}}'$ are non-stiff. This can be readily seen from the expression of the non-stiff flux $\widehat{\boldsymbol{G}}_1$ (and similarly $\widehat{\boldsymbol{G}}_2$)

$$\widehat{\boldsymbol{G}}_1 = \begin{bmatrix} 0 \\ \dfrac{m_1^2}{z-b} + \dfrac{z^2-2zb}{2\varepsilon^2} - \dfrac{\overline{m_1}^2}{\overline{z}-b} - \dfrac{\overline{z}^2-2\overline{z}b}{2\varepsilon^2} + \dfrac{\overline{m_1}^2 v_1 \varepsilon^2}{(\overline{z}-b)^2} - \dfrac{2\overline{m_1}v_2}{\overline{z}-b} - (\overline{z}-b)v_1 \\ \dfrac{m_1 m_2}{z-b} - \dfrac{\overline{m_1 m_2}}{\overline{z}-b} + \dfrac{\overline{m_1 m_2} v_1 \varepsilon^2}{(\overline{z}-b)^2} - \dfrac{\overline{m_1}v_3}{\overline{z}-b} - \dfrac{\overline{m_2}v_2}{\overline{z}-b} \end{bmatrix}, \tag{7d}$$

as, after simplification, it does not contain any $\mathcal{O}(1/\varepsilon)$ term.

Denoting the central discretization of the first and second derivatives in the x-direction by $\nabla_{h,x}$ and $\Delta_{h,x}$ respectively, the RS-IMEX scheme can be written as

$$\boldsymbol{V}_{ij}^{n+\frac{1}{2}} = \boldsymbol{V}_{ij}^{n} - \Delta t \left(\nabla_{h,x} \widehat{\boldsymbol{G}}_{1,ij}^{n} + \nabla_{h,y} \widehat{\boldsymbol{G}}_{2,ij}^{n} \right) + \Delta t \frac{\widehat{\alpha} \Delta x}{2} \Delta_{h,\boldsymbol{x}} \boldsymbol{V}_{ij}^{n}, \tag{8a}$$

$$\boldsymbol{V}_{ij}^{n+1} = \boldsymbol{V}_{ij}^{n+\frac{1}{2}} - \Delta t \left(\nabla_{h,x} \widetilde{\boldsymbol{G}}_{1,ij}^{n+1} + \nabla_{h,y} \widetilde{\boldsymbol{G}}_{2,ij}^{n+1} \right) + \Delta t \widetilde{\boldsymbol{Z}}_{ij}^{n+1} - \Delta t \overline{\boldsymbol{T}}_{ij}^{n+1}, \tag{8b}$$

for each cell $(i, j) \in \{1, 2, \ldots, N\}^2$ in the square computational domain Ω_N with spatial steps $\Delta x = \Delta y$ and the time step Δt, where $\widetilde{\boldsymbol{Z}}_{ij}^{n+1}$ is the central discretization of the source term (7c), and $\overline{\boldsymbol{T}}_{ij}^{n+1}$ is the central discretization of the scaled residual (2) computed as

$$\overline{\boldsymbol{T}}_{ij}^{n+1} = D^{-1} \frac{\overline{\boldsymbol{U}}_{ij}^{n+1} - \overline{\boldsymbol{U}}_{ij}^{n}}{\Delta t} + \nabla_{h,x} \overline{\boldsymbol{G}}_{1,ij}^{n+1} + \nabla_{h,y} \overline{\boldsymbol{G}}_{2,ij}^{n+1} - \overline{\boldsymbol{Z}}_{ij}^{n+1}. \tag{9}$$

The reference solution is chosen as the zero-Froude limit, which is the solution of the so-called *lake equations* (cf. [3] for a formal derivation):

$$\begin{cases} \partial_t \boldsymbol{m} - \operatorname{div}_{\boldsymbol{x}} \left(\dfrac{\boldsymbol{m} \otimes \boldsymbol{m}}{b} \right) - b \nabla_{\boldsymbol{x}} \pi = \boldsymbol{0}, \\ \operatorname{div}_{\boldsymbol{x}} \boldsymbol{m} = 0. \end{cases} \tag{10}$$

So, considering the solution of (10) as $\overline{\boldsymbol{U}}$ with a constant (in time and space) $\overline{z}$ and a solenoidal $\overline{\boldsymbol{m}}$, one can write $\overline{\boldsymbol{T}}$ block-wise as $\overline{\boldsymbol{T}}_{\Delta}^{n+1} := [\overline{\boldsymbol{T}}_{1,\Delta}^{n+1}, \overline{\boldsymbol{T}}_{2,\Delta}^{n+1}, \overline{\boldsymbol{T}}_{3,\Delta}^{n+1}]^T$ with

$$
\begin{aligned}
\overline{\boldsymbol{T}}_{1,ij}^{n+1} &= \left(\nabla_{h,x}\overline{m}_{1ij}^{n+1} + \nabla_{h,x}\overline{m}_{2ij}^{n+1}\right)/\varepsilon^2,\\
\overline{\boldsymbol{T}}_{2,ij}^{n+1} &= D_t\overline{m}_{1ij}^{n} + \nabla_{h,x}\left(\frac{\overline{m}_{1ij}^{n+1,2}}{\overline{z}-b_{ij}}\right) + \nabla_{h,y}\left(\frac{\overline{m}_{1ij}^{n+1}\overline{m}_{2ij}^{n+1}}{\overline{z}-b_{ij}}\right),\\
\overline{\boldsymbol{T}}_{3,ij}^{n+1} &= D_t\overline{m}_{2ij}^{n} + \nabla_{h,x}\left(\frac{\overline{m}_{1ij}^{n+1}\overline{m}_{2ij}^{n+1}}{\overline{z}-b_{ij}}\right) + \nabla_{h,y}\left(\frac{\overline{m}_{1ij}^{n+1,2}}{\overline{z}-b_{ij}}\right).
\end{aligned}
\tag{11}
$$

So far, the scheme for computing the scaled perturbation has been introduced. The remaining point to be clarified is how to solve the equations for the reference solution (10), which is needed to compute $\overline{\boldsymbol{T}}$. In fact, there exist several numerical methods for the lake equations. Here, we employ the so-called Chorin's *projection method* [4] because of its simplicity and applicability to collocated grids. We wish to mention that the Poisson problem (in the projection method) for a doubly-periodic domain has an infinite number of solutions differed by a constant. To solve it numerically, we use the Discrete Fourier Transform (DFT) for the flat bottom case, while for the non-flat bottom case, we regularize the problem by a time derivative in the pseudo-time τ and seek the stationary solution.

3 Main Result: Asymptotic Analysis of the Scheme

Theorem 1. *Consider the shallow water equations (5) with topography in a periodic domain and with well-prepared initial data* $(z_{0,\varepsilon}, \boldsymbol{m}_{0,\varepsilon})$ *such that*

$$
z(0,\cdot) = z_{0,\varepsilon} = z^0_{(0)} + \varepsilon^2 z^0_{(2),\varepsilon}, \qquad \boldsymbol{m}(0,\cdot) = \boldsymbol{m}_{0,\varepsilon} = \boldsymbol{m}^0_{(0)} + \varepsilon \boldsymbol{m}^0_{(1),\varepsilon},
$$

where $z^0_{(0)}$ *is a constant and* $\boldsymbol{m}^0_{(0)}$ *satisfies the lake equations (10). Then, the RS-IMEX scheme (8a)–(8b) is solvable, i.e., it has a unique solution for all* $\varepsilon > 0$, *if* $\widetilde{\alpha}$ *is constant. Also, the scheme is consistent with the asymptotic limit in the fully-discrete settings, i.e., it is asymptotically consistent.*

3.1 Solvability

Assuming $\Delta x = \Delta y$ and $\widetilde{\alpha} = 0$ for simplicity, the linear system of the implicit step (8b) with the companion matrix J_ε can be written as $J_\varepsilon := \mathbb{I}_{3N^2} + \beta\,\Xi_\varepsilon$, where $\beta := \frac{\Delta t}{2\Delta x}$ and Ξ_ε is a matrix not depending on β. It is plausible to conclude that for a suitable choice of β, none of the eigenvalues of $\beta\,\Xi_\varepsilon$ are equal to -1; so J_ε is non-

singular, and the implicit step (so the whole scheme) is solvable. The proof for $\widetilde{\alpha} \neq 0$ is likewise.

3.2 Asymptotic Consistency

The asymptotic consistency analysis is often done formally in the literature, namely by putting the Poincaré expansion ansatz into the scheme and by balancing the equal powers of ε. For the present work, we adopt the same approach.

Firstly, we show that the explicit step is "ε-*stable*", i.e., $\|\boldsymbol{V}_{\Delta}^{n+\frac{1}{2}}\| = \mathscr{O}(1)$. Given $\|\boldsymbol{V}_{\Delta}^{n}\| = \mathscr{O}(1)$, which is compatible with the well-prepared initial data, and since $\widehat{\boldsymbol{G}}_{1,1} = \widehat{\boldsymbol{G}}_{2,1} = 0$, one can immediately conclude that $\|\boldsymbol{V}_{1,\Delta}^{n+\frac{1}{2}}\| = \mathscr{O}(1)$. For $\boldsymbol{V}_{2,\Delta}$ (and similarly $\boldsymbol{V}_{3,\Delta}$), one can simply confirm that

$$\lim_{\varepsilon \to 0} \left(\nabla_{h,x} \widehat{\boldsymbol{G}}_{1,2,ij}^{n} + \nabla_{h,y} \widehat{\boldsymbol{G}}_{2,2,ij}^{n} \right) = \mathscr{O}(1), \tag{12}$$

since

$$\begin{aligned} \lim_{\varepsilon \to 0} \Bigg[\nabla_{h,x} \Bigg(& \frac{m_1^2}{z-b} + \frac{z^2 - 2zb}{2\varepsilon^2} - \frac{\overline{m_1}^2}{\overline{z}-b} - \frac{\overline{z}^2 - 2\overline{z}b}{2\varepsilon^2} + \frac{\overline{m_1}^2 v_1 \varepsilon^2}{(\overline{z}-b)^2} - \frac{2\overline{m_1} v_2}{\overline{z}-b} - (\overline{z}-b) v_1 \Bigg) \\ & + \nabla_{h,y} \Bigg(\frac{m_1 m_2}{z-b} - \frac{\overline{m_1 m_2}}{\overline{z}-b} + \frac{\overline{m_1 m_2} v_1 \varepsilon^2}{(\overline{z}-b)^2} - \frac{\overline{m_1} v_3}{\overline{z}-b} - \frac{\overline{m_2} v_2}{\overline{z}-b} \Bigg) \Bigg] = \mathscr{O}(1). \end{aligned}$$

So, the explicit step does not change the leading order of $\boldsymbol{V}_{2,\Delta}^{n}$ (and $\boldsymbol{V}_{3,\Delta}^{n}$). This concludes the ε-stability proof of the explicit step.

Completing the asymptotic consistency analysis, we show that the implicit step is consistent with the limit. We *assume* that $\|\boldsymbol{V}_{\Delta}^{n+1}\| = \mathscr{O}(1)$ to justify the use of Poincaré expansion and will discuss this assumption somewhere else. From the v_1-update, (11) and (7a)–(7d), the momentum field (up to $\mathscr{O}(\varepsilon^2)$) is solenoidal, i.e.,

$$\nabla_{h,x} \left(\overline{m_1} + v_2\right)_{ij}^{n+1} + \nabla_{h,y} \left(\overline{m_2} + v_3\right)_{ij}^{n+1} = \mathscr{O}(\varepsilon^2). \tag{13}$$

Since the consistency of the evolution of the leading order of the momentum is clear, the asymptotic consistency of the scheme is concluded, but only up to possible oscillations for the momentum field in the null space of central difference operators $\nabla_{h,x}$ and $\nabla_{h,y}$ which may lead to checker-board oscillations.

Remark 1. The Eq. (13), combined with the v_1-update, immediately implies that possible checker-board oscillations for the surface perturbation z are small, i.e., $\mathscr{O}(\varepsilon^2)$. This seems to solve the problem in [15] regarding the checker-board oscillations in a periodic domain and suggests that it may not be necessary to add a large diffusion in order to preclude oscillations.

Table 1 Experimental order of convergence with CFL $= 0.45$ and for different ε. Error e is defined with the exact solution in ℓ_∞-norm

N	$\varepsilon = 0.8$				N	$\varepsilon = 10^{-6}$			
	e_{z,ℓ_∞}	EOC_{z,ℓ_∞}	e_{u_1,ℓ_∞}	EOC_{u_1,ℓ_∞}		e_{z,ℓ_∞}	EOC_{z,ℓ_∞}	e_{u_1,ℓ_∞}	EOC_{u_1,ℓ_∞}
20	2.61e-2	–	1.04e-1	–	20	4.08e-14	–	1.04e-1	–
40	2.00e-2	0.38	6.80e-2	0.61	40	3.13e-14	0.38	6.80e-2	0.61
80	1.23e-2	0.70	3.63e-2	0.91	80	1.92e-14	0.71	3.63e-2	0.91
160	6.20e-3	0.99	1.65e-3	1.14	160	9.69e-15	0.99	1.65e-3	1.14

4 Numerical Results

We discuss the traveling vortex example [2] to verify the quality of the solutions computed by the RS-IMEX scheme. We consider a well-prepared initial condition in the periodic domain $\Omega = [0, 1)^2$:

$$
\begin{aligned}
z(x, y, 0) &= \mathbf{1}_{[r \le \frac{\pi}{\omega}]} \left(\frac{\Gamma \varepsilon}{\omega} \right)^2 (g(\omega r) - g(\pi)) , \\
u_1(x, y, 0) &= u_0 + \mathbf{1}_{[r \le \frac{\pi}{\omega}]} \Gamma \, (1 + \cos(\omega r)) \, (y_c - y), \\
u_2(x, y, 0) &= \mathbf{1}_{[r \le \frac{\pi}{\omega}]} \Gamma \, (1 + \cos(\omega r)) \, (x - x_c),
\end{aligned}
$$

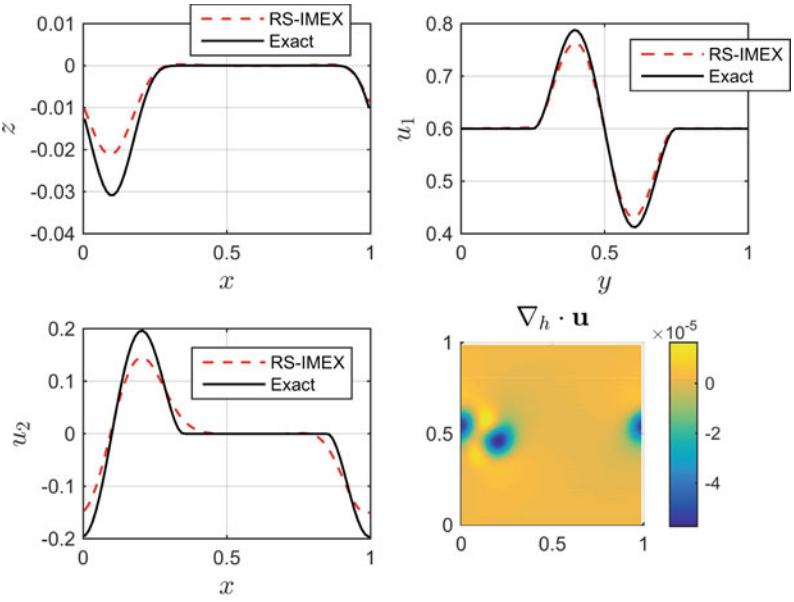
with $H_{mean} = 110$, $u_0 = 0.6$, $\boldsymbol{x}_c = (0.5, 0.5)^T$, $\Gamma = 1.4$, $\omega = 4\pi$, $r := \|\boldsymbol{x} - \boldsymbol{x}_c\|$ and

$$
g(r) := 2 \cos r + 2r \sin r + \frac{1}{8} \cos 2r + \frac{r}{4} \sin 2r + \frac{3}{4} r^2 .
$$

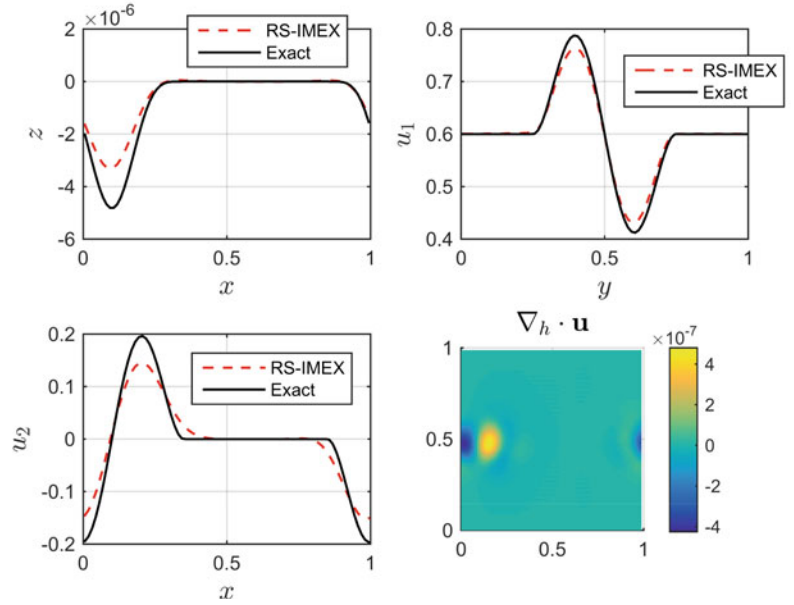
We choose the time step as $\Delta t := \text{CFL} \, \Delta x / \widehat{\alpha}$. The exact solution is the initial condition advected by u_0 with time-periodicity $T_\pi = \frac{5}{3}$ such that $w(x, y, t) = w(x - u_0 t, y, 0)$ for $w \in \{z, u_1, u_2\}$. Using this exact solution, Table 1 shows the experimental order of convergence (EOC) for the final time $T_f = 1$ and for different ε; it is clear that the EOC is close to one uniformly in ε and the scheme is accurate for all $\varepsilon > 0$. We also illustrate this fact in Fig. 1, where both exact and numerical solutions are plotted on centerlines of the domain.

Figure 2a illustrates the computed solution for an small ε, in particular $\varepsilon = 10^{-6}$. There is a very good agreement between the result of the RS-IMEX scheme and the exact solution. It is also clear that there is no checker-board oscillation for the momentum and surface perturbation. These suggest that the scheme is asymptotically consistent and stable. Moreover, Fig. 2b shows that the scaled perturbation is bounded in terms of ε; so, the formal asymptotic consistency analysis is justified.

Acknowledgements The research was supported by RWTH Aachen University through *Graduiertenförderung nach Richtlinien zur Förderung des wissenschaftlichen Nachwuchses (RFwN).*

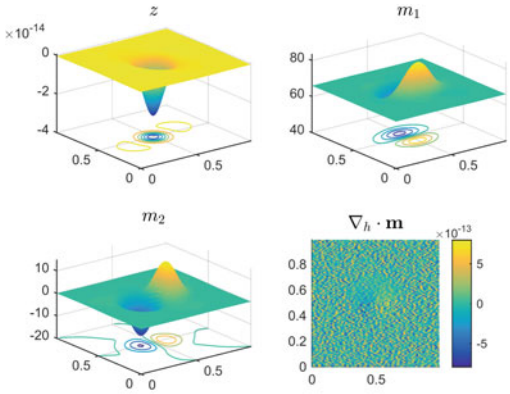


(a) $\varepsilon = 0.8$.

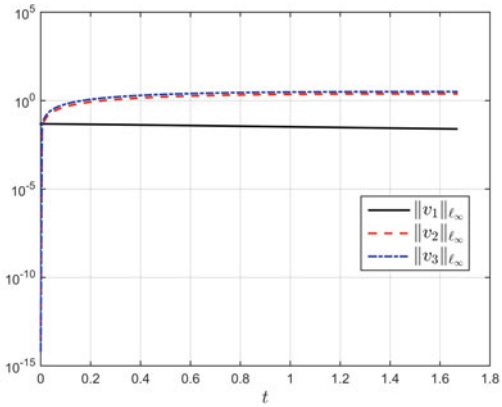


(b) $\varepsilon = 0.01$.

Fig. 1 Error of the RS-IMEX scheme for different ε on the 80×80 grid, with CFL $= 0.45$ and $T_f = 1$



(a) Solution of the RS-IMEX scheme for $\varepsilon = 10^{-6}$.



(b) Time evolution of the norm of the perturbation from the incompressible solution for $\varepsilon = 10^{-6}$. The figure is almost the same for $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-2}$.

Fig. 2 Behavior of the scheme on the 100×100 grid with CFL $= 0.45$ and $T_f = T_\pi$

References

1. G. Bispen, IMEX finite volume methods for the shallow water equations. Ph.D. thesis, Johannes Gutenberg-Universität, 2015
2. G. Bispen, K.R. Arun, M. Lukáčová-Medvid'ová, S. Noelle, IMEX large time step finite volume methods for low Froude number shallow water flows. Commun. Comput. Phys. **16**, 307–347 (2014)
3. D. Bresch, R. Klein, C. Lucas, Multiscale analyses for the shallow water equations, in *Computational Science and High Performance Computing IV* (Springer, 2011), pp. 149–164
4. A.J. Chorin, Numerical solution of the Navier-Stokes equations. Math. Comput. **22**, 745–762 (1968)
5. P. Degond, M. Tang, All speed scheme for the low Mach number limit of the isentropic Euler equation. Commun. Comput. Phys. **10**, 1–31 (2011)
6. S. Dellacherie, Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. J. Comput. Phys. **229**, 978–1016 (2010)

7. L. Even-Dar Mandel, S. Schochet, Convergence of solutions to finite difference schemes for singular limits of nonlinear evolutionary PDEs. ESAIM Math. Model. Numer. Anal. Modél. Math. et Anal. Numér. (2016). https://doi.org/10.1051/m2an/2016029
8. L. Even-Dar Mandel, S. Schochet, Uniform discrete Sobolev estimates of solutions to finite difference schemes for singular limits of nonlinear PDEs. ESAIM Math. Model. Numer. Anal. Modél. Math. et Anal. Numér. (2016). https://doi.org/10.1051/m2an/2016038
9. F. Filbet, S. Jin, A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. J. Comput. Phys. **229**, 7625–7648 (2010)
10. J. Giesselmann, Low Mach asymptotic-preserving scheme for the Euler-Korteweg model. IMA J. Numer. Anal. **35**, 802–833 (2015)
11. F.X. Giraldo, M. Restelli, High-order semi-implicit time-integrators for a triangular discontinuous Galerkin oceanic shallow water model. Int. J. Numer. Methods Fluids **63**, 1077–1102 (2010)
12. J. Haack, S. Jin, J.-G. Liu, An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations. Commun. Comput. Phys. **12**, 955–980 (2012)
13. S. Jin, Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. SIAM J. Sci. Comput. **21**, 441–454 (1999)
14. S. Jin, Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. Lecture Notes for Summer School on "Methods and Models of Kinetic Theory" (M&MKT), Porto Ercole (Grosseto, Italy) (2010), pp. 177–216
15. K. Kaiser, J. Schütz, R. Schöbel, S. Noelle, A new stable splitting for the isentropic Euler equations. J. Sci. Comput. 1–18 (2016)
16. N. Masmoudi, Examples of singular limits in hydrodynamics. Handb. Differ. Eqn. Evol. Eqn. **3**, 195–275 (2007)
17. S. Noelle, G. Bispen, K.R. Arun, M. Lukáčová-Medvid'ová, C.-D. Munz, A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics. SIAM J. Sci. Comput. **36**, B989–B1024 (2014)
18. H. Zakerzadeh, Asymptotic analysis of the RS-IMEX scheme for the shallow water equations in one space dimension, HAL: hal-01491450, IGPM report 455 (RWTH Aachen University, Submitted for publication, 2016)
19. H. Zakerzadeh, On the Mach-uniformity of the Lagrange-projection scheme. ESAIM Math. Model. Numer. Anal. Modél. Math. et Anal. Numér. (2016). https://doi.org/10.1051/m2an/2016064

Class of Space-Time Entropy Stable DG Schemes for Systems of Convection–Diffusion

Georg May and Mohammad Zakerzadeh

Abstract In this work, we present a family of entropy stable discontinuous Galerkin methods for systems of convection–diffusion with nonlinear convective and viscous fluxes. The discretization presented here is based on a mixed formulation and is designed to preserve entropy stability of an already existing entropy stable discretization for a hyperbolic system of conservation laws. The fully discrete version of the entropy stability is proven in the framework of space–time formulation with variants of several known schemes, including the method of Bassi–Rebay and symmetric interior penalty method.

Keywords Discontinuous Galerkin · Entropy stable · Convection–diffusion

1 Introduction

In analyzing the stability of numerical schemes for nonlinear systems of hyperbolic conservation laws, entropy stability is often the framework of choice [2, 5, 9, 12, 13, 16]. This is typically done by realization of the numerical scheme in terms of the so-called entropy variables.

Compared to entropy stability analysis of hyperbolic conservation laws, much less work has been done in extending entropy stability of discontinuous Galerkin (DG) schemes to nonlinear convection–diffusion systems. To the best knowledge of the authors, the only available results in this direction are [5, 8, 14, 15]. More specifically, in [5], the symmetric/non-symmetric interior penalty DG (SIPG/NIPG) formulation has been presented for the one-dimensional Navier–Stokes equations realized in terms of entropy variables, and the entropy stability has been proved in the semi-discrete form. In [14, 15], a formulation of SIPG/NIPG as well as a

G. May · M. Zakerzadeh (✉)
AICES Graduate School, RWTH Aachen University, 52062 Aachen, Germany
e-mail: may@aices.rwth-aachen.de

M. Zakerzadeh
e-mail: zakerzadeh@aices.rwth-aachen.de

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_51

variant of LDG method [7] have been proposed for the one-dimensional systems of convection–diffusion and in space–time framework. Moreover, the proof of entropy stability in space–time fully discrete settings is provided. The authors of [8] present an entropy stable space–time formulation for turbulent computations using the entropy variables; nevertheless, they did not provide a rigorous proof of the entropy stability.

In this work, we consider a general class of DG formulations for multidimensional systems of convection–diffusion. This is done by using a mixed method for discretizing the viscous flux, combined with an entropy stable formulation for the convective flux which already proposed in [11, 17]. Moreover, we have already published a similar formulation in [19] without specifying the time marching technique and with a proof of semi-discrete entropy stability. Here, and by the reformulation in a space-time framework, we extend the stability result of [17] to convection–diffusion systems and of [19] to fully discrete form. The current work is also similar to [14, 15] in using space-time framework; however, unlike them, the different viscous formulations in this work are obtained from a canonical framework. Here, we only consider a version of SIPG and BR2 [4] methods and refer to [19] for more formulations.

We consider general systems of convection–diffusion as the following form

$$\frac{\mathrm{d}u}{\mathrm{d}t} + \nabla \cdot \big(f_c(u) - f_v(u, \nabla u)\big) = 0, \qquad \text{in } \Omega, \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ is bounded. Here, $u \in \mathbb{R}^m$ is the vector of conservative variables, and by f_c and f_v, we denote the convective and viscous fluxes. Also, we consider that f_v is linear with respect to its second argument. Hence, one can write (1) as

$$u_t + A_i(u)\nabla_i u - \nabla \cdot (K_{i,j}(u)\nabla_j u) = 0, \qquad i, j = 1, \cdots, d, \tag{2}$$

where $A_i(u) = \partial_u f_c^i(u)$ and $K_{i,j} \in \mathbb{R}^{m\times m}$ and the repeated indices denote the summation.

In general, the matrices $A_i(u)$ and $K(u) = \big[K_{i,j}(u)\big]$ are not symmetric which makes the energy analysis somewhat cumbersome. An idea to detour this issue is the transform to the entropy variables; i.e., using the change of variables $v(u) = U_u(u)$, where U is a strictly convex entropy function of the corresponding hyperbolic system, i.e., when $f_v \equiv 0$.

This symmetrization is a well-known approach for hyperbolic systems of conservation laws (see [16]), and here, we assume that it also symmetrizes the diffusion matrix $K(u)$. Consequently, and by abusing the notation A and K to denote the transformed matrices of (2), the symmetric representation of (2) realized in terms of entropy variables reads

$$u_v v_t + A_i(v)\nabla_i v - \nabla \cdot (K_{i,j}(v)\nabla_j v) = 0, \qquad i, j = 1, \cdots, d, \tag{3}$$

such that $A_i(v)$ is symmetric and $K(v)$ is symmetric positive semi-definite. Note that u_v is symmetric positive definite by the properties of the convex entropy function.

The system of compressible Navier–Stokes equations is an example of systems that accepts such a nice symmetric form as (3). We are going to discuss it briefly later in Sect. 4, and we refer for more details to [12].

Henceforth, we realize the functions in terms of entropy variables v which are the basic unknowns, and the dependent conservative variables are derived via mapping $u(v)$. This mapping is sometimes omitted in notation, e.g., $f(v)$ rather than $f(u(v))$.

The outline of this paper is as follows: The DG discretization is introduced in Sect. 2. This section includes the explicit form of the convective and different viscous discretizations. The stability analysis is presented in Sects. 3, and 4 is reserved for numerical examples.

2 Space–Time Discontinuous Galerkin Scheme

In this section, we present the space–time DG discretization of (1). Let us define the space–time coordinates $\overline{x} = (x_0, x_1, \ldots, x_d)$ with $x_0 \equiv t$. Now, with the notation of $\overline{\nabla}$ as the space–time gradient, one might reformulate the symmetrized problem (3) in the space–time conservative form as

$$\overline{\nabla} \cdot \left(\bar{f}_c(v) - \bar{f}_v(v, \overline{\nabla} v) \right) = 0, \qquad \text{in } \Omega \times [0, \infty), \tag{4}$$

where the space–time convective and viscous fluxes are

$$\bar{f}_c(v) = \begin{bmatrix} u(v) \\ f(u(v)) \end{bmatrix}, \qquad \bar{f}_v(v, \overline{\nabla} v) = \begin{bmatrix} 0 & 0 \\ 0 & K \end{bmatrix} \overline{\nabla} v = \bar{K}(v) \overline{\nabla} v. \tag{5}$$

For arbitrary final time $T > 0$, consider a sequence of time instances $0 \equiv t_0 < t_1 < \cdots < t_N \equiv T$, with corresponding time intervals $I_n = (t_n, t_{n+1})$. Let us consider $\partial\Omega$ be a polygon and $\mathscr{T}_h^n = \{\kappa\}$ be a shape-regular subdivision of the space–time slab $I_n \times \Omega$ into disjoint $(d+1)$-simplices. In order to avoid technicalities of boundary conditions, we consider periodic boundary conditions on $\partial\Omega \times [0, T]$. The treatment of the temporal boundaries of the space–time slab, i.e., $\{t_n, t_{n+1}\} \times \Omega$, will be discussed later.

Define $\mathscr{T}_h := \bigcup_{n=0}^{N-1} \mathscr{T}_h^n$, and let $h := \sup_{\kappa \in \mathscr{T}_h} h_\kappa$, where $h_\kappa := diam(\kappa)$. Also, we denote n_κ to be the outward normal to $\partial\kappa$, and $d' = d + 1$ the dimension of space–time.

We assume that $\mathscr{T}_h$ is of *bounded variation*; that is, there exists a constant $l > 1$ such that $l^{-1} \leq \frac{h_\kappa}{h_{\kappa'}} \leq l$, where $\kappa, \kappa' \in \mathscr{T}_h$ share an edge. This property means that there is an upper bound for the number of neighboring elements, denoted by N_l.

We denote the *skeleton* of the triangulation, i.e., the set of all d-dimensional faces of $\kappa \in \mathscr{T}_h$, by $\mathscr{E}_h = \{e\}$ and the diameter of e by h_e. Also, we denote the set of temporal faces as $\mathscr{E}_{h,t}$ and $\mathscr{E}_{h,i} = \mathscr{E}_h \setminus \mathscr{E}_{h,t}$, respectively.

Let us fix the definition of the jump and average of discontinuous functions on the skeleton $\mathscr{E}_h$. For any $e \in \mathscr{E}_{h,i}$, where e is the common face of κ, κ', with $w_{\kappa,e} = w_\kappa|_e$, we set

$$\{w\} = \frac{1}{2}(w_{\kappa,e} + w_{\kappa',e}), \qquad [\![w]\!] = w_{\kappa,e} \otimes n_\kappa + w_{\kappa',e} \otimes n_{\kappa'} \tag{6}$$

for all $w \in \prod_{\kappa\in\mathscr{T}_h} \left[L_2(\partial\kappa)\right]^m$. Similarly for all $\tau \in \prod_{\kappa\in\mathscr{T}_h} \left[L_2(\partial\kappa)\right]^{m\times d'}$ we set

$$\{\tau\} = \frac{1}{2}(\tau_{\kappa,e} + \tau_{\kappa',e}), \qquad [\![\tau]\!] = \tau_{\kappa,e} \cdot n_\kappa + \tau_{\kappa',e} \cdot n_{\kappa'}. \tag{7}$$

Moreover, for any boundary edge $e \in \mathscr{E}_{h,t}$, we define

$$[\![w]\!] = w_{\kappa,e} \otimes n_\kappa, \qquad \{\tau\} = \tau_{\kappa,e}. \tag{8}$$

We will use several different notations for inner product; $\langle w, v\rangle$ denotes the inner product between $w, v \in \mathbb{R}^m$, while $a \cdot b$ defines the inner product for $a, b \in \mathbb{R}^{d'}$. Moreover, for the Frobenius inner product, we use the notation $\tau : \zeta = \sum_{i,j} \tau_{i,j}\zeta_{i,j}$ for $\tau, \zeta \in \mathbb{R}^{m\times d}$. For $w \in \mathbb{R}^m$ and $a \in \mathbb{R}^d$, we define the outer product $w \otimes a \in \mathbb{R}^{m\times d}$ as $[w \otimes a]_{i,j} = w_i a_j$.

Furthermore, the finite dimensional space for the approximate solution is

$$V_{h,q}(\mathscr{T}_h) := \{w^h \in [L_2(\Omega)]^m : w^h|_\kappa \in [\mathscr{P}^q(\kappa)]^m, \quad \forall \kappa \in \mathscr{T}_h\}, \tag{9}$$

where $\mathscr{P}^q(\kappa)$ is the space of polynomials of at most degree q on a domain κ.

The proposed DG method has the following semi-linear variational form: Find $v^h \in V_{h,q}$ such that

$$\mathscr{B}(v^h, w^h) := \mathscr{B}^c(v^h, w^h) + \mathscr{B}^v(v^h, w^h) = 0, \qquad \forall w^h \in V_{h,q}. \tag{10}$$

Here, $\mathscr{B}^c$ and $\mathscr{B}^v$ correspond to the convective and viscous discretization of (4), respectively. We are going to present the details of these discretizations later in this section.

In some related works like [11, 15, 17], one or two stabilization terms, in form of *shock capturing* and *streamline diffusion*, were added to the DG discretization (10). Despite their beneficial role in alleviating the oscillations, here, we only focus on smooth solutions and neglect these terms.

2.1 Convective Discretization

Introducing the shorthand notation $\sum_{n,\kappa} := \sum_{n=0}^{N-1} \sum_{\kappa\in\mathscr{T}_h^n}$, we define the convective semi-linear form as

$$\mathscr{B}^c(v^h, w^h) := \sum_{n,\kappa} \left\{ \int_\kappa \langle \overline{\nabla} \cdot \bar{f}_c(v^h), w^h\rangle \,\mathrm{d}x + \int_{\partial\kappa} \langle \hat{f}_c(v^h) - \bar{f}(v^h_{\kappa,e}) \cdot n_\kappa, w^h\rangle \,\mathrm{d}s \right\}, \tag{11}$$

for $w^h \in V_{h,q}$. The numerical flux function $\hat{f}_c(v^h) \equiv \hat{f}_c(v^h_{\kappa,e}, v^h_{\kappa',e}; n_\kappa)$ is set to be Lipschitz continuous, conservative, and consistent with $\bar{f}_c$. On faces where $n_\kappa = (\pm 1, 0, \ldots, 0)^t$, this numerical flux should reduce to pure upwinding (i.e. in time). This upwinding leads to decoupling the time slabs from each other.

In general, $\hat{f}_c$ is required to be entropy stable; i.e., it has the following viscosity form, for any $e \in \mathscr{E}_h^\circ$

$$\hat{f}_c(v^h_{\kappa,e}, v^h_{\kappa',e}; n_\kappa) = f^*(v^h_{\kappa,e}, v^h_{\kappa',e}; n_\kappa) - \frac{1}{2} D(v^h) [\![v^h]\!], \tag{12}$$

where f^* denotes the *entropy conservative flux*, and D is an entropy dissipation matrix required to be symmetric and uniformly positive definite.

We refer to the literature for more details on the entropy stable fluxes, as well as explicit forms of such fluxes for the Euler and shallow water equations [9, 13, 16].

2.2 Viscous Discretization

For the discretization of the viscous flux in (4), we follow the approach presented in [6] and consider a first-order mixed formulation of (4) for three unknown variables; v, $\theta = \overline{\nabla} v$, and $\sigma = \bar{K}\theta$ as

$$-\overline{\nabla} \cdot \sigma = R, \qquad \sigma = \bar{K}\theta, \qquad \theta = \overline{\nabla} v, \qquad x \in \Omega \times [0, T], \tag{13}$$

where $R := -\overline{\nabla} \cdot \bar{f}_c(v)$ is the remainder of (4). In this section, we present a primal formulation of (13) which can easily fit into (10). We approximate the solution of (13) by discrete functions $(\sigma^h, \theta^h, u^h)$ in the finite element space $(\Sigma_{h,p} \times \Sigma_{h,p} \times V_{h,q})$, where $V_{h,q}$ is defined by (9) and $\Sigma_{h,p}$ is

$$\Sigma_{h,p}(\mathscr{T}_h) := \{\theta^h \in [L_2(\Omega)]^{m \times d'} : \theta^h|_\kappa \in [\mathscr{P}^p(\kappa)]^{m \times d'}, \quad \forall \kappa \in \mathscr{T}_h\}, \tag{14}$$

with $q \geq 1$ and $p = q$ or $p = q - 1$, in order to satisfy the property $\nabla V_{h,q} \subset \Sigma_{h,p}$ (cf. [1]). In the computational code, we use $p = q$.

Let us consider the following weak formulation of (13)

$$\sum_{n,\kappa} \int_\kappa \bar{K}\theta^h : \zeta^h \, \mathrm{d}\overline{x} = \sum_{\kappa \in \mathscr{T}_h} \int_\kappa \sigma^h : \zeta^h \, \mathrm{d}\overline{x}, \tag{15}$$

$$\sum_{n,\kappa} \int_\kappa \theta^h : \tau^h \, \mathrm{d}\overline{x} + \sum_{\kappa \in \mathscr{T}_h} \int_\kappa \langle v^h, \nabla \cdot \tau^h \rangle \, \mathrm{d}\overline{x} = \sum_{\kappa \in \mathscr{T}_h} \int_{\partial\kappa} \langle \hat{v}, \tau^h \cdot n_\kappa \rangle \, \mathrm{d}s, \tag{16}$$

$$\sum_{n,\kappa} \int_\kappa \sigma^h : \overline{\nabla} w^h \, \mathrm{d}\overline{x} - \sum_{\kappa \in \mathscr{T}_h} \int_{\partial\kappa} \langle \hat{\sigma} \cdot n_\kappa, w^h \rangle \, \mathrm{d}s = (R, w^h), \tag{17}$$

for any (ζ^h, τ^h, w^h) in $\Sigma_{h,p} \times \Sigma_{h,p} \times V_{h,q}$ and by noting that $(R, w^h) = -\mathscr{B}^c(v^h, w^h)$. We present the explicit form of the numerical fluxes $\hat{v}$ and $\hat{\sigma}$ later.

From [1], we have the following identity; for any $v \in \prod_{\kappa \in \mathscr{T}_h} \left[L_2(\partial\kappa)\right]^m$ and $\xi \in \prod_{\kappa \in \mathscr{T}_h} [L_2(\partial\kappa)]^{m \times d'}$, the following holds

$$\sum_{n,\kappa} \int_{\partial\kappa} \langle v, \xi \cdot n_\kappa \rangle \, \mathrm{d}s = \sum_{e \in \mathscr{E}_{h,i}} \int_e \langle \{v\}, [\![\xi]\!] \rangle \, \mathrm{d}s + \sum_{e \in \mathscr{E}_h} \int_e [\![v]\!] : \{\xi\} \, \mathrm{d}s. \tag{18}$$

Applying (18) in (16) and (17), one can write

$$\sum_{n,\kappa} \int_\kappa \theta^h : \tau^h \, \mathrm{d}\overline{x} = \sum_{n,\kappa} \int_\kappa \overline{\nabla} v^h : \tau^h \, \mathrm{d}\overline{x} - \sum_{e \in \mathscr{E}_{h,i}} \int_e \langle \{v^h - \hat{v}\}, [\![\tau^h]\!] \rangle \, \mathrm{d}s - \sum_{e \in \mathscr{E}_h} \int_e [\![v^h - \hat{v}]\!] : \{\tau^h\} \, \mathrm{d}s, \tag{19}$$

$$\sum_{n,\kappa} \int_\kappa \sigma^h : \overline{\nabla} w^h \, \mathrm{d}\overline{x} - \sum_{e \in \mathscr{E}_h} \int_e \{\hat{\sigma}\} : [\![w^h]\!] \, \mathrm{d}s - \sum_{e \in \mathscr{E}_{h,i}} \int_e \langle [\![\hat{\sigma}]\!], \{w^h\} \rangle \, \mathrm{d}s = (R, w^h). \tag{20}$$

Let us define the following Galerkin projection $\mathscr{G}_h : [L_2(\Omega)]^{m \times d'} \to \Sigma_{h,p}$; for all $\xi \in [L_2(\Omega)]^{m \times d'}$

$$\sum_{n,\kappa} \int_\kappa \xi : \tau \, \mathrm{d}\overline{x} = \sum_{n,\kappa} \int_\kappa \mathscr{G}_h(\xi) : \tau \, \mathrm{d}\overline{x}, \qquad \forall \tau \in \Sigma_{h,p}. \tag{21}$$

Moreover, we need a global and an edge-wise lifting operator, $r : [L_2(\mathscr{E}_{h,i})]^{m \times d'} \to \Sigma_{h,p}$ and $r^e : [L_2(e)]^{m \times d'} \to \Sigma_{h,p}$ as,

$$\sum_{n,\kappa} \int_\kappa r(\varphi) : \tau \, \mathrm{d}\overline{x} = -\sum_{e \in \mathscr{E}_{h,i}} \int_e \varphi : \{\tau\} \, \mathrm{d}s, \quad \sum_{n,\kappa} \int_\kappa r^e(\varphi) : \tau \, \mathrm{d}\overline{x} = -\int_e \varphi : \{\tau\} \, \mathrm{d}s, \tag{22}$$

for any $\tau \in \Sigma_{h,p}$ and $e \in \mathscr{E}_{h,i}$. The difference of the definition (22) with the more standard definition in [19] is in excluding the temporal boundary terms which do not add any viscous contributions; see the structure of $\bar{K}$ in (5). Note that $r(\varphi) = \sum_{e \in \mathscr{E}_{h,i}} r^e(\varphi)$.

Here, we consider two different options for the numerical fluxes $\hat{v}$ and $\hat{\sigma}$. Other choices such as BR1 [3] and LDG can be done similarly (cf. [19]).

(i) BR2: Here, $\hat{v}$ and $\hat{\sigma}$ are set as [4], for $e \in \mathscr{E}_h$

$$\hat{v} = \begin{cases} \{v^h\} & e \in \mathscr{E}_{h,i} \\ v^h_{\kappa,e} & e \in \mathscr{E}_{h,t} \end{cases}, \quad \hat{\sigma} = \begin{cases} \{\mathscr{G}_h\big(\bar{K}(v^h)(\nabla v^h + \eta_e r^e([\![v^h]\!]))\big)\} & e \in \mathscr{E}_{h,i} \\ 0 & e \in \mathscr{E}_{h,t} \end{cases}.$$

The parameter η_e depends only on the properties of the triangulation. The appropriate choice for this parameter will be presented later in Sect. 3.

(ii) SIPG: In this formulation, we set for $e \in \mathscr{E}_h$

$$\hat{v} = \begin{cases} \{v^h\} & e \in \mathscr{E}_{h,i} \\ v^h_{\kappa,e} & e \in \mathscr{E}_{h,t} \end{cases}, \qquad \hat{\sigma} = \begin{cases} \{\mathscr{G}_h(\bar{K}(v^h)\nabla v^h)\} - \dfrac{\mu_e}{h_e}[\![v^h]\!] & e \in \mathscr{E}_{h,i} \\ 0 & e \in \mathscr{E}_{h,t} \end{cases}$$

with some $\mu_e > 0$ dependent on the type of the triangulation, polynomial order, and the diffusion matrix. We state the criterion for μ_e in Sect. 3.

Remark 1. The choice of the boundary flux on $\mathscr{E}_{h,t}$ stems from the fact that there is no diffusion in the time direction. Hence, no coupling should be enforced.

Using the definition of the numerical fluxes (of both BR2 and SIPG) in (19) and (20), one might simplify the weak formulation as

$$\sum_{n,\kappa} \int_\kappa \theta^h : \tau^h \, \mathrm{d}\overline{x} = \sum_{n,\kappa} \int_\kappa \overline{\nabla} v^h : \tau^h \, \mathrm{d}\overline{x} + \sum_{e \in \mathscr{E}_{h,i}} \int_e [\![v^h]\!] : \{\tau^h\} \, \mathrm{d}s = 0, \tag{23}$$

$$\sum_{n,\kappa} \int_\kappa \sigma^h : \overline{\nabla} w^h \, \mathrm{d}\overline{x} - \sum_{e \in \mathscr{E}_h} \int_e \hat{\sigma} : [\![w^h]\!] \, \mathrm{d}s = (R, w^h). \tag{24}$$

In order to obtain the primal formulation, using (15) and (21), one can solve σ^h as

$$\sigma^h = \mathscr{G}_h\big(\bar{K}(v^h)\theta^h\big), \tag{25}$$

and consequently, using (22), (23) and (25) give

$$\theta^h = \overline{\nabla} v^h + r([\![v^h]\!]), \qquad \sigma^h = \mathscr{G}_h\Big(\bar{K}(v^h)\big(\overline{\nabla} v^h + r([\![v^h]\!])\big)\Big). \tag{26}$$

Now (θ^h, σ^h) can be solved locally in terms of v^h by inserting (26) in (24), and the corresponding primal formulations are obtained as the following, for BR2

$$\begin{aligned} \mathscr{B}^v(v^h, w^h) = &\sum_{n,\kappa} \int_\kappa \bar{K}(v^h)\overline{\nabla} v^h : \overline{\nabla} w^h \, \mathrm{d}\overline{x} + \sum_{e \in \mathscr{E}_{h,i}} \eta_e \sum_{n,\kappa} \int_\kappa \bar{K}(v^h) r^e([\![v^h]\!]) : r^e([\![w^h]\!]) \, \mathrm{d}\overline{x} \\ &+ \sum_{n,\kappa} \int_\kappa \Big(\bar{K}(v^h) r([\![v^h]\!]) : \overline{\nabla} w^h + \bar{K}(v^h)\overline{\nabla} v^h : r([\![w^h]\!])\Big) \, \mathrm{d}\overline{x}, \end{aligned} \tag{27}$$

and for SIPG

$$\mathscr{B}^v(v^h, w^h) = \sum_{n,\kappa} \int_\kappa \bar{K}(v^h)\overline{\nabla} v^h : \overline{\nabla} w^h \, \mathrm{d}\overline{x} + \sum_{e\in\mathscr{E}_{h,i}} \frac{\mu_e}{h_e} \int_e [\![v^h]\!] : [\![w^h]\!] \, \mathrm{d}s.$$
$$+ \sum_{n,\kappa} \int_\kappa \left(\bar{K}(v^h) r([\![v^h]\!]) : \overline{\nabla} w^h + \bar{K}(v^h)\overline{\nabla} v^h : r([\![w^h]\!]) \right) \mathrm{d}\overline{x}. \quad (28)$$

Let us remark that due to the discrete nature of the projection (21) and lifting operators (22), the viscous formulations presented here are inconsistent with the exact solution of (1) and consequently adjoint inconsistent. However, in the asymptotic limit of the mesh refinement, the consistency and adjoint consistency can be recovered, provided that the exact solution is sufficiently smooth [18].

3 Entropy Stability

Similar to [19], and by taking inner product of (1) with respect to the entropy variables, the following *global entropy inequality* can be formally obtained using the positive semi-definiteness of $\bar{K}(v)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega U(u) \, \mathrm{d}x \leq 0, \quad (29)$$

on a periodic domain Ω.

This property can be seen as the most generic stability notion for systems of conservation laws. It is also desirable to retain this stability for the approximate solution. In [19], this was proved in the semi-discrete form. The space–time framework we adopted here provides us with a stronger fully discrete version:

Theorem 1. *Let us consider v^h as the approximate solution of (1) produced by scheme (10). Also, assume that the following holds for the symmetric positive semi-definite diffusion matrix $\bar{K}(v)$; there exists $\Lambda > 0$ such that for any $w \neq 0$*

$$0 \leq \langle w, \bar{K} w \rangle \leq \Lambda \langle w, w \rangle. \quad (30)$$

Also, let us set the stabilization parameters in the viscous discretization as the following

$$BR2: \quad \eta_e \geq N_l(\mathscr{T}_h), \qquad SIPG: \quad \mu_e \geq C_p(\mathscr{T}_h)\Lambda q^2, \quad (31)$$

where C_p is only dependent on the type of triangulation. Then, the following holds

$$\sum_{\kappa\in\mathscr{T}_h} \int_\kappa U(v^h(x,T)) \, \mathrm{d}x \leq \sum_{\kappa\in\mathscr{T}_h} \int_\kappa U(v^h(x,0)) \, \mathrm{d}x, \quad (32)$$

i.e., the method (10) is entropy stable in the fully discrete form.

The proof of this theorem follows the arguments already provided in [17, 19]. The relation (32) comes directly from the convective part as discussed in [17]. Showing the positiveness of the viscous discretization, $\mathscr{B}^v(v^h, v^h) \geq 0$, in a very similar way to [19], concludes the proof.

4 Numerical Results

In this section, we provide some numerical results, for both BR2 and SIPG formulations, to test and validate the methods and their convergence behavior. First in Sect. 4.1, we look into the advection–diffusion problem in the scalar settings. In Sect. 4.2, we apply our formulation to the compressible Navier–Stokes equations.

4.1 Scalar Advection-Diffusion

For scalar cases, the entropy variables coincide with the conservative ones, by choosing the entropy function as $U(u) = \frac{u^2}{2}$. We consider the following linear advection–nonlinear diffusion problem on a periodic domain $\Omega = [0, 1]$, by setting

$$f_c(u) = cu, \qquad f_v(u, \nabla u) = \epsilon(1 + u)\nabla u$$

in (1). Here, $\epsilon > 0$ is some constant, and $c = (1, 1)^t$ is the velocity field. Also, a source term is added to the right-hand side of (1) such that the exact solution of the problem is

$$u(x, t) = \frac{1}{2}\sin(2\pi x)\sin(2\pi t). \tag{33}$$

The numerical flux $\hat{f_c}$ is set to Lax–Friedrichs flux, which is entropy stable, combined with either BR2 or SIPG for discretizing the viscous flux. Moreover, we choose $\epsilon = 1$ and the corresponding stabilization parameters for SIPG and BR2 as $\mu_e = 10q^2$ and $\eta_e = 4$, respectively. The final time of computation is set to $T = 1$.

The convergence results for different DG polynomial degree are provided in Table 1. The results show that the scheme (approximately) achieves the optimal $q + 1$ order of accuracy in the asymptotic mesh refinement limit for both methods.

4.2 Navier–Stokes Equations

Let us consider the compressible Navier–Stokes equations in one dimension in the form of (1)

Table 1 Convergence table for advection–diffusion problem with BR2 and SIPG, $\epsilon = 1$[a]

BR2								
	$q=1$		$q=2$		$q=3$		$q=4$	
#Elements	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order
6	2.37e-01		1.13e-01		4.56e-02		6.79e-03	
24	5.25e-02	2.172	3.25e-02	1.806	1.22e-03	5.22	8.48e-04	3.001
96	3.67e-02	0.5168	3.21e-03	3.338	2.19e-04	2.479	2.20e-05	5.268
384	1.23e-02	1.577	3.60e-04	3.156	1.24e-05	4.136	7.52e-07	4.870
1536	3.64e-03	1.756	4.34e-05	3.051	7.27e-07	4.103	2.46e-08	4.936
6144	1.02e-03	1.838	5.39e-06	3.008	4.34e-08	4.067	7.86e-10	4.968
SIPG								
	$q=1$		$q=2$		$q=3$		$q=4$	
#Elements	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order
6	2.49e-01		1.31e-01		4.82e-02		8.27e-03	
24	5.37 e-02	2.213	4.20e-02	1.636	1.49e-03	5.014	1.09e-03	2.929
96	3.55e-02	0.5691	3.88e-03	3.436	2.72e-04	2.455	2.69e-05	5.331
384	1.20e-02	1.563	3.87e-04	3.328	1.53e-05	4.155	8.05e-07	5.067
1536	3.62e-03	1.733	4.43e-05	3.124	9.17e-07	4.058	2.51e-08	5.007
6144	1.02e-03	1.831	5.43e-06	3.029	5.68e-08	4.013	9.43e-10	4.73

[a] $\|e\|_{L_2}$ is calculated at the final time T

$$u = \begin{bmatrix} \rho \\ \rho V \\ E \end{bmatrix}, \qquad f_c = Vu + p \begin{bmatrix} 0 \\ 1 \\ V \end{bmatrix}, \qquad f_v = \begin{bmatrix} 0 \\ \tau \\ \tau V + q \end{bmatrix} \tag{34}$$

when V is the x-velocity, and ρ and E are the density and the internal energy, respectively. Also, p denotes the static pressure defined as $p = (\gamma - 1)(E - \frac{1}{2}\rho V^2)$ where γ is the *heat capacity ratio*. For air, we have $\gamma = 1.4$.

Moreover, τ is the viscous shear stress tensor and for Newtonian fluid in one dimension is $\tau = \frac{4}{3}\mu\partial_x V$. The heat flux q is defined by the Fourier's law as $q = \kappa\partial_x T$, where κ is the heat conductivity and is equal to $\kappa = \frac{\mu c_p}{Pr}$. Here, Pr is the Prandtl number, which for air at moderate conditions has a constant value of about $Pr = 0.72$. Also, the temperature T is defined by the ideal gas law.

According to [12], in order to arrive at the symmetric form (3) in case of nonzero heat flux, one should choose an affine function of the specific entropy $s = \log(\frac{p}{\rho^\gamma})$, e.g., $U = -\frac{\rho s}{\gamma - 1}$. For explicit form of v and $K(v)$, we refer to [12].

First, in order to see the accuracy of the method, we consider the following manufactured solution similar to [10], as

$$(\rho, \rho V, E) = (4 + \sin(kx - wt), \frac{4 + 0.2\sin(kx - wt)}{3}, \big(4 + \sin(kx - wt)\big)^2)$$

Table 2 Convergence table for the Navier–Stokes problems with BR2 and SIPG, $\mu = 0.1$[a]

BR2								
	$q = 1$		$q = 2$		$q = 3$		$q = 4$	
#Elements	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order
6	3.87e0		7.47e-01		2.91e-01		1.35e-01	
24	8.97e-01	2.11	9.51e-02	2.974	2.79e-02	3.38	3.85e-03	5.135
96	1.94e-01	2.21	1.32e-02	2.847	1.26e-03	4.471	1.13e-04	5.095
384	4.74e-02	2.033	1.66e-03	2.988	7.96e-05	3.987	4.56e-06	4.629
1536	1.20e-02	1.976	2.18e-04	2.929	5.08e-06	3.969	1.72e-07	4.731
SIPG								
	$q = 1$		$q = 2$		$q = 3$		$q = 4$	
#Elements	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order	$\|e\|_{L_2}$	order
6	5.04e0		9.70e-01		3.32e-01		1.52e-01	
24	8.84e-01	2.511	1.09e-01	3.154	3.16e-02	3.396	4.31e-03	5.136
96	1.87e-01	2.238	1.64e-02	2.733	1.37e-03	4.528	1.26e-04	5.088
384	5.03e-02	1.897	2.07e-03	2.991	8.17e-05	4.067	5.66e-06	4.486
1536	1.33e-02	1.919	2.73e-04	2.922	4.96e-06	4.042	2.37e-08	4.581

[a] $\|e\|_{L_2}$is calculated at the final time T

Table 3 Convergence table for the Functional J, with BR2 and SIPG scheme, $\mu = 0.1$

BR2						
	$q = 1$		$q = 2$		$q = 3$	
#Elements	ΔJ	order	ΔJ	order	ΔJ	order
6	7.31e-03		8.55e-03		−3.84e-04	
24	−3.99e-02	−2.451	1.86e-03	2.201	−2.77e-04	0.4695
96	−5.65e-03	2.824	−1.33e-04	3.805	5.79e-06	5.581
384	−1.42e-03	1.996	−3.78e-06	5.136	2.36e-08	7.938
1536	−3.45e-04	2.039	−2.22e-07	4.089	3.66e-10	6.013
SIPG						
	$q = 1$		$q = 2$		$q = 3$	
#Elements	ΔJ	order	ΔJ	order	ΔJ	order
6	−2.58e-02		1.17e-02		3.725e-04	
24	−4.85e-02	−0.9093	1.74e-03	2.751	−2.34e-04	0.667
96	−7.00e-03	2.792	−1.04e-04	4.06	6.17e-06	5.246
384	−1.71e-03	2.034	−3.87e-06	4.752	2.31e-08	8.058
1536	−4.01e-04	2.092	−2.63e-07	3.879	3.51e-10	6.044

where $k = 2\pi$, $w = 0.5$ on a periodic domain $\Omega = [0, 1]$, with the final time $T = 1$. Also, we set the viscosity $\mu = 0.1$.

The numerical convective flux is Lax–Friedrichs, and the method is tested with $\eta_e = 4$ for BR2 and with $\mu_e = 20q^2$ for SIPG.

The results presented in Tables 2 show almost the optimal order of convergence $q+1$ in the asymptotic mesh refinement limit. However, for $q=4$, the rate looks somehow sub-optimal.

We are also interested in measuring the error in terms of a given target functional $J(\cdot)$. We consider a weighted mean value of the density as

$$J(u) = \int_0^T \int_\Omega \rho(x,t)\sin(2\pi x)\,\mathrm{d}x\,\mathrm{d}t. \tag{35}$$

This functional is calculated for the same settings as for the results in Table 2, and the convergence rates of the error $\Delta J = J(u^h) - J(u)$ are reported in Table 3. The convergence rates are similar for both SIPG and BR2 and show (approximately) the convergence order of $2q$ for $q=1,2,3$ in the mesh refinement limit, which is the expected value (see [10]). Before reaching the asymptotic regime, one might observe irregularities in the convergence behavior, like the increase of the error at the first refinement for BR2 and SIPG with $q=1$.

Acknowledgements The research of the authors was supported by the Deutsche Forschungsgemeinschaft (German Research Association) through grant GSC 111.

References

1. D.N. Arnold, F. Brezzi, B. Cockburn, L.D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**, 1749–1779 (2002)
2. T.J. Barth, On discontinuous Galerkin approximations of Boltzmann moment systems with Levermore closure. Comput. Methods Appl. Mech. Eng. **195**, 3311–3330 (2006)
3. F. Bassi, S. Rebay, A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. J. Comput. Phys. **131**, 267–279 (1997)
4. F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, M. Savini, A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows, in *Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, Technologisch Instituut, Antwerpen, Belgium (1997), pp. 99–109
5. P. Chandrashekar, Discontinuous Galerkin method for Navier-Stokes equations using kinetic flux vector splitting. J. Comput. Phys. **233**, 527–551 (2013)
6. B. Cockburn, C. Dawson, Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multidimensions, in *The Mathematics of Finite Elements and Applications, X, MAFELAP, Uxbridge*, vol. 2000 (Elsevier, Oxford, 1999), pp. 225–238
7. B. Cockburn, C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems. SIAM J. Numer. Anal. **35**, 2440–2463 (1998)
8. L. Diosady, S. Murman, Higher-order methods for compressible turbulent flows using entropy variables, in *53rd AIAA Aerospace Sciences Meeting* (2015), p. 294
9. U.S. Fjordholm, S. Mishra, E. Tadmor, Arbitrarily high-order accurate entropy stable essentially nonoscillatory schemes for systems of conservation laws. SIAM J. Numer. Anal. **50**, 544–573 (2012)
10. R. Hartmann, P. Houston, An optimal order interior penalty discontinuous Galerkin discretization of the compressible Navier-Stokes equations. J. Comput. Phys. **227**, 9670–9685 (2008)
11. A. Hiltebrand, S. Mishra, Entropy stable shock capturing space-time discontinuous Galerkin schemes for systems of conservation laws. Numer. Math. **126**, 103–151 (2014)

12. T.J. Hughes, L. Franca, M. Mallet, A new finite element formulation for computational fluid dynamics: I. symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. Comput. Methods Appl. Mech. Eng. **54**, 223–234 (1986)
13. F. Ismail, P.L. Roe, Affordable, entropy-consistent Euler flux functions II: Entropy production at shocks. J. Comput. Phys. **228**, 5410–5436 (2009)
14. S. May, Spacetime discontinuous Galerkin methods for solving convection-diffusion systems, Technical Report 2015-05, Seminar for Applied Mathematics, ETH Zürich, Switzerland (2015)
15. S. May, Spacetime discontinuous Galerkin methods for convection-diffusion equations. Bull. Braz. Math. Soc. New Series **47**, 561–573 (2016)
16. E. Tadmor, Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. Acta Numerica **12**, 451–512 (2003)
17. M. Zakerzadeh, G. May, On the convergence of a shock capturing discontinuous Galerkin method for nonlinear hyperbolic systems of conservation laws. SIAM J. Numer. Anal. **54**, 874–898 (2016)
18. M. Zakerzadeh, G. May, *Analysis of mixed discontinuous Galerkin formulations for quasilinear elliptic problems*, arXiv preprint arXiv:1702.02733 (2017)
19. M. Zakerzadeh, G. May, Entropy stable discontinuous Galerkin scheme for the compressible Navier–Stokes equations, in *55th AIAA Aerospace Sciences Meeting* (2017), p. 0084

Invariant Manifolds for a Class of Degenerate Evolution Equations and Structure of Kinetic Shock Layers

Kevin Zumbrun

Abstract We describe recent results with A. Pogan developing dynamical systems tools for a class of degenerate evolution equations arising in kinetic theory, including the steady Boltzmann and BGK equations. These yield information on structure of large- and small-amplitude kinetic shocks, the first steps in a larger program toward time-evolutionary stability and asymptotic behavior.

Keywords Invariant manifolds · Steady Boltzmann equation · Kinetic shock profile

1 Introduction

In these notes, we describe recent results [39, 40] with Alin Pogan developing a set of dynamical systems tools suitable for the study of existence and structure of shock and boundary layer solutions arising in Boltzmann's equation and related kinetic models. These represent the first steps in a larger program to develop dynamical systems methods like those used in the study of finite-dimensional viscous and relaxation shocks in [13, 29, 46–49, 51, 52], suitable for treatment of one- and multi-dimensional stability of large-amplitude kinetic shock and boundary layers.

1.1 Equations and Assumptions

Our goal is the study of shock or boundary layer solutions

$$\mathbf{u}(x,t) = \check{\mathbf{u}}(x), \qquad \lim_{x\to\pm\infty} \check{\mathbf{u}}(x) = \mathbf{u}^{\pm}, \tag{1.1}$$

Research of K. Zumbrun was partially supported under NSF grant No. DMS-0300487.

K. Zumbrun (✉)
Indiana University, Bloomington 47405, IN, USA
e-mail: kzumbrun@indiana.edu

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems II*, Springer Proceedings in Mathematics & Statistics 237, https://doi.org/10.1007/978-3-319-91548-7_52

of kinetic-type relaxation systems

$$A^0\mathbf{u}_t + A\mathbf{u}_x = Q(\mathbf{u}) \tag{1.2}$$

on a Hilbert space $\mathbb{H}$, where A^0 and A are constant bounded linear operators, and Q, the *collision operator*, is a bounded bilinear map. This leads us to the study of the associated *steady equation*

$$A\mathbf{u}' = Q(\mathbf{u}). \tag{1.3}$$

Following [31, 39, 40], we make the following structural assumptions.

Hypothesis (H1) (i) The linear operator A is bounded, self-adjoint, and one-to-one on the Hilbert space $\mathbb{H}$, but *not boundedly invertible*. (ii) There exists $\mathbb{V}$ a proper, closed subspace of $\mathbb{H}$ with $\dim \mathbb{V}^\perp < \infty$ and $B : \mathbb{H} \times \mathbb{H} \to \mathbb{V}$ is a bilinear, symmetric, continuous map such that $Q(\mathbf{u}) = B(\mathbf{u}, \mathbf{u})$.

Hypothesis (H2) There exists an equilibrium $\overline{\mathbf{u}} \in \ker Q$ satisfying

(i) $Q'(\overline{\mathbf{u}})$ is self-adjoint and $\ker Q'(\overline{\mathbf{u}}) = \mathbb{V}^\perp$;
(ii) There exists $\delta > 0$ such that $Q'(\overline{\mathbf{u}})_{|\mathbb{V}} \leq -\delta I_{\mathbb{V}}$;

The class of system so described includes in particular our main example, of *Boltzmann's equation* with hard-sphere potential, written in appropriate coordinates [31]; see Sect. 2. As regards (1.3), the main novelty is that A by (H1)(i) has an *essential singularity*, i.e., essential spectrum at the origin, hence (1.3) is a *degenerate evolution equation* to which invariant manifold results of standard dynamical systems theory do not immediately apply. Our purpose here is precisely the construction of invariant manifolds for the class of degenerate Eq. (1.3) satisfying (H1)-(H2), and the application of these tools toward existence and structure of kinetic shock and boundary layers.

Remark 1.1. We do not assume as in [31] the "genuine coupling" or "Kawashima" condition that no eigenvector of A lies in the kernel of $Q'(\overline{\mathbf{u}})$. The assumption A one-to-one implies (trivially) the weaker condition, sufficient for our analysis, that no zero eigenvector of A lies in the kernel of $Q'(\overline{\mathbf{u}})$.

1.2 Chapman–Enskog Expansion and Canonical Form

Our starting point is the formal *Chapman–Enskog* expansion designed to approximate near-equilibrium flow [23]. Near $\overline{\mathbf{u}}$, (H1)-(H2) yields by the Implicit Function Theorem existence of a (Fréchet) C^∞ manifold of equilibria

$$\mathscr{E} = \ker Q, \qquad \dim \mathscr{E} = \dim \mathbb{V}^\perp =: r, \tag{1.4}$$

tangent to $\mathbb{V}^\perp$ at $\bar{\mathbf{u}}$, expressible in coordinates $\mathbf{w} := \mathbf{u} - \bar{\mathbf{u}}$ as a C^∞ graph

$$v_* : \mathbb{V}^\perp \to \mathbb{V}. \tag{1.5}$$

Denote $u = P_{\mathbb{V}^\perp}\mathbf{u}$, $v = P_{\mathbb{V}}\mathbf{u}$, where $P_{\mathbb{V}^\perp}$ and $P_{\mathbb{V}}$ are the orthogonal projections onto $\mathbb{V}^\perp$ and $\mathbb{V}$ associated with the decomposition $\mathbb{H} = \mathbb{V}^\perp \oplus \mathbb{V}$. The second-order Chapman–Enskog approximation, or "hydrodynamic limit," of (1.2) is then $h_*(u)_t + f_*(u)_x = D_* u_{xx}$, with associated steady equation

$$f_*(u)_x = D_* u_{xx}, \tag{CE}$$

where $h_*(u) := P_{\mathbb{V}^\perp} A^0(u^T, v_*(u)^T)^T$ and

$$f_*(u) := P_{\mathbb{V}^\perp} A(u^T, v_*(u)^T)^T, \quad D_* := A_{12} E^{-1} A_{12}^T, \tag{1.6}$$

with $A_{12} := P_{\mathbb{V}^\perp} A P_{\mathbb{V}}$ and $E := Q'(\bar{\mathbf{u}})_{|\mathbb{V}}$. See [23, 31, 40] for further details.

From (H1)(ii), $P_{\mathbb{V}^\perp}(A\mathbf{u})' = P_{\mathbb{V}^\perp} Q \equiv 0$. Integrating, we find that (1.3) admits a *conservation law*

$$P_{\mathbb{V}^\perp} A\mathbf{u} \equiv q = \text{constant}. \tag{1.7}$$

By the definition of f_*, v_*, equilibria $\mathbf{u}_\pm = (u^T, v_*(u)^T)_\pm^T$ satisfy the *Rankine–Hugoniot* condition

$$f_*(u_+) = f_*(u_-) = q \tag{RH}$$

associated with viscous shock profiles of the Chapman–Enskog system (CE), giving a rigorous connection at the inviscid level between shock or boundary layer profiles of the two systems (1.2) and (CE). A further connection, between the types of the equilibria $\bar{\mathbf{u}} = (\bar{u}^T, v_*(\bar{u})^T)^T$ and $\bar{u}$ with respect to their associated flows, is given by the following key observation proved in Sect. 2.

Lemma 1.2. *System* (1.3) *may, by an invertible change of coordinates, be put in canonical form*

$$\begin{aligned} w_c' &= J w_c + \tilde{Q}_c(w_c, w_h) \\ \Gamma_0 w_h' &= -w_h + \tilde{Q}_h(w_c, w_h), \end{aligned} \tag{1.8}$$

w_c and w_h parametrizing center and hyperbolic (i.e., stable/unstable) subspaces, $\dim w_c = m + r$, $m = \dim \ker f_*'(\bar{u})$, $r = \dim \mathbb{V}^\perp$, *where* $J = \begin{pmatrix} 0 & I_m & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ *is a nilpotent block-Jordan form,* Γ_0 *is a constant, bounded symmetric operator, and* $\tilde{Q}_j(w_c, w_h) = O(|w_c, w_h|^2)$. *In case* $m = 0$, J, $\tilde{Q}_c \equiv 0$.

One may compute that the perturbation equations for (CE) about $\bar{u}$ have the same canonical form (noting $f_*'(\bar{u}) = P_{\mathbb{V}^\perp} A P_{\mathbb{V}^\perp}$, D_* symmetric) with Γ_0 finite-dimensional, invertible [28, 32].

1.3 Dichotomies Versus Direct L^p Estimate

Lemma 1.2 effectively reduces the study of near-equilibrium flow of (1.3) to understanding the hyperbolic operator $(\Gamma_0\partial_x + \mathrm{Id})$, specifically, obtaining bounds on solutions of the degenerate inhomogeneous linear evolution system

$$(\Gamma_0\partial_x + \mathrm{Id})w_c = g, \tag{1.9}$$

where Γ_0 is bounded, symmetric, and one-to-one, but (by (H1)) *not boundedly invertible:* formally,

$$(\partial_x + \Gamma_0^{-1})w_c = \tilde{g}, \tag{1.10}$$

where Γ_0^{-1} is an unbounded self-adjoint operator and $\tilde{g} := \Gamma_0^{-1}g$. As Γ_0 is indefinite, (1.10) is *ill-posed* with respect to the Cauchy problem, featuring unbounded growth in both directions.

Ill-posed equations, and the derivation of associated resolvent bounds, have been treated in a variety of contexts via *generalized exponential dichotomies*: for example, modulated waves on cylindrical domains [33, 36, 37], Morse theory [1, 2, 34], PDE Hamiltonian systems [35], and the functional-differential equations of mixed type [27]. It is not difficult to see, either by spectral decomposition of Γ_0, or by Galerkin approximation, that $(\partial_x + \Gamma_0^{-1})$ generates a *stable bi-semigroup* [5, 19], the infinite-dimensional analog of an exponential dichotomy, that is, there exist bounded projections on whose range the homogeneous flow is exponentially decaying in forward/backward direction, in this case with rate $|\Gamma_0|_{\mathbb{H}}^{-1}$, where $|\cdot|_{\mathbb{H}}$ denotes operator norm; see [39] for details.

This, however, yields only $\|u\| \le C\|\tilde{g}\| = \|\Gamma_0^{-1}g\|$, the intervention of the unbounded operator Γ_0^{-1} making these bounds useless for our analysis. Thus, the present problem differs from the above-mentioned ones in that *exponential dichotomies are inadequate to bound the resolvent* $(\Gamma_0\partial_x + \mathrm{Id})^{-1}$. Indeed, we have the following striking result obtained by direct estimate in Sect. 3, showing that our situation is one of *maximal regularity*. In this sense, our analysis is related in flavor to construction of center manifolds for quasilinear systems; see [16, 30] and references therein.

Lemma 1.3. *Assuming (H1)-(H2),* $|(\Gamma_0\partial_x + \mathrm{Id})^{-1}|_{L^p(\mathbb{R})} < \infty$ *for* $1 < p < \infty$*, but* not *for* $p = 1, \infty$.

An important consequence is that usual weighted L^∞ constructions of invariant manifolds are unavailable. We work instead in *weighted* H^1 *spaces*, with accompanying new technical issues.

1.4 Results

We are now ready to state our main results. Assuming (H1)-(H2), from (1.8) and symmetry of Γ_0 we readily obtain a decomposition $\mathbb{H} = \mathbb{H}_s \oplus \mathbb{H}_c \oplus \mathbb{H}_u$ of $\mathbb{H}$ into stable, center, and unstable subpaces invariant under the homogeneous linearized flow of (1.3) about the equilibrium $\bar{\mathbf{u}}$. Let $H^1_\eta(\mathbb{R}, \mathbb{H})$ denote the space of functions bounded in the exponentially weighted H^1 norm

$$\|f\|_{H^1_\eta(\mathbb{R},\mathbb{H})} := \|e^{\eta\langle\cdot\rangle} f(\cdot)\|_{L^2(\mathbb{R},\mathbb{H})} + \|e^{\eta\langle\cdot\rangle} f'(\cdot)\|_{L^2(\mathbb{R},\mathbb{H})}, \tag{1.11}$$

where $\langle x\rangle := (1+|x|^2)^{1/2}$ and $\eta \in \mathbb{R}$ may be positive or negative according to our needs. Following [19], we define solutions of (1.3) using Lemma 1.3 as H^1_{loc} solutions of the fixed-point equation $w_h = (\Gamma_0\partial_x + \mathrm{Id})^{-1} g_c(w)$ and the finite-dimensional ODE $(\partial_x - J)w_c = g_c(w)$ in w_c; see [39, 40].

1.4.1 H^1 Stable Manifold and Exponential Decay of Large-Amplitude Shock and Boundary Layers

Our first observation is that for singular Γ_0 the H^1 stable subspace of (1.8), defined as the trace at $x = 0$ of solutions w_h bounded in $H^1(\mathbb{R}^+, \mathbb{H})$, *is a dense proper subspace of* $\mathbb{H}_s$, related to the domain of the generator Γ_0^{-1} of the bi-semigroup associated with homogeneous linearized flow.

Lemma 1.4. *Assuming (H1)-(H2), the H^1 stable subspace of the linearized equations of (1.3) about $\bar{\mathbf{u}}$ (equivalently, the linearization of (1.8) about 0) is* $\mathrm{dom}(|\Gamma_0|^{-1/2}) \cap \mathbb{H}_s \subset \mathbb{H}_s$.

Proof. The H^1 stable subspace consists of $f \in \mathbb{H}_s$ such that $\int_0^\infty \langle \partial_x e^{\Gamma_0^{-1}x} f, \partial_x e^{\Gamma_0^{-1}x} f\rangle\, dx < \infty$, or, equivalently, $-(1/2)\int_0^\infty \partial_x \langle e^{\Gamma_0^{-1}x}|\Gamma_0|^{-1/2} f, e^{\Gamma_0^{-1}x}|\Gamma_0|^{-1/2} f\rangle dx < \infty$. Integrating, and observing that the boundary term at infinity vanishes, gives condition $\langle |\Gamma_0|^{-1/2} f, |\Gamma_0|^{-1/2} f\rangle < \infty$. Alternatively, this may be deduced by spectral decomposition of Γ_0 and direct computation [39]. □

We have accordingly the following modification of the usual stable manifold theorem.

Theorem 1.5. *Assuming (H1)-(H2), for any $0 < \alpha < \tilde{\nu} < \nu := |\Gamma_0|_{\mathbb{H}}^{-1}$, there exists a local stable manifold $\mathscr{M}_s$ near $\bar{\mathbf{u}}$, expressible in coordinates $w = \mathbf{u} - \bar{\mathbf{u}}$ as a C^1 embedding tangent to $\mathbb{H}_s$ of $\mathrm{dom}(\Gamma_0^{-1/2}) \cap \mathbb{H}_s$ with (graph) norm induced by $\Gamma_0^{-1/2}$ into $\mathbb{H}$, locally invariant under the forward flow of (1.3), containing the orbits of all solutions w with $H^1_\alpha(\mathbb{R}_+, \mathbb{H})$ norm sufficiently small, with solutions w initiating in $\mathscr{M}_s$ at $x = 0$ lying in $H^1_{\tilde{\nu}}(\mathbb{R}_+, \mathbb{H})$. In case* $\det f'_*(u_+) \neq 0$, *α may be taken to be zero.*

We obtain as a consequence exponential decay of noncharacteristic shock or boundary layers.

Corollary 1.6. *Assuming (H1)-(H2), let* $\bar{\mathbf{u}}$ *be a noncharacteristic equilibrium in the sense of* (CE)*,* $\det f_*'(\bar u) \neq 0$*, and* $\tilde\nu < \nu = 1/|\Gamma_0|_{\mathbb{H}}$*. Then, for any solution* $\check{\mathbf{u}}$ *of* (1.3) *converging to* $\bar{\mathbf{u}}$ *as* $x \to +\infty$ *in the sense that* $\check{\mathbf{u}} - \bar{\mathbf{u}}$ *is eventually bounded in* $H^1([x,\infty),\mathbb{H})$*, we have* exponential decay*:*

$$|\check u - \bar{\mathbf{u}}|_{\mathbb{H}}(x) \lesssim e^{-\tilde\nu x} \quad \text{as } x \to +\infty. \tag{1.12}$$

1.4.2 Center Manifold and Structure of Small-Amplitude Shock Layers

We have, similarly, the following modification of the usual center manifold theorem (cf. [7, 16, 44, 45]).

Theorem 1.7. *Let* $\bar{\mathbf{u}}$ *be an equilibrium satisfying (H1)-(H2). Then, for any integer* $k \geq 2$ *there exists local to* $\bar{\mathbf{u}}$ *a* C^k *center manifold* $\mathscr{M}_{\mathrm{c}}$*, tangent at* $\bar{\mathbf{u}}$ *to* $\mathbb{H}_{\mathrm{c}}$*, expressible in coordinates* $\mathbf{w} := \mathbf{u} - \bar{\mathbf{u}}$ *as a* C^k *graph* $\mathscr{J}_{\mathrm{c}} : \mathbb{H}_{\mathrm{c}} \to \mathbb{H}_{\mathrm{s}} \oplus \mathbb{H}_{\mathrm{u}}$*, that is locally invariant under the flow of* (1.3) *and contains all solutions that remain sufficiently close to* $\bar{\mathbf{u}}$ *in forward and backward* x*. Moreover,* $\mathscr{M}_{\mathrm{c}}$ *has the* H^1 exponential approximation property*: For any* $0 < \tilde\nu < \nu = 1/|\Gamma_0|_{\mathbb{H}}$*, a solution* $\mathbf{u}$ *of* (1.3) *with* $\|\mathbf{u} - \bar{\mathbf{u}}\|_{H^1_{-\alpha}\cap L^\infty([M,\infty),\mathbb{H})}$ *and* $\alpha > 0$ *sufficiently small approaches a solution* $\mathbf{z}$ *with orbit lying in* $\mathscr{M}_{\mathrm{c}}$ *as* $x \to +\infty$ *at exponential rate* $\|\mathbf{u} - \mathbf{z}\|_{\mathbb{H}} \lesssim e^{-\tilde\nu x}$*, with also* $\|\mathbf{u} - \mathbf{z}\|_{H^1_{\tilde\nu}([M,\infty),\mathbb{H})} < \infty$.

Here, the only difference from the standard center manifold theorem [7] is the weakened, H^1, version of the exponential approximation property. For applications involving normal form reduction, they are essentially equivalent; in particular, the formal Taylor expansion for center graph $w_h = \Xi(w_c)$ may be computed to arbitrary order in coordinates (1.8) by successively matching terms of increasing order in the defining relation $\Gamma_0\Xi(w_c)' = -\Xi(w_h) + \tilde Q_h$, or equivalently $\Xi(w_c) = -\Gamma_0\Xi'(w_c)(Jw_c + \tilde Q_c) + \tilde Q_h$, exactly as in the usual (nonsingular A, Γ_0) case [10, 16].

Remark 1.8. In the noncharacteristic case, the center manifold, by dimensional count and the fact that it must contain all local equilibria, is uniquely determined as the manifold of equilibria $\mathscr{E}$. In this case, the exponential approximation property improves slightly the result of Corollary 1.6, yielding that solutions $\check{\mathbf{u}}$ of (1.3) lying sufficiently close to $\bar{\mathbf{u}}$ in $L^\infty(\mathbb{R}^+,\mathbb{H})$ and sufficiently slowly exponentially growing in H^1 converge to an equilibrium at exponential rate $e^{-\tilde\nu x}$, $0 < \tilde\nu < 1/|\Gamma_0|_{\mathbb{H}}$.

Denote the characteristics of Chapman–Enskog system (CE), or eigenvalues of $f_*'(u)$, by

$$\lambda_1(u) \leq \cdots \leq \lambda_r(u).$$

The *noncharacteristic case* $f_*'(\bar u) \neq 0$ is the case that no characteristic velocity $\lambda_j(\bar u)$ vanishes, in which case, by the Inverse Function Theorem, the Rankine–Hugoniot equations (RH) admit a single nearby solution for each value of q, hence no local

shock connections occur. To study small-amplitude shock profiles, we focus therefore on the *characteristic case* $f_*'(\bar u) = 0$, specifically on the generic case that $\lambda_j(\bar u) = 0$ for a single characteristic velocity λ_p, with associated unit eigenvector $\bar{\mathbf r}$, that is *genuinely nonlinear* in the sense of Lax [21, 41]:

$$\Lambda := \bar{\mathbf r} \cdot f_*''(\bar u)(\bar{\mathbf r}, \bar{\mathbf r}) \neq 0. \tag{GNL}$$

In this case, it is well known [21, 28, 41] that there exists a family of small-amplitude shock profiles $\breve{\bar u}$ of (CE) connecting endstates $\bar u_\pm \to \bar u$, with $(\bar u_+ - \bar u_-)$ lying in approximate direction $\bar{\mathbf r}$, with $\lambda := \lambda_p(\breve{\bar u})$ satisfying an approximate Burgers equation

$$\delta\lambda' = -\varepsilon^2 + \lambda^2/2 + O(|\varepsilon, \lambda|^3), \tag{1.13}$$

Λ as in (GNL), $\varepsilon > 0$ parametrizing amplitude, provided there holds the *stable viscosity criterion* $\delta := \bar{\mathbf r} \cdot D_*\bar{\mathbf r} > 0$, as may be readily seen to hold for D_* using (1.6) and (H1) (cf. Remark 1.1).

Our final result gives a corresponding characterization of small-amplitude kinetic shocks of (1.3) bifurcating from a simple genuinely nonlinear eigenvalue of $f_*'(\bar u)$. The complementary case of bifurcation from a multiple, linearly degenerate eigenvalue of $f_*'(\bar u)$ [21, 41] is treated also in [40, Theorem 1.5] (not stated here); in that case, no nontrivial shock or boundary layer connections exist.

Corollary 1.9. *Let $\bar{\mathbf u}$ be an equilibrium satisfying (H1)-(H2) in the characteristic case* (GNL)*, $\lambda_p(\bar u) = 0$ a simple eigenvalue, and k an integer ≥ 2. Then, local to $\bar{\mathbf u}$, $\bar u$, each pair of points $u_\pm$ satisfying the Rankine–Hugoniot condition* (RH) *has a corresponding viscous shock solution u_{CE} of* (CE) *and relaxation shock solution $\mathbf u_{REL} = (u_{REL}, v_{REL})$ of* (1.3)*, satisfying for all $j \leq k-2$:*

$$\begin{aligned} \big|\partial_x^j(u_{REL} - u_{CE})\big| &\leq C\varepsilon^{j+2}e^{-\mu\varepsilon|x|}, \\ \big|\partial_x^j\big(v_{REL} - v_*(u_{CE})\big)\big| &\leq C\varepsilon^{j+2}e^{-\mu\varepsilon|x|}, \\ |\partial_x^j(u_{REL} - u_\pm)| &\leq C\varepsilon^{j+1}e^{-\mu\varepsilon|x|}, \quad x \gtrless 0, \end{aligned} \tag{1.14}$$

$\mu > 0, C > 0, \varepsilon := |u_+ - u_-|$, unique up to translation, with $\lambda_p(u_{REL})$ and $\lambda_p(u_{CE})$ both satisfying approximate Burgers equations (1.13)*: in particular, both monotone decreasing in x.*

1.5 Discussion and Open Problems

Corollary 1.9 recovers under slightly weakened assumptions, the result of [31, Proposition 5.4], which, applied to Boltzmann's equation, in turn recovers and sharpens the fundamental result [8] of existence of small-amplitude Boltzmann shocks with standard, square-root Maxwellian-weighted L^2 norm in velocity [15]. With further

effort, one may show [40, Proposition 1.8] (not stated here) that the center manifold of Theorem 1.7, hence also the small-amplitude shock profiles obtained in Corollary 1.9, are contained in a stronger space of near-Maxwellian-weighted L^2 norm in velocity, recovering the strongest current existence result for Boltzmann shocks [31, Theorem 1.1], plus the additional dynamical information of (1.13) and monotonicity of $\lambda_p(u_{REL}(x))$- neither of the latter of which appears to be available by the Sobolev-based fixed-point iteration arguments of [8, 31].

To our knowledge, Theorems 1.5 and 1.7 are the first results on existence of invariant manifolds for any system of form 1.2, (H1)-(H2) in either Hilbert or Banach space setting, in particular for the steady Boltzmann equation with hard-sphere potential. Liu and Yu [25] have studied existence of invariant manifolds for Boltzmann's equation in a weighted L^∞ (in both velocity and x) Banach space setting, using rather different methods of time-regularization and detailed pointwise bounds, pointing out that monotonicity of $\lambda_p(\bar{u})$ follows from center manifold reduction and describing physical applications of center manifold theory to condensation and subsonic/supersonic transition in Milne's problem. However, their claimed linearized bounds, based on exponential dichotomies, hence also their arguments for existence of invariant manifolds, were incorrect [50]; see Remark 3.3. Our results among other things repair this gap, validating their larger program/physical conclusions.

A longer-term program is to develop further dynamical systems tools for kinetic systems (1.2) with structure (H1)-(H2), sufficient to treat *time-evolutionary stability* of shock and boundary layers by the methods used for viscous/relaxation shocks in [13, 29, 46–49, 51, 52]. Besides unification/simplification, this approach has the advantage of applying in principle to multi-dimensional and/or large-amplitude waves, each of these long-standing open problems in the area.

These techniques have the further advantages of separating the issues of existence, spectral stability, and linearized/nonlinear stability, with the first two often treated by a combination of analytical and numerical methods, up to and including (see, e.g., [3, 4]) interval arithmetic-based rigorous numerical proof. The development of numerical and or analytical methods for the treatment of existence of large-amplitude kinetic shocks we regard as a further, very interesting open problem.

Indeed, the *structure problem* discussed by Truesdell, Ruggeri, Boillat, and others, of existence and description of large-amplitude Boltzmann shocks, is perhaps *the* fundamental open problems in the theory, and one of the main motivations for their study. As discussed, e.g., in [6], Navier–Stokes theory well describes the behavior of shocks of Mach number $M \lessapprox 2$, but inaccurately predicts shock width/structure at large Mach numbers; by contrast, Boltzmann's equation (numerically and via various formal approximations) appears to match experiment in the large-M regime.

2 Reductions and Main Example

We begin by carrying out various reductions, first from Boltzmann's equation to the abstract form (1.3), (H1)-(H2), then the abstract equation to the canonical form (1.8).

2.1 Boltzmann's Equation

(Following [31]) Our main interest is Boltzmann's equation with hard-sphere potential (or Grad hard cutoff potential as in [8]):

$$f_t + \xi_1 \partial_x f = \mathscr{Q}(f, f), \tag{2.1}$$

where $f(x, t, \xi) \in \mathbb{R}$ is the distribution of velocities $\xi \in \mathbb{R}^3$ at $x, t \in \mathbb{R}$, and

$$\mathscr{Q}(g, h) := \int \big(g(\xi')h(\xi'_*) - g(\xi)h(\xi_*)\big)C(\Omega, \xi - \xi_*)d\Omega d\xi_* \tag{2.2}$$

is the collision operator, with collision kernel $C(\Omega, \xi) = \big|\Omega \cdot \xi\big|$ for hard-sphere case.

The space of *collision invariants* $\langle\psi\rangle$, $\int_{\mathbb{R}^3} \psi(\xi)\mathscr{Q}(g, g)(\xi)d\xi \equiv 0$, of (2.1) is spanned by

$$Rf := \int \Psi(\xi) f(\xi) d\xi \in \mathbb{R}^5, \qquad \Psi(\xi) = (1, \xi_1, \xi_2, \xi_3, \frac{1}{2}|\xi|^2)^T. \tag{2.3}$$

(Here, we are assuming that distributions $f(x, t, \cdot)$ are confined to a space $\mathbb{H}$ to be specified later such that the integral converges.) The associated macroscopic (fluid-dynamical) variables are

$$\mathrm{u} := Rf =: (\rho, \rho v_1, \rho v_2, \rho v_3, \rho E)^T, \tag{2.4}$$

where ρ denotes density, $v = (v_1, v_2, v_3)$ velocity, $E = e + \frac{1}{2}|v|^2$ total energy density, and e internal energy density. The set of *equilibria* ($\ker \mathscr{Q}$) consists of the *Maxwellian distributions*:

$$M_{\mathrm{u}}(\xi) = \frac{\rho}{\sqrt{(4\pi e/3)^3}} e^{-\frac{|\xi - v|^2}{4e/3}}. \tag{2.5}$$

2.1.1 Symmetry, Boundedness, and Spectral Gap

Boltzmann's H-theorem [11, 14, 15] (equivalent to existence of a thermodynamical entropy in the sense of [12]) asserts the variational principle

$$\int \log f \mathscr{Q}(f, f) d\xi \le 0,$$

with equality on the set of Maxwellians $\underline{M}$. Taylor expanding about a local maximum $\underline{M}$, we obtain symmetry and nonnegativity of the Hessian $\int \underline{M}^{-1}(\partial\mathscr{Q}|_{\underline{M}}h)hd\xi \le 0$,

giving symmetry and nonnegativity of $\partial\mathscr{Q}|_{\underline{M}}$ on the space $\mathbb{H}$ defined by the square-root Maxwellian-weighted norm

$$\|f\|_{\mathbb{H}} := \|f\underline{M}^{-1/2}\|_{L^2(\mathbb{R}^3)}. \tag{2.6}$$

Making the coordinate change

$$\mathbf{u} = \langle\xi\rangle^{1/2} f, \quad Q(\mathbf{u}) := \langle\xi^{-1/2}\rangle^{-1}\mathscr{Q}(\langle\xi\rangle^{-1/2}\mathbf{u}), \qquad \langle\xi\rangle := \sqrt{1+|\xi|^2}, \tag{2.7}$$

and defining multiplication operators $A^0 = \langle\xi\rangle^{-1}$ and $A = \xi_1/\langle\xi\rangle$, we find that (2.1) may be put in form (1.2), for $\mathbf{u} \in \mathbb{H}$, with A^0, A evidently symmetric and bounded, $A^0 > 0$, and $Q'(\overline{\mathbf{u}})$ symmetric nonpositive at any equilibrium $\overline{\mathbf{u}} = \langle\xi\rangle^{1/2}\underline{M}$. By [31, Corollary 2.4], Q is bounded as a bilinear map on $\mathbb{H}$. Moreover, by [31, Proposition 3.5], $Q'(\overline{\mathbf{u}})$ is negative definite with respect to $\mathbb{H}$ on its range, this last being a straightforward consequence of Carleman's theorem [9] that $\partial\mathscr{Q}|_{\underline{M}}$ acting on $\mathbb{H}$ may be decomposed as the sum of a multiplication operator $\nu(\xi) \sim \langle-\xi\rangle$ and a compact operator K, whence $Q'(\overline{\mathbf{u}})$ is the sum of a multiplication operator $\tilde{\nu}(\xi) \sim -1$ and the compact operator $\tilde{K} = \langle\xi\rangle^{-1/2}K\langle\xi\rangle^{-1/2}$, Weyl's Theorem thereby implying existence of a spectral gap.

Collecting information, we find that we have reduced to a system of form (1.2) satisfying (H1)-(H2), with $\mathbb{V} := \langle\xi\rangle^{1/2}(\mathrm{Range}\,R)^{\perp}$, R as in (2.3), $\dim\mathbb{V}^{\perp} = 5$, and $\overline{\mathbf{u}} = \langle\xi\rangle^{1/2}\underline{M}$ for any Maxwellian $\underline{M}$. Note that A has no kernel on $\mathbb{H}$, but essential spectra $\xi_1/\langle\xi\rangle \to 0$ as $\xi_1 \to 0$: an *essential singularity*. A consequence is that *small velocities* $\xi_1 \to 0$ *constitute the main difficulties in our analysis,* large-velocities issues having been subsumed in the reduction [31] to form (1.2).

2.1.2 Hydrodynamic Limit

The formal Chapman–Enskog expansion (CE), or hydrodynamic limit, being independent of coordinate representation, is the same in our variables $\mathbf{u}$, Q as in the standard Boltzmann variables f, $\mathscr{Q}$. As computed, e.g., in [11, 25], this appears in fluid variables (2.4) as the *compressible Navier–Stokes* equations with temperature-dependent viscosity and heat conduction:

$$\begin{aligned}
\rho_t + (\rho v_1)_x &= 0,\\
(\rho v_1)_t + (\rho v_1^2 + p)_x &= ((4/3)\mu v_{1,x})_x,\\
(\rho_2)_t + (\rho v_1 v_2)_x &= (\mu v_{2,x})_x,\\
(\rho_3)_t + (\rho v_1 v_3)_x &= (\mu v_{3,x})_x,\\
(\rho E)_t + (\rho v_1 \rho E + v_1 p)_x &= (\kappa T_x + (4/3)\mu v_1 v_{1,x})_x,
\end{aligned} \tag{cNS}$$

where T denotes temperature, with monatomic equation of state $p = \Gamma\rho e, T = c_v^{-1}e$, with

$$\Gamma = 2/3, \quad c_v = 3/4, \quad \mu = \mu(T) = (5/16)\sqrt{T/\pi}, \quad \kappa = \kappa(T) = (75/16)\sqrt{T/\pi}. \tag{2.8}$$

As computed in, e.g., [41], the hyperbolic (i.e., left-hand side) part of (cNS) has characteristics

$$\lambda_1 = v_1 - c, \quad \lambda_2 = \lambda_3 = \lambda_4 = v_1, \quad \lambda_5 = v_1 + c, \tag{2.9}$$

where $c := \sqrt{\Gamma(1+\Gamma)e} > 0$ denotes sound speed, with "acoustic modes" $v_1 \pm c$ simple and satisfying (GNL), and "entropic/vorticity modes" v_1 multiplicity three and linearly degenerate in the sense of Lax [21, 41] (not addressed here; see [40] for discussion of the linearly degenerate case).

2.2 Macro–Micro Decomposition

Next, starting with form (1.3), (H1)-(H2), coordinatize as in Sect. 1.2 $\mathbf{u}$ as (u, v), $u = P_{\mathbb{V}^\perp}\mathbf{u}$, $v = P_{\mathbb{V}}\mathbf{u}$, where $P_{\mathbb{V}^\perp}$ and $P_{\mathbb{V}}$ are the orthogonal projections associated with orthogonal decomposition $\mathbb{H} = \mathbb{V}^\perp \oplus \mathbb{V}$, to obtain the block decomposition

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}' = \begin{pmatrix} 0 & 0 \\ 0 & E \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ f \end{pmatrix}, \tag{2.10}$$

into "macro" and "micro" variables u and v similarly as in [25, 31], with forcing term $f = B(\mathbf{u}, \mathbf{u})$, where B is a bounded bilinear map and $E < 0$ is symmetric negative definite on $\mathbb{H}$. The following further reduction greatly simplifies computations later on; hereafter we take $E = -\mathrm{Id}$.

Observation 2.1. By the change of variables $v \to (-E)^{1/2}$ combined with left-multiplication of the v-equation by $(-E)^{-1}$, we may take without loss of generality $E = \mathrm{Id}$.

2.3 Reduction to Canonical Form

Since A and $P_{|\mathbb{V}^\perp}$ are self-adjoint on $\mathbb{H}$, $A_{11} = P_{\mathbb{V}^\perp}A_{|\mathbb{V}^\perp}$ is self-adjoint on $\mathbb{V}^\perp$, hence $\mathbb{V}^\perp = \ker A_{11} \oplus \mathrm{im}A_{11}$. Denote by $P_{\ker A_{11}}$ and $P_{\mathrm{im}A_{11}}$ the associated orthogonal projections onto $\ker A_{11}$ and $\mathrm{im}A_{11}$, and $\widetilde{A}_{12} : \mathbb{V} \to \mathrm{im}A_{11}$ and $T_{12} : \mathbb{V} \to \ker A_{11}$ the operators defined by $\widetilde{A}_{12} = P_{\mathrm{im}A_{11}}A_{12}$ and $T_{12} = P_{\ker A_{11}}A_{12}$. From the assumption that A is one-to-one, we readily obtain the following; see [40, Lemma 2.1] for details.

Lemma 2.2. *Assuming (H1)-(H2), (i)* $\ker T_{12}^* = \{0\}$, $\mathrm{im} T_{12} = \ker A_{11}$, $\ker T_{12} \neq \{0\}$, *and (ii) The linear operator* $\widetilde{A}_{11} = (A_{11})_{|\mathrm{im} A_{11}}$ *is self-adjoint and invertible on* $\mathrm{im} A_{11}$.

Introduce now orthogonal subspaces $\mathbb{V}_1 = \mathrm{im} T_{12}^*$ and $\widetilde{\mathbb{V}} = \ker T_{12}$ decomposing $\mathbb{V}$, with associated projectors $P_{\mathbb{V}_1}$ and $P_{\widetilde{\mathbb{V}}}$. Denoting

$$u_1 = P_{\ker A_{11}} u, \quad \widetilde{u} = P_{\mathrm{im} A_{11}} u, \quad v_1 = P_{\mathbb{V}_1} v, \quad \text{and } \widetilde{v} = P_{\widetilde{\mathbb{V}}} v, \tag{2.11}$$

and applying $P_{\ker A_{11}}$ and $P_{\mathrm{im} A_{11}}$ to the first equation of (2.10) we obtain

$$T_{12} v' = 0, \quad \widetilde{A}_{11} \widetilde{u}' + \widetilde{A}_{12} \widetilde{v}' = 0. \tag{2.12}$$

Moreover, by $(A_{21})_{|\ker A_{11}} = T_{12}^*$, $(A_{21})_{|\mathrm{im} A_{11}} = \widetilde{A}_{12}^*$ the second equation of (2.10) is equivalent to

$$T_{12}^* u_1' + \widetilde{A}_{12}^* \widetilde{u}' + A_{22} v' = E v + f. \tag{2.13}$$

Since $v_1 \in \mathbb{V}_1 = \mathrm{im} T_{12}^*$, from (2.12) we conclude $v_1' = 0$. In addition, since $\widetilde{A}_{11}$ is invertible on $\mathrm{im} A_{11}$ by Lemma 2.2(ii), we have $\widetilde{u}' = -\widetilde{A}_{11}^{-1} \widetilde{A}_{12} \widetilde{v}'$. Summarizing, (2.12) is equivalent to

$$v_1' = 0, \quad (\widetilde{u} + \widetilde{A}_{11}^{-1} \widetilde{A}_{12} \widetilde{v})' = 0. \tag{2.14}$$

Next, taking without loss of generality $E = \mathrm{Id}$, we obtain from (2.13) evidently

$$T_{12}^* u_1' + P_{\mathbb{V}_1} (A_{22} - \widetilde{A}_{12}^* \widetilde{A}_{11}^{-1} \widetilde{A}_{12}) \widetilde{v}' = -v_1 + P_{\mathbb{V}_1} f \tag{2.15}$$

and

$$P_{\widetilde{\mathbb{V}}} (A_{22} - \widetilde{A}_{12}^* \widetilde{A}_{11}^{-1} \widetilde{A}_{12}) \widetilde{v}' = -\widetilde{v} + P_{\widetilde{\mathbb{V}}} E v_1 + P_{\widetilde{\mathbb{V}}_1} f. \tag{2.16}$$

From Lemma 2.2(i), $(T_{12}^*)^{-1}$ is well-defined and bounded, hence we obtain from (2.15)

$$(u_1 - \Gamma_1 \widetilde{v})' = -(T_{12}^*)^{-1} v_1 + (T_{12}^*)^{-1} P_{\mathbb{V}_1} f, \quad \Gamma_0 \widetilde{v}' = \widetilde{v} + P_{\widetilde{\mathbb{V}}} f, \tag{2.17}$$

where $\Gamma_1 = (T_{12}^*)^{-1} (\widetilde{A}_{12}^* \widetilde{A}_{11}^{-1} \widetilde{A}_{12} - A_{22}) \in \mathscr{B}(\mathbb{V}, \ker A_{11})$ and

$$\Gamma_0 = P_{\widetilde{\mathbb{V}}} (A_{22} - \widetilde{A}_{12}^* \widetilde{A}_{11}^{-1} \widetilde{A}_{12})_{|\widetilde{\mathbb{V}}} \in \mathscr{B}(\widetilde{\mathbb{V}}) \text{ issymmetric.} \tag{2.18}$$

Summarizing, we have that (2.10) is equivalent to the system

$$(u_1 - \Gamma_1 \widetilde{v})' = (T_{12}^*)^{-1} v_1 + (T_{12}^*)^{-1} P_{\mathbb{V}_1} f, \quad (\widetilde{u} + \widetilde{A}_{11}^{-1} \widetilde{A}_{12} \widetilde{v})' = 0, \quad v_1' = 0, \quad \Gamma_0 \widetilde{v}' = \widetilde{v} + P_{\widetilde{\mathbb{V}}} f. \tag{2.19}$$

By the invertible change of coordinates

$$w_c = \Big((u_1 - \Gamma_1 \tilde{v})^T, (-(T_{12}^*)^{-1} v_1)^T, (\tilde{u} + \tilde{A}_{11}^{-1} \tilde{A}_{12} \tilde{v})^T\Big)^T, \quad w_h = \tilde{v}, \tag{2.20}$$

we reduce (2.10) finally to the canonical form of Lemma 1.2:

$$w_c' = J w_c + g_c, \quad \Gamma_0 w_h' = w_h + g_h, \tag{2.21}$$

where $J = \begin{pmatrix} 0 & I_m & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $g_c = \tilde{Q}_c(w, w)$ and $g_h = \tilde{Q}_h(w, w)$ are bounded bilinear maps.

Observation 2.3. We record for later that the tangent subspace $(u, v) = (\zeta, 0)$ to equilibrium manifold $\mathscr{E} = \{(u, v) \in \mathbb{V}^\perp \oplus \mathbb{V} : Q(u, v) = 0\}$ is given in coordinates (2.20) by $w_c = (\zeta_1, 0, \tilde{\zeta})$, $w_h = 0$, as can also be seen by computing the subspace of equilibria of (1.8) with $g = (g_c, g_h) = 0$.

3 Linear Resolvent Estimates

The starting point for construction of invariant manifolds is the study of the solution operator for the decoupled linear inhomogeneous Eq. (2.21) with arbitrary forcing terms g_c, g_h. The "center," w_c equation is of standard finite-dimensional type, so may be treated by usual methods. Evidently, then, the key issue is treatment of the degenerate "hyperbolic," w_h equation.

3.1 Symmetric Degenerate Evolution Equations

Consider a degenerate inhomogeneous evolution equation $(\Gamma_0 \partial_x + \mathrm{Id}) w_c = g$, with Γ_0 (recalling (2.18) and (H1)-(H2)) symmetric and one-to-one but not boundedly invertible, with the goal to obtain bounds on the resolvent operator

$$\mathscr{R} := (\Gamma_0 \partial_x - \mathrm{Id})^{-1}. \tag{3.1}$$

As discussed in the introduction, the inhomogeneous flow $u' + \Gamma_0^{-1} u = 0$ possesses generalized exponential dichotomies, but the resulting bounds on $(\partial_x - \Gamma_0^{-1})^{-1}$ are insufficient to bound the inhomogeneous solution operator $(\Gamma_0 \partial_x - \mathrm{Id})^{-1} = (\partial_x - \Gamma_0^{-1})^{-1} \Gamma_0^{-1}$.

That is, (1.9) represents an interesting new class of symmetric degenerate evolution equations for which construction of dichotomies is inadequate to bound the resolvent (3.1).

A key observation of [39] is that L^2 bounds may be obtained *directly*, using symmetry. In [39], we use for technical reasons a frequency domain/Fourier transform formulation following [19, 20]; however, this can be seen at formal level through an a priori energy estimate

$$|\langle u, g\rangle| = |\langle u, \Gamma_0 u'\rangle - \langle u, u\rangle| = |\langle u, u\rangle| = \|u\|^2 \Rightarrow \|u\| \le C\|g\|, \tag{3.2}$$

reminiscent of Friedrichs estimates for symmetric hyperbolic PDE, where $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote L^2 norm and inner product; indeed, one could view (1.9) as a "symmetric hyperbolic" analog for ODE. As in the PDE setting, the crucial property of symmetry of Γ_0 is guaranteed by existence of a convex entropy for (1.2) [12], e.g., the Boltzmann H-Theorem as discussed in Sect. 2.1.1.

3.2 Details/Counter Examples

Viewing the constant-coefficient operator $\mathscr{R}$, (3.1), as a Fourier multiplier with symbol $\hat{\mathscr{R}}(\omega) = (i\omega\Gamma_0 - \mathrm{Id})^{-1}$, and computing the uniform estimates

$$|\hat{\mathscr{R}}(\omega)| \le C, \quad |\hat{\mathscr{R}}'(\omega)| = |-\hat{\mathscr{R}} i\Gamma_0 \hat{\mathscr{R}}| \le C_2(1+|\omega|)^{-1}, \tag{3.3}$$

we find by the Mikhlin–Hormander multiplier theorem that $\mathscr{R}$ is bounded on L^p, $1 < p < \infty$.

Further detail may be obtained by spectral decomposition of Γ_0, converting $(\Gamma_0\partial_x + \mathrm{Id})w_c = g$ into a family of scalar equations $(\alpha_\lambda\partial_x - 1)u_\lambda = g_\lambda$, with u_λ the coordinate associated with spectrum α_λ and $\|u\|_{\mathbb{H}}^2 = \int |u_\lambda|^2 d\mu_\lambda$. The associated (scalar) resolvent operators $\mathscr{R}_\lambda = (\alpha_\lambda\partial_x + 1)^{-1}$ have explicit kernels

$$R_\lambda(\theta) = \alpha_\lambda^{-1} e^{(\theta)/\alpha_\lambda^{-1}},\ \theta\alpha_\lambda < 0; \qquad (\mathscr{R}_\lambda h)(x) = \int_{\mathbb{R}} R_\lambda(x-y)h(y)dy, \tag{3.4}$$

that are evidently integrable with respect to x, so bounded coordinate-wise on any $L^p(\mathbb{R}_+)$. However, explicit example [39, Example 4.7, p. 23] shows that the full operator $\mathscr{R}$ is not bounded on $L^\infty(\mathbb{R},\mathbb{H})$ (resp. $L^1(\mathbb{R},\mathbb{H})$); that is, *it is not an L^∞ (resp. L^1) multiplier.* This has the important consequence that our dynamical theory must be carried out in H^1 (bounding L^∞) rather than the usual $C^0(\mathbb{R})$ setting costing a surprising amount of technical difficulty.

The above shows also that the full resolvent kernel $R(\theta)$ determined by (3.4), considered as an operator-valued function from $\mathbb{H} \to \mathbb{H}$, *is not integrable*, since otherwise $\mathscr{R}$ by standard convolution bounds would be a bounded multiplier on all L^p. Likewise, the computation

$$|R(\theta)|_{\mathbb{H}} = \sup_{\alpha_\lambda(\theta<0} |\alpha_\lambda^{-1} e^{-\theta/\alpha_\lambda^{-1}}| \sim C/|\theta| \text{ as } \theta \to 0 \tag{3.5}$$

shows that $|R(\theta)|_{\mathbb{H}}$ *is not bounded*. This indicates the delicacy of, and cancellation involved in, the bounds on $\mathscr{R}$ obtained above through energy estimate (3.2)/resolvent bounds (3.3).

Remark 3.1. We emphasize that L^p multiplier theory/spectral decomposition is used here only to construct counterexamples, our construction of invariant manifolds relying simply on Parseval's identity.

3.3 The Banach Space Setting

(Following [50]) Weighted L^∞ spaces $L^\infty_{r,\xi}$ in velocity ξ, defined by norms $\|f\|_{L^\infty_{r,\xi}} := \sup_{\xi\in\mathbb{R}^3}(1+|\xi|)^r M(\xi)^{-1/2}|f(\xi)|, r \geq 0$, where $M(\xi) = e^{-c_0|\xi-v|^2}$ is the Maxwellian corresponding to equilibrium $\overline{\mathbf{u}}$, have been used in the study of Boltzmann's equation in, e.g., [25, 26]. Though resolvent bounds appear more difficult to obtain in this context, we can establish that $|R(\theta)|_{L^\infty_{r,\xi}}$ is not bounded, similarly as in the Hilbert case.

Recall [15] that the linearized collision operator L appearing in the linearized inhomogeneous steady Boltzmann equation $\xi_1 f' - Lf = \tilde{g}$ may be decomposed as $\tilde{L} = \tilde{\nu}(\xi) + \tilde{K}$, where $\tilde{\nu}(\xi)$ is a multiplication operator with $\tilde{\nu}(\xi) \sim \langle\xi\rangle$ and $\tilde{K}$ has kernel $|\tilde{k}(\xi,\xi_*)| \leq C|\xi-\xi_*|^{-1}e^{-c|\xi-\xi_*|^2}$, $(\tilde{K}h)(\xi) = \int_{\mathbb{R}^3}\tilde{k}(\xi,\xi_*)h(\xi_*)d\xi_*$. By $\||\xi|^{-1}e^{-c|\xi|^2}\|_{L^1} < \infty$ and standard convolution bounds, $\tilde{K}$ is bounded on $L^\infty(\xi)$, hence, by $\langle\xi\rangle/\langle\xi-\xi_*\rangle \leq C\langle\xi_*\rangle$, on $L^\infty_{r,\xi}$. In our coordinates (2.7), $Af' - Q'f = g$, $A = \frac{\xi_1}{\langle\xi\rangle}$, $Q' = -\nu(\xi) + K$, where $\nu(\xi) \sim 1$ and $K = \langle\xi\rangle^{-1/2}\hat{K}\langle\xi\rangle^{-1/2}$ is bounded from $L^\infty_{r,\xi} \to L^\infty_{r,\xi}$. The reduced equation $\Gamma_0 u' - Eu = g$ of (1.9) corresponds to the restriction of g to a finite-codimension subspace Σ of "hyperbolic modes," where $E := Q'|_\Sigma < 0$ [39]. That

$$|R(\cdot)|_{L^\infty_{r,\xi}} = \infty \tag{3.6}$$

thus follows (by contradiction, using standard convolution bounds) from the following slightly stronger statement.

Lemma 3.2 (adapted from [50]). *The solution of $Au' - Q'(\overline{\mathbf{u}})u = g$ with data g valued in a finite-codimension subspace Σ of $L^\infty_{r,\xi}$ does not satisfy a uniform bound $|u|_{L^\infty(x,L^\infty_{r,\xi})} \leq C|g|_{L^1(x,L^\infty_{r,\xi})}$.*

Proof Defining $\mathscr{S} = \big((\xi_1/\langle\xi\rangle)\partial_x - \nu(\xi)\big)^{-1}$, we have the explicit solution formula

$$(\mathscr{S}g)(x,\xi) = \int_{\mathbb{R}} S_\xi(x-y)g(y,\xi)dy; \qquad S_\xi(\theta) = (\xi_1/\langle\xi\rangle)^{-1}e^{-\nu(\xi)/(\xi_1/\langle\xi\rangle)^{-1}}, \tag{3.7}$$

where scalar kernels $S_\xi(\cdot)$ are integrable, hence $\mathscr{S}$ is bounded on $L^\infty(x, L^\infty_{r,\xi}) = L^\infty_r(\xi, (L^\infty(x)$. Writing $Au' - Q'(\overline{\mathbf{u}})u = g$ as $\big((\xi_1/\langle\xi\rangle)\partial_x - \nu(\xi)\big)u = Ku + g$, applying $\mathscr{S}$, and rearranging, we obtain $\mathscr{S}g = u - \mathscr{S}Ku$, hence $|\mathscr{S}g|_{L^\infty(x,L^\infty_{r,\xi})} \leq$

$C|u|_{L^\infty(x,L^\infty_{r,\xi})}$ by boundedness of $|K|_{L^\infty_{r,\xi}}$, $|\mathscr{S}|_{L^\infty(x,L^\infty_{r,\xi})}$. Thus, $|u|_{L^\infty(x,L^\infty_{r,\xi})} \le C|g|_{L^1(x,L^\infty_{r,\xi})}$ would imply $|\mathscr{S}g|_{L^\infty(x,L^\infty_{r,\xi})} \le C|g|_{L^1(x,L^\infty_{r,\xi})}$, or, taking $g \to \delta(x)h(\xi)$, $|S(\theta)|_{L^\infty_{r,\xi}} \le C$ for the full kernel S of $\mathscr{S}$. But, direct calculation as in (3.5) shows $|S(\theta)|_{L^\infty_{r,\xi}} \sim |\theta|^{-1}$ as $\theta \to 0$, a contradiction. □

Remark 3.3. In our notation, the bound asserted in [25] is $|R(\theta)|_{L^\infty_{5/2,\xi}} \le Ce^{-\beta|\theta|}$, in contradiction with (3.6). We conjecture that $|R(\theta)|_{L^\infty_{r,\xi}} \gtrsim |\theta|^{-1}$ as $\theta \to 0$ similarly as for its principal part $S(\theta)$, and similarly as in the Hilbert space setting (3.5), so that $|R(\cdot)|_{L^\infty_{r,\xi}} \notin L^p(\mathbb{R})$ for any $1 \le p \le \infty$.

4 H^1 Stable Manifold Theorem

We now outline the argument for construction of the stable manifold; for details, see [39].

Proof of Theorem 1.5. For clarity, we first treat the *noncharacteristic case* $m = 0$, $\dim w_c = \dim \mathscr{E} = r$, for which (1.8) becomes $w_c \equiv$ constant, $(\Gamma_0\partial_x + \mathrm{Id})w_h = \tilde{Q}_c(w)$, and the equation for the stable manifold reduces to $w_c \equiv 0$ and $(\Gamma_0\partial_x + \mathrm{Id})w_h = B(w_h, w_h)$, $B(w_h, w_h) := \tilde{Q}_c((0, w_h)$ a bounded bilinear map. Inverting, we deduce (see [39]) the fixed-point formulation

$$u(\tau) = T_S(\tau)\Pi_S u_0 + (\Gamma_0\partial_x - \mathrm{Id})^{-1}B(u,u)(\tau), \tag{4.1}$$

where Π_S, T_S denote projection and semigroup associated with the stable subspace of homogeneous flow $\Gamma_0 u' = -u$, so that $T_S(\tau)\Pi_S u_0$ is a homogeneous solution with data $\Pi_S u_0$ lying in the stable subspace at $\tau = 0$, and $\Pi_S u(0) = \Pi_S u_0$. This can be recognized as a concise, frequency-domain version of the usual variation of constants formula for finite-dimensional ODE.

However, significant new difficulties arise from the fact that, due to the properties of $(\Gamma_0\partial_x + \mathrm{Id})^{-1}$ described in Sect. 3, we must carry out the analysis in weighted H^1 rather than standard L^∞ spaces. For example, for the unbounded formal generator $-\Gamma_0^{-1}$, the H^1-stable subspace is strictly contained in the L^2-stable one, so that we must seek a graph not over the entire stable subspace but only the H^1 part, conveniently characterized as $\mathrm{dom}(\Pi_S(-\Gamma_0)^{-1/2})$. Moreover, differentiating the equation gives $u'(\tau) = T_S'(\tau)\Pi_S(u_0 - B(u(0),u(0)))(\Gamma_0\partial_x - \mathrm{Id})^{-1}B(u,u)'(\tau)$ (noting $u'(0) = -\Gamma_0^{-1}\big(u(0) - B(u(0),u(0))\big)$ by the equation) so that the "homogeneous term" involving T_S' lies in L^2 when $v_0 := u_0 - B(u(0),u(0))$, not u_0, lies in the H^1-stable subspace.

Our solution is to introduce the *modified fixed-point equation*

$$u(\tau) = T_S(\tau)\Pi_S\big(v_0 - B(u(0),u(0))\big) + (\Gamma_0\partial_x - \mathrm{Id})^{-1}B(u,u)(\tau) \tag{4.2}$$

parametrized by elements v_0 in the H^1-stable subspace, for which the derivative equation is the harmless $u'(\tau) = T_S'(\tau)\Pi_S v_0 + (\Gamma_0\partial_x - \mathrm{Id})^{-1}B(u,u)'(\tau)$. Observing that the trace $u \to u(0)$ is bounded on H^1 by 1D Sobolev embedding, as is $(\Gamma_0\partial_x + \mathrm{Id})^{-1}$ by L^2-boundedness plus commutation of constant coefficient operators with derivatives, we find that (4.2) is contractive, yielding existence/uniqueness in H^1 (and exponentially weighted H^1) norm, and thereby existence of an (exponentially decaying) stable manifold expressed as a graph over the H^1 stable subspace, Fréchet-differentiable from $\mathrm{dom}(-(\Pi_S\Gamma_0)^{-1/2})$ with norm induced by $(-\Pi_S\Gamma_0)^{-1/2}$ to the full space $\mathbb{H}$ with its original norm. A novel aspect is that the graph lies above the H^1-stable subspace not only in unstable directions, but also in stable directions lying in the stable but not H^1-stable subspace.

In the characteristic case, there is a nontrivial center equation $w_c' = Jw_c + B_c(w,w)$, coupled to the hyperbolic equation $\Gamma_0 w_h' = -w_h + B_h(w,w)$. This may be treated, setting $w = (z,u)$, by the larger fixed-point equation appending to (4.2) a standard finite-dimensional z equation:

$$\begin{aligned} z(\tau) &= -\int_\tau^{+\infty} e^{J(\tau-\theta)}B_c(w,w)(\theta)d\theta, \\ u(\tau) &= T_S(\tau)\Pi_S\big(v_0 - B_h(u(0),u(0))\big) + (\Gamma_0\partial_x - \mathrm{Id})^{-1}B_h(w,w)(\tau). \end{aligned} \tag{4.3}$$

□

Proof of Corollary 1.6. Because the stable manifold contains the forward orbits of all solutions with $H^1(\mathbb{R}_+,\mathbb{H}))$ norm sufficiently small, it contains the orbit on $\mathbb{R}_+$ of $\check{\mathbf{u}}_M := \check{\mathbf{u}}(\cdot + M)$ for M sufficiently large, whence $\check{\mathbf{u}}_M \in H^1_{\tilde\nu}(\mathbb{R}_+,\mathbb{H})$ by Theorem 1.5. It follows that $e^{\tilde\nu|\cdot|}\mathbf{u}_M \in H^1(\mathbb{R}_+,\mathbb{H})$, hence, by Sobolev embedding, $|e^{\tilde\nu|x|}\mathbf{u}(x)| \le C$, or $|\mathbf{u}(x)| \le C|e^{\tilde\nu|x|}$, for $x \ge M$. □

Remark 4.1 The key technical points in the above construction are the use of H^1 rather than sup norms to bound the resolvent, and the "integration by parts" parametrization by v_0 in (4.2).

5 Existence of a Center Manifold

Next, we outline the argument for existence of a an H^1 center manifold; for details, see [40]. The translation from standard C^0 to H^1 framework again introduces interesting new difficulties: surprisingly, different from those encountered in the stable manifold case.

Proof of Theorem 1.7. Following the standard approach to construction of center manifolds [7, 42, 44], we first replace $\tilde Q$ by a truncated nonlinearity $N_\varepsilon(w) :=$

$\rho(w/\varepsilon)\tilde{Q}(w)$, where ρ is a smooth cutoff function equal to 1 for $|w| \le 1$ and 0 for $|w| \ge 2$. The truncated nonlinearity satisfies bounds

$$|N_\varepsilon| \le c\varepsilon^2, \quad |N_\varepsilon'| \le c\varepsilon, \quad |N_\varepsilon''| \le c \qquad (|\cdot| = |\cdot|_{\mathbb{H}}), \tag{5.1}$$

and agrees with the original one locally to $\overline{\mathbf{u}}$.

Translating the usual sup-norm approach to the H^1 setting, we seek solutions to the modified (truncated) equation in a negatively weighted space $H^1_{-\alpha}$, for $\alpha > 0$ sufficiently small. Similarly as in (4.1), this yields the fixed-point formulation

$$w(\tau) = T_c(\tau)\Pi_c w_0 + \int_0^\tau T_c(\tau-\theta)\Pi_c N_\varepsilon(w(\theta))d\theta + (\Gamma_0\partial_x + \mathrm{Id})^{-1}\Pi_h N_\varepsilon(w)(\tau) \tag{5.2}$$

for solution $w = (w_c^T, w_h^T)^T$, where Π_c denotes projection onto the w_c component, $T_c(\cdot) = e^{J(\cdot)}$ the associated (nondegenerate) flow, and Π_h denotes projection onto the w_h component.

The difficulty in this case is not with the "homogeneous" term $T_c(\tau)\Pi_c w_0$ as in the stable manifold case (since derivatives on Σ_c are bounded) nor $\int_0^\tau T_c(\tau-\theta)\Pi_c N_\varepsilon(w(y))dy$, but the formerly harmless $(A\partial_x - \mathrm{Id})^{-1}\Pi_H N_\varepsilon(w,w)(\tau)$, specifically, the "substitution operator" $\mathcal{N}_\varepsilon : w \to N_\varepsilon(w)$. Bounds (5.1) yield readily that (5.2) is contractive in $L^2_{-\alpha}$ and bounded in $H^1_{-\alpha}$, $\|f\|_{H^s_{-\alpha}} := \|e^{-\alpha\langle\cdot\rangle}f(\cdot)\|_{H^s}$, giving existence and uniqueness of a $C^{0+1/2}$ center manifold $\Pi_c w_0 \to w(0)$ via the trace map $w \to w(0)$ and the 1-d Sobolev estimate $|f(0)| \le \|f\|^{1/2}_{L^2_{-\alpha}}\|\partial_x f\|^{1/2}_{L^2_{-\alpha}}$.

However, higher (even Lipschitz) regularity seems to require contraction in $\|\cdot\|_{H^1_{-\alpha}}$, the difficulty lying in term

$$\|\partial_x(N_\varepsilon(v_1) - N_\varepsilon(v_2))\|_{L^2_{-\alpha}} \sim \|\max_j(|N_\varepsilon''(v_j)||\partial_x v_j|)|v_2 - v_1|\|_{L^2_{-\alpha}} \sim \sum_j \||\partial_x v_j||v_2 - v_1|\|_{L^2_{-\alpha}},$$

for which the obvious Sobolev embedding estimate gives $\|v_1 - v_2\|_{H^1_{-\alpha}}\sum_j(\int_{\mathbb{R}}|\partial_x v_j|^2)^{1/2} = +\infty$.

A key observation is that, for $0 < \alpha_1 \ll \alpha \ll \alpha_2 \ll 1$, (5.2) is contractive in the *mixed norm*

$$\|f\| := \|f\|_{L^2_{-\alpha}} + \|\partial_x f\|_{L^2_{-\alpha_2}} \tag{5.3}$$

and bounded in $H^1_{-\alpha_1}$ for $\|w\|_{H^1_{-\alpha_1}} \ll 1$. For, the Sobolev bound

$$e^{-2\alpha_2\langle x\rangle}|f(x)|^2 \le \|f\|_{L^2_{-\alpha_2}(x,\infty)}\|\partial_x f\|_{L^2_{-\alpha_2}(x,\infty)} \le e^{-(\alpha_2-\alpha)\langle x\rangle}\|f\|_{L^2_{-\alpha}(x,\infty)}\|\partial_x f\|_{L^2_{-\alpha_2}(x,\infty)}$$

gives

$$\|\partial_x(N_\varepsilon(v_1)-N_\varepsilon(v_2))\|^2_{L^2_{-\alpha_2}} \lesssim \int_{\mathbb{R}} e^{-2\alpha_2\langle x\rangle}|v_1(x)-v_2(x)|^2|\partial_x v_1(x)|^2dx$$
$$\lesssim \Big(\int_{\mathbb{R}} e^{-(\alpha_2-\alpha)\langle x\rangle}|\partial_x v_1(x)|^2dx\Big)\|v_1-v_2\|^2 \lesssim \|\partial_x v_1(x)\|^2_{H^1_{-\alpha_1}}\|v_1-v_2\|^2. \tag{5.4}$$

With this observation, working in norm $\|\cdot\|$, we obtain essentially immediately existence and uniqueness of a global center manifold for the truncated equation/ local center manifold for the exact equation that is *Lipschitz continuous,* as a graph over the center subspace Σ_c. C^r (Fréchet) regularity, $r \geq 1$ may then be obtained similarly as in the finite-dimensional case [7, 16, 42, 44], by a bootstrap argument, using a nested sequence of mixed-weight norms together with a general result on smooth dependence with respect to parameters of a fixed-point mapping $y = T(x, y)$ that is Fréchet differentiable in y from a stronger to a weaker Banach space, with differential T_y extending to a bounded, contractive map on the weaker space [44, Lemma 2.5, p. 53] ([46, Lemma 3, p. 132]). See [40, Appendix A], for further details. The H^1 exponential approximation property (not discussed in [40]) follows by transcription to the H^1 setting of the finite-dimensional argument given in [7, Step 7, p. 9]. □

Remark 5.1. The estimate (5.4), and introduction of norm (5.3), we view as the crucial technical points in our construction of center manifolds, and the main novelty in this part of the analysis.

6 Structure of Small-Amplitude Kinetic Shocks

Given existence of a center manifold, one may in principle obtain an arbitrarily accurate description of near-equilibrium dynamics via formal Taylor expansion/reduction to normal form. We give here a particularly simple normal form argument describing bifurcation of stationary shock profiles from a simple genuinely nonlinear characteristic equilibrium, adapting more general center manifold arguments of [28, 29] in the finite-dimensional case. Similarly as in [28, 29], the main idea is to use the fact that equilibria are predicted by the Rankine–Hugoniot shock conditions (RH) to deduce normal form information from the structure of the Chapman–Enskog approximation (CE).

Lemma 6.1. *Let* $\overline{\mathbf{u}} \in \ker Q$ *be an equilibrium satisfying (H1)-(H2). In the simple genuinely nonlinear characteristic case* (GNL), $m = 1$, *the center manifolds of* (1.3) *and* (CE) *both consist of the union of one-dimensional fibers parametrized by* $q \in \mathbb{R}^r$ *as in* (RH) *and coordinatized by* u_1 *as in* (2.11), *satisfying an approximate Burgers flow: without loss of generality*

$$\tilde{q} = 0, \qquad u_1' = \delta^{-1}\big(-q_1 + \Lambda u_1^2/2\big) + O(|u_1|^3 + |q_1||u_1| + |q_1|^2), \tag{6.1}$$

where $\delta := \bar{\mathbf{r}}^T D_* \bar{\mathbf{r}} > 0$ *with* $\bar{\mathbf{r}}$, D_* *as in* (GNL), (CE). *In particular, under the normalization* $\tilde{q} = 0$, *there exist local heteroclinic (Lax shock) connections for* $q_1\Lambda < 0$ *between endstates* $u_1^{\pm} \approx \sqrt{-2q_1/\Lambda}$.

Proof. First, note that $T_{12}v_1$ in the original coordinates of (2.19) is exactly the first component q_1 of q in (RH), or $v_1 = T_{12}^{-1}q_1$. By Observation 2.3 and the Implicit Function Theorem, we may take without loss of generality $\tilde{q} = 0$ by a shift along equilibrium manifold $\mathscr{E}$ of the background equilibrium $\bar{\mathbf{u}}$. By (1.8), therefore, the flow on the $(r+1)$-dimensional center manifold has an r-dimensional constant of motion

$$\big(w_{c,2}, w_{c,3}\big) \equiv (\zeta, \gamma) = \big(-(T_{12}^*)^{-1}v_1, \tilde{q}\big) = \big((-T_{12}^*)^{-1}T_{12}^{-1}q_1, 0\big), \tag{6.2}$$

w as in (2.20), with flow along one-dimensional fibers coordinatized by $w_{c,1} = u_1 - \Gamma_1\tilde{v} = u_1 - \Gamma_1 w_{\mathrm{h}}$ given by the $w_{\mathrm{c},1}$ equation of (1.8):

$$w_{c,1}' = \zeta + \phi(w_{c,1}, \zeta), \quad \phi(w_{c,1}, \zeta) := g_{c,1}\big((w_{c,1}, \zeta, 0), \Xi(w_{c,1}, \zeta, 0)\big) = \mathscr{O}(\|w_{c,1}\|^2, \|\zeta\|^2). \tag{6.3}$$

The factor $(T_{12}^*)^{-1}T_{12}^{-1} > 0$ in term $\zeta = -(T_{12}^*)^{-1}T_{12}^{-1}q_1$ is easily recognized as δ^{-1}, where $\delta := T_{12}T_{12}^* > 0$, or, using $\bar{\mathbf{r}} = e_1$, $\delta = \bar{\mathbf{r}} \cdot D_*\bar{\mathbf{r}}$ with D_* as in (CE). Using the fact that $w_{\mathrm{h}} = \mathscr{J}(w_{\mathrm{c}}) = O(|w_{\mathrm{c}}|^2)$ along the center manifold to trade $w_{c,1}$ for u_1 by an invertible coordinate change preserving the order of error terms, we may thus rewrite (6.3) as

$$u_1' = \delta^{-1}(-q_1 + \delta\chi u_1^2) + O(|u_1|^3 + |u_1||q_1| + |q_1|^2), \tag{6.4}$$

where χ, hence the product $\delta\chi$, is yet to be determined. On the other hand, performing Lyapunov–Schmidt reduction for the equilibrium problem (RH), we obtain the normal form

$$0 = (-q_1 + \frac{1}{2}\Lambda u_1^2) + O(|u_1|^3 + |u_1||q_1| + |q_1|^2),$$

where Λ is as in (GNL). Using the fact that equilibria for (1.3) and (RH) agree, we find that $\delta\chi$ must be equal to $\frac{1}{2}\Lambda$, yielding a final normal form consisting of the approximate Burgers flow (1.13). A similar computation yields the same normal form for fibers of the center manifold of the formal viscous problem (CE); see also the more detailed computations of [28] yielding the same result.

For $q_1\Lambda > 0$, the scalar Eq. (1.13) evidently possesses equilibria $\sim \mp\sqrt{2q_1/\Lambda}$, connected (since the equation is scalar) by a heteroclinic profile. Since $\operatorname{sgn} u_1' = -\operatorname{sgn}\Lambda$ for u_1 between the equilibria, so that $\big(\lambda(u)\big)' \sim \Lambda u_1'$ has sign of $-\Lambda^2 < 0$, the connection is in the direction of decreasing characteristic $\lambda(u)$, corresponding to a Lax-type solution of (RH) (cf. [28, 29]). □

Remark 6.2. Using $\lambda(u_1) \sim \Lambda u_1$, we may rewrite (6.1) as (1.13) as in the introduction, eliminating the $\tilde{q}$-dependent term Λ. However, the "effective viscosity" δ remains dependent on $\tilde{q}$.

Having determined the normal form (1.13), we establish closeness of profiles of (1.3) and (CE) by comparing their u_1 coordinates, separately, to an exact Burgers shock, then showing that differences in remaining, slaved, coordinates, since vanishing at both endstates, are negligibly small.

Lemma 6.3 ([22, 38]). *Let $\eta \in \mathbb{R}^1$ be a heteroclinic connection of an approximate Burgers equation*

$$\delta\eta' = \frac{1}{2}\Lambda(-\varepsilon^2 + \eta^2) + S(\varepsilon, \eta), \quad S = O(|\eta|^3 + |\varepsilon|^3) \in C^{k+1}(\mathbb{R}^2), \quad k \geq 0, \tag{6.5}$$

and $\bar{\eta} := -\varepsilon \tanh(\Lambda\varepsilon x/2\delta)$ a connection of the exact Burgers equation $\delta\bar{\eta}' = \frac{1}{2}\Lambda(-\varepsilon^2 + \bar{\eta}^2)$. Then,

$$\begin{aligned} |\eta_\pm - \bar{\eta}_\pm| &\leq C\varepsilon^2, \\ |\partial_x^k(\bar{\eta} - \bar{\eta}_\pm)(x)| &\sim \varepsilon^{k+1} e^{-\delta\varepsilon|x|}, \quad x \gtrless 0, \quad \delta > 0, \\ \left|\partial_x^k\big((\eta - \eta_\pm) - (\bar{\eta} - \bar{\eta}_\pm)\big)(x)\right| &\leq C\varepsilon^{k+2} e^{-\delta\varepsilon|x|}, \quad x \gtrless 0, \end{aligned} \tag{6.6}$$

uniformly in $\varepsilon > 0$, where $\eta_\pm := \eta(\pm\infty)$, $\bar{\eta}_\pm := \bar{\eta}(\pm\infty) = \mp\varepsilon$ denote endstates of the connections.

Proof. (From [40], following [22]) Rescaling $\eta \to \eta/\varepsilon$, $x \to \Lambda\varepsilon\tilde{x}/\beta$, we obtain the blowup equations

$$\eta' = \frac{1}{2}(\eta^2 - 1) + \varepsilon\tilde{S}(\eta, \varepsilon) \quad \tilde{S} \in C^{k+1}(\mathbb{R}^2)$$

and $\bar{\eta}' = \frac{1}{2}(\bar{\eta}^2 - 1)$, for which estimates (6.5) translate to

$$\begin{aligned} |\eta_\pm - \bar{\eta}_\pm| &\leq C\varepsilon, \\ |\partial_x^k(\bar{\eta} - \bar{\eta}_\pm)(x)| &\sim C\varepsilon^k e^{-\theta|x|}, \quad x \gtrless 0, \quad \theta > 0, \\ |\partial_x^k\big((\eta - \eta_\pm) - (\bar{\eta} - \bar{\eta}_\pm)\big)(x)| &\leq C\varepsilon^{k+1} e^{-\theta|x|}, \quad x \gtrless 0. \end{aligned} \tag{6.7}$$

The estimates (6.7) follow readily from the implicit function theorem and stable manifold theorems together with smooth dependence on parameters of solutions of ODE, giving the result. □

Setting $q_1 = \Lambda\varepsilon^2/2$, and either $\eta = u_{REL,1}$ or $\eta = u_{CE,1}$, we obtain approximate Burgers equation (6.5), and thereby estimates (6.6) relating $\eta = u_{REL,1}, u_{CE,1}$ to an exact Burgers shock $\bar{\eta}$.

Corollary 6.4 ([40]). *Let $\bar{\mathbf{u}} \in \ker Q$ be an equilibrium satisfying (H1)-(H2), in the characteristic case* (GNL)*, and k and integer ≥ 2. Then, local to $\bar{\mathbf{u}}$ ($\bar{u}$), each pair of points $u_\pm$ corresponding to a standing Lax-type shock of* (RH) *has a corresponding viscous shock solution u_{CE} of* (CE) *and relaxation shock solution $\mathbf{u}_{REL} = (u_{REL}, v_{REL})$ of* (1.3)*, satisfying for all $j \leq k-2$:*

$$\begin{aligned} |\partial_x^j (u_{REL,1} - u^\pm_{REL,1})(x)| &\sim C\varepsilon^j e^{-\theta|x|}, \quad x \gtrless 0, \quad \theta > 0, \\ |\partial_x^j (u_{REL,1} - u_{CE,1})(x)| &\leq C\varepsilon^{j+1} e^{-\theta|x|}, \quad x \gtrless 0. \end{aligned} \tag{6.8}$$

Proof. Immediate, by (6.7), Lemma 6.3 and the triangle inequality, together with the observation that, as equilibria of (CE) and (1.3), hence solutions of (RH), endstates $u^\pm_{REL,1} = u^\pm_{CE,1}$ agree. □

Proof of Corollary 1.9. ([40]) Noting that the $\mathrm{im}A_{11}$ and $\mathbb{V}$ components of $\mathbf{u}_{REL}$ are the C^2 functions $\Psi(u_{REL,1})$, $\Phi(u_{REL,1})$ of $u_{REL,1}$ along the fiber (1.13), we obtain (1.14)(iii) immediately from (6.8)(i). Denote by Ψ_{CE} the map describing the dependence of $\mathrm{im}A_{11}$ component of u_{CE} on $u_{CE,1}$ on the corresponding fiber of (CE). Since $\Psi - \Psi_{CE}$ and $\Phi - v_*$ both vanish at the endstates $u^\pm_{REL,1}$, we have by smoothness of Ψ, Ψ_{CE}, Φ, v_* that

$$|\Psi - \Psi_{CE}|, \ |\Phi - v_*| = \mathscr{O}(|u_{REL,1} - u^+_{REL,1}|, |u_{REL,1} - u^-_{REL,1}|),$$

giving (1.14)(i)–(ii) by (6.8)(i)–(ii). □

Remark 6.5. Applied to Boltzmann's equation, Corollary 1.9 yields existence/convergence to hydrodynamic shock profiles in the square-root Maxwellian-weighted norm (2.6). Using a bootstrap argument analogous to that of [31, Proposition 3.1], one can show [40, Proposition 1.8] that the center manifold of Theorem 1.7 lies in the stronger spaces determined by near-Maxwellian-weighted norms $\|f\|_{\mathbb{H}^s} := \|f\underline{M}^{-s}\|_{L^2(\mathbb{R}^3)}$, $1/2 \leq s < 1$, yielding further information on localization of velocity in small-amplitude shock profiles. This and the streamlined proof of existence above are the main novelties in our treatment by center manifold techniques of existence and structure of kinetic shocks.

References

1. A. Abbondandolo, P. Majer, Ordinary differential operators in Hilbert spaces and Fredholm pairs. Math. Z. **243**, 525–562 (2003)
2. A. Abbondandolo, P. Majer, Morse homology on Hilbert spaces. Comm. Pure Appl. Math. **54**, 689–760 (2001)
3. B. Barker, Numerical proof of stability of roll waves in the small-amplitude limit for inclined thin film flow. J. Diff. Eq. **257**(8), 2950–2983 (2014)
4. B. Barker, K. Zumbrun, Numerical proof of stability of viscous shock profiles. Math. Models Meth. Appl. Sci. (to appear)

5. H. Bart, I. Gohberg, M.A. Kaashoek, Wiener-Hopf factorization, inverse Fourier transforms and exponentially dichotomous operators. J. Funct. Anal. **68**(1), 1–42 (1986)
6. G. Boillat, T. Ruggeri, On the shock structure problem for hyperbolic system of balance laws and convex entropy. Continuum Mech. Thermodyn. **10**(5), 285–292
7. A. Bressan, *A Tutorial on the Center Manifold Theorem, Appendix A, Hyperbolic Systems of Balance Laws*, Lecture Notes in Mathematical, vol. 1911, (Springer-Verlag, Heidelberg, 2007)
8. R. Caflisch, B. Nicolaenko, Shock profile solutions of the Boltzmann equation. Comm. Math. Phys. **86**(2), 161–194 (1982)
9. T. Carleman, *Sur la theorie des equations integrales et ses applications*, Verhandl. des Internat. Math. Kong., I, Zurich (1932), pp. 138–151
10. J. Carr, *Applications of Centre Manifold Theory*. Applied Mathematical Sciences, vol. 35 (Springer-Verlag, New York-Berlin, 1981), vi+142 pp. ISBN: 0-387-90577-4
11. C. Cercignani, *The Boltzmann Equation and Its Applications*, Applied Mathematical Sciences, vol. 67. (Springer-Verlag, New York, 1988), xii+455 pp. ISBN: 0-387-96637-4
12. G.Q. Chen, C.D. Levermore, T.P. Liu, Hyperbolic conservation laws with stiff relaxation terms and entropy. Comm. Pure Appl. Math. **47**, 787–830 (1994)
13. R.A. Gardner, K. Zumbrun, The gap lemma and geometric criteria for instability of viscous shock profiles. Comm. Pure Appl. Math. **51**(7), 797–855 (1998)
14. R. Glassey, *The Cauchy Problem in Kinetic Theory* (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996), xii+241 pp. ISBN: 0-89871-367-6
15. H. Grad, *Asymptotic theory of the Boltzmann equation. II*, in *Proceedings of the 3rd International Conference on Rarefied Gases, Palais de l'UNESCO, Paris*, vol. I (Academic Press, New York, 1962), pp. 26–59
16. M. Haragus and G. Ioos, *Local Bifurcations, Center Manifolds, and Normal Forms in Infinite-dimensional Dynamical Systems*, Universitext. (Springer-Verlag London, Ltd., London; EDP Sciences, Les Ulis, 2011), xii+329 pp. ISBN: 978-0-85729-111-0; 978-2-7598-0009-4
17. J. Humpherys, G. Lyng, K. Zumbrun, Multidimensional stability of large-amplitude Navier-Stokes shocks, arXiv:1603.03955
18. S. Kawashima, Systems of a hyperbolic–parabolic composite type, with applications to the equations of magnetohydrodynamics, thesis, Kyoto University, 1983
19. Y. Latushkin, A. Pogan, The dichotomy theorem for evolution bi-families. J. Diff. Eq. **245**(8), 2267–2306 (2008)
20. Y. Latushkin, A. Pogan, The infinite dimensional evans function. J. Funct Anal. **268**(6), 1509–1586 (2015)
21. P.D. Lax, Hyperbolic systems of conservation laws and the mathematical theory of shock waves, in *Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics*, No. 11 (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1973), v+48 pp
22. Y. Li, Scalar Green function bounds for instantaneous shock location and one-dimensional stability of viscous shock waves. Quart. App. Math. (to appear)
23. T.-P. Liu, Hyperbolic conservation laws with relaxation. Comm. Math. Phys. **108**(1), 153–175 (1987)
24. T.P. Liu, S.H. Yu, Boltzmann equation: micro-macro decompositions and positivity of shock profiles. Comm. Math. Phys. **246**(1), 133–179 (2004)
25. T.P. Liu, S.H. Yu, Invariant manifolds for steady boltzmann flows and applications. Arch. Rational Mech. Anal. **209**, 869–997 (2013)
26. T.-P. Liu, S.-H. Yu, The Greens function and large-time behavior of solutions for the one-dimensional Boltzmann equation. Comm. Pure Appl. Math. **57**(7), 841–876 (2004)
27. J. Mallet-Paret, The Fredholm alternative for functional-differential equations of mixed type. J. Dyn. Diff. Eq. **11**, 1–47 (1999)
28. A. Majda, R. Pego, Stable viscosity matrices for systems of conservation laws. J. Diff. Eqs. **56**, 229–262 (1985)
29. C. Mascia, K. Zumbrun, Pointwise Green's function bounds and stability of relaxation shocks. Indiana Univ. Math. J. **51**(4), 773–904 (2002)

30. A. Mielke, Reduction of quasilinear elliptic equations in cylindrical domains with applications. Math. Methods Appl. Sci. **10**, 51–66 (1988)
31. G. Métivier, K. Zumbrun, Existence and sharp localization in velocity of small-amplitude Boltzmann shocks. Kinet. Relat. Models **2**(4), 667–705 (2009)
32. R.L. Pego, Stable viscosities and shock profiles for systems of conservation laws. Trans. Amer. Math. Soc. **282**, 749–763 (1984)
33. D. Peterhof, B. Sandstede, A. Scheel, Exponential dichotomies for solitary-wave solutions of semilinear elliptic equations on infinite cylinders. J. Diff. Eq. **140**, 266–308 (1997)
34. J. Robbin, D. Salamon, The spectral flow and the Maslov index. Bull. Lond. Math. Soc. **27**, 1–33 (1995)
35. B. Sandstede, Stability of traveling waves, in *Handbook of Dynamical Systems*, vol. 2, (North-Holland, Amsterdam, 2002), pp. 983–1055
36. B. Sandstede, A. Scheel, On the structure of spectra of modulated traveling waves. Math. Nachr. **232**, 39–93 (2001)
37. B. Sandstede, A. Scheel, Relative Morse indices, Fredholm indices, and group velocities. Discrete Contin. Dyn. Syst. A **20**, 139–158 (2008)
38. R. Plaza, K. Zumbrun, Evans function approach to spectral stability of small-amplitude shock profiles. Discrete Contin. Dyn. Syst. **10**, 885–924 (2004)
39. A. Pogan, K. Zumbrun, Stable manifolds for a class of degenerate evolution equations and exponential decay of kinetic shocks, arXiv:1607.03028
40. A. Pogan, K. Zumbrun, Center manifolds of degenerate evolution equations and existence of small-amplitude kinetic shocks, arXiv:1612.05676
41. J. Smoller, *Shock Waves and Reaction–Diffusion Equations*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 258, 2nd edn. (Springer-Verlag, New York, 1994), xxiv+632 pp. ISBN: 0-387-94259-9
42. A. Vanderbauwhede and G. Iooss, *Center manifold theory in infinite dimensions*, in *Dynamics Reported: Expositions in Dynamical Systems*, vol. 1 (Springer, Heidelberg, 1992), pp. 125–163
43. W.-A. Yong, Basic structures of hyperbolic relaxation systems. Proc. R. Soc. Edinb. Sect. A **132**(5), 1259–1274 (2002)
44. K. Zumbrun, Conditional stability of unstable viscous shocks. J. Diff. Eq. **247**(2), 648–671 (2009)
45. K. Zumbrun, *Ordinary Differential Equations*, Indiana University, Lecture notes for graduate ODE (2009)
46. K. Zumbrun, Multidimensional stability of planar viscous shock waves, in *Advances in the Theory of Shock Waves, Progress in Nonlinear Differential Equations and Their Applications*, vol. 47 (Birkhäuser Boston, Boston, MA, 2001), pp. 307–516
47. K. Zumbrun, H.K. Jenssen, G. Lyng, Stability of large-amplitude shock waves of compressible Navier–Stokes equations, in *Handbook of Mathematical Fluid Dynamics*, vol. III (North-Holland, Amsterdam, 2004), pp. 311–533
48. K. Zumbrun, Planar stability criteria for viscous shock waves of systems with real viscosity, in *Hyperbolic Systems of Balance Laws*, Lecture Notes in Mathematics, vol. 1911, (Springer, Heidelberg, 2007), pp. 229–326
49. K. Zumbrun, *Stability and dynamics of viscous shock waves*, in *Nonlinear Conservation Laws and Applications, The IMA Volumes in Mathematics and its Applications*, vol. 153, (Springer, New York, 2011), pp. 123–167
50. K. Zumbrun, L^∞ resolvent estimates for steady Boltzmann's equation, arXiv:1612.06916
51. K. Zumbrun, P. Howard, Pointwise semigroup methods and stability of viscous shock waves. Indiana Math. J. **47**, 741–871 (1998); Errata. Indiana Univ. Math. J. **51**(4), 1017–1021 (2002)
52. K. Zumbrun, D. Serre, Viscous and inviscid stability of multidimensional planar shock fronts. Indiana Univ. Math. J. **48**, 937–992 (1999)