CONTENTS

Preface ................................................................. i

Part 1: Plenary Lectures 1–107

Camillo De Lellis, Elisabetta Chiodaroli and Ondřej Kreml
Surprising solutions to the isentropic Euler system of gas dynamics .......... 1

Eduard Feireisl
Relative entropies, dissipative solutions, and singular limits of complete fluid systems ................................................................. 11

David Gérard-Varet and Nader Masmoudi
Recent progress in the theory of homogenization with oscillating Dirichlet data ................................................................. 29

David Gérard-Varet and Nader Masmoudi
Relative entropies, dissipative solutions, and singular limits of complete fluid systems ................................................................. 11

Franz Fuchs, Andrew McMurry, Nils Henrik Risebro and Siddhartha Mishra
Explicit and implicit finite volume schemes for radiation MHD and the effects of radiation on wave propagation in stratified atmospheres .............. 41

Sebastiano Boscarino and Giovanni Russo
Implicit-explicit Runge-Kutta schemes for hyperbolic systems with stiff relaxation and applications ......................................................... 61

Robert Saye and James Sethian
New interface methods for tracking multiphase multiphysics .................... 81

Yannick Privat, Emmanuel Trélat and Enrique Zuazua
Optimal sensor location for wave and Schrödinger equations .................. 89

Part 2: Invited Lectures 109–232

Gianluca Crippa
Ordinary differential equations and singular integrals .............................. 109

Jean-Luc Guermond and Bojan Popov
Entropy viscosity for the Euler equations and questions regarding parabolic regularization ................................................................. 119

Soohyun Bae, Sun-Ho Choi and Seung-Yeal Ha
Dynamic instability of the Vlasov-Poisson-Boltzmann system in high dimensions ................................................................. 125

Rinaldo M. Colombo and Michael Herty
Nodal conditions for hyperbolic systems of balance laws .......................... 147

Corrado Lattanzio and Athanasios E. Tzavaras
Relative entropy methods for hyperbolic and diffusive limits .................. 163

Pierre Degond, Amic Frouvelle and Jian-Guo Liu
A note on phase transitions for the Smoluchowski equation with dipolar potential ................................................................. 179

Katrin Grunert, Helge Holden and Xavier Raynaud
Lipschitz metric for the two-component Camassa–Holm system ............. 193
Moritz Reintjes
Points of shock wave interaction are ‘regularity singularities’ in spacetime .... 209

Michael Shearer
Two fluid flow in porous media ................................................................. 221

Part 3: Contributed Talks 233–1066

Eduardo Abreu
Numerical simulation of three-phase flow in heterogeneous media with
spatially varying nonlinear hyperbolic-parabolic flux functions ............. 233

Franz Achleitner, Sabine Hittmeir and Christian Schmeiser
On nonlinear conservation laws regularized by a Riesz-Feller operator ...... 241

Fabio Ancona, Olivier Glass and Khai T. Nguyen
On quantitative compactness estimates for hyperbolic conservation laws .... 249

Boris Andreianov
One-dimensional conservation law with boundary conditions: General results
and spatially inhomogeneous case ............................................................ 259

Denise Aregba-Driollet
The Riemann problem for Kerr equations and non-uniqueness of selfsimilar
entropy solutions ....................................................................................... 269

Thomas Auphan, Philippe Angot and Olivier Guès
Penalty methods for edge plasma transport in a tokamak ....................... 277

Hantaek Bae and Konstantina Trivisa
On the Doi model for the suspensions of rod-like molecules in compressible
fluids ........................................................................................................... 285

Luís Almeida and Patrizia Bagnerini
Multi-scale tissular-cellular model for wound healing ............................. 293

Joshua Ballew
Low Mach number limits to the Navier-Stokes-Smoluchowski system ....... 301

Caroline Bauzet
Method for solving a stochastic conservation law .................................... 309

Stefan Berres, Frank Huth, Hartmut Schwandt and
Günter Bärwolf
A multi-phase model for pedestrian flow with strategic and tactic component . 317

Luigi C. Berselli
Some results on the two-dimensional dissipative Euler equations ........... 325

Raul Borsche and Jochen Kall
ADER-schemes on networks of scalar conservation laws .......................... 333

Laurent Boudin and Julien Mathiaud
Asymptotic behavior of a diffusive scheme solving the inviscid one-dimensional
pressureless gases system ......................................................................... 341

Benjamin Boutin, Frédéric Coquel and Philippe G. LeFloch
Coupling techniques for nonlinear hyperbolic equations .......................... 349

Alberto Bressan
Globally optimal and Nash equilibrium solutions for traffic flow on networks . 357
M. Briani, D. Aregba-Driollet and R. Natalini
How to improve the decay of the numerical error for large times: The case of dissipative BGK systems ................................................................. 365

Buğra Kabil
On the asymptotics of solutions to resonator equations ......................... 373

Raimund Bürger, Pep Mulet and Luis M. Villada
Adaptive mesh refinement for spectral WENO schemes for efficient simulation of polydisperse sedimentation processes ................................ 381

Sunčica Čanić and Boris Muha
A nonlinear moving-boundary problem of parabolic-hyperbolic-hyperbolic type arising in fluid-multi-layered structure interaction problems ............. 389

Giovanni Alberti, Stefano Bianchini and Laura Caravenna
Reduction on characteristics for continuous solutions of a scalar balance law .. 399

Filipe Carvalho and Ana Jacinta Soares
Detonation wave problems: Modeling, numerical simulations and linear stability ............................................................................................... 407

Pablo Castañeda, Dan Marchesin and Frederico Furtado
On singular points for convex permeability models .................................... 415

Fausto Cavalli
Linearly implicit schemes for convection-diffusion equations ................. 423

Nikolai Chemetov and Fernanda Cipriano
The inviscid limit for slip boundary conditions ........................................ 431

Koottungal Revi Arun, Guoxian Chen and Sebastian Noelle
A finite volume evolution Galerkin scheme for acoustic waves in heterogeneous media .......................................................... 439

Fabio Ancona and Giuseppe Maria Coclite
On the asymptotic stabilization of a generalized hyperelastic-rod wave equation ................................................................................................. 447

Armando Coco and Giovanni Russo
Boundary treatment in ghost point finite difference methods for compressible gas dynamics in domain with moving boundaries ................. 455

Rinaldo M. Colombo, Mauro Garavello, Magali Lecureux-Mercier and Nikolay Pogodaev
Conservation laws in the modeling of moving crowds .................................. 467

Andrea Corli and Haitao Fan
A hyperbolic model for phase transitions in porous media ...................... 475

Kristian Debrabant and Espen R. Jakobsen
Semi-Lagrangian schemes for linear and fully non-linear Hamilton-Jacobi-Bellman equations ................................................................. 483

A.I. Delis and I.K. Nikolos
On a solution reconstruction and limiting procedure for unstructured finite volumes .......................................................................................... 491

Maria Laura Delle Monache and Paola Goatin
A strongly coupled PDE-ODE system modeling moving density
A priori analysis of asymptotic preserving schemes with the modified equation

Liviu Florin Dinu
Wave-wave interactions of a gasdynamic type

Donatella Donatelli and Pierangelo Marcati
Analysis of oscillations and defect measures in plasma physics

Renjun Duan
Asymptotic stability of kinetic plasmas for general collision potentials

Arnaud Duran, Fabien Marche, Christophe Berthon and Rodolphe Turpault
Numerical discretizations for shallow water equations with source terms on unstructured meshes

Jan Ernest
Schemes with Well-Controlled Dissipation (WCD) for scalar conservation laws with pseudo-parabolic regularization

Adriano Festa and Maurizio Falcone
$L^1$ convergence of a SL scheme for the eikonal equation with discontinuous coefficients

Hermano Frid
A brief note on recent developments on divergence-measure fields

Raul Borsche, Rinaldo M. Colombo and Mauro Garavello
Mixed systems with boundaries

Mohamed-Gazibo Karimou
Degenerate convection-diffusion equation with a Robin boundary condition

Christoph Gersbacher
Explicit higher order schemes for the coupling of dimensionally heterogeneous free-surface flow models

Jan Giesselmann and Athanasios E. Tzavaras
On cavitation in elastodynamics

Christophe Chalons, Mathieu Girardin and Samuel Kokh
Operator-splitting-based asymptotic preserving scheme for the gas dynamics equations with stiff source terms

Christian Bourdarias, Marguerite Gisclon and Stéphane Junca
Blow-up or not blow-up at the hyperbolic boundary for a system from chemistry?

Hyunkyung Lim, Yijie Zhou, Valmor F. de Almeida and James Glimm
Fully developed turbulent mixing in an annular sector

Frédéric Coquel, Edwige Godlewski and Nicolas Seguin
Fluid systems: Approximation by relaxation and coupling
Debora Amadori, Rinaldo M. Colombo, Graziano Guerra and Wen Shen
Slow erosion of granular flow: Continuous and discontinuous profiles ........ 641

Maren Hantke and Ferdinand Thein
Numerical solutions to the Riemann problem for compressible isothermal
Euler equations for two phase flows with and without phase transition ........ 651

Itsuko Hashimoto
Asymptotic behavior of solutions for damped wave equations with non-convex
convection term on the half line ......................................................... 659

Boris Haspot
Porous media equations, fast diffusions equations and the existence of global
weak solution for the quasi-solutions of compressible Navier-Stokes equations 667

Christiane Helzel and Marsha J. Berger
Cartesian grid embedded boundary methods for hyperbolic problems ........... 675

Zhenying Hong, Guangwei Yuan and Xuedong Fu
Time and space discrete scheme to suppress numerical solution oscillation
for the neutron transport equations .................................................... 685

François James and Nicolas Vauchelet
Numerical simulation of a hyperbolic model for chemotaxis after blow up ........ 693

Quansen Jiu, Yi Wang and Zhouping Xin
Global well-posedness of 2D compressible Navier-Stokes equations with large
data and vacuum ................................................................. 701

Pierre Castelli and Stéphane Junca
Oscillating waves and optimal smoothing effect for one-dimensional
nonlinear scalar conservation laws .................................................... 709

Dante Kalise
A WENO-TVD finite volume scheme for the approximation of atmospheric
phenomena ................................................................. 717

Friedemann Kemm
An enhancement to the AUFS flux splitting scheme by Sun and Takayama ... 725

K. Sudarshan Kumar, G. D. Veerappa Gowda and C. Praveen
On polymer flooding problem in oil reservoir simulation .......................... 733

Christian Klingenberg and Ujjwal Koley
An error estimate for viscous approximate solutions to degenerate anisotropic
convection-diffusion equations ....................................................... 741

Michael Benfield, Helge Kristian Jenssen and Irina A. Kogan
1-D conservative systems: A geometric approach ................................. 749

Ujjwal Koley and Nils Henrik Risebro
Convergence of a finite difference scheme for $2 \times 2$ Keyfitz-Kranzer system ... 759

Mirko Kränkel
A local discontinuous Galerkin scheme for compressible phase field flow ...... 767

Gerhard Dziuk, Dietmar Kröner and Thomas Müller
Existence and uniqueness for scalar conservation laws on moving
hypersurfaces ................................................................. 775
Alexei A. Mailybaev
Renormalization and universal structure of blowup in 1D conservation laws ... 783

Vítor Matos, Pablo Castañeda and Dan Marchesin
Classification of the umbilic point in immiscible three-phase flow in porous media ... 791

Thomas Müller and Axel Pfeiffer
Well-balanced simulation of geophysical flows via the shallow water equations with bottom topography: Consistency and numerical computation ... 801

Martin Nolte
Approximation of the effective Hamiltonian through a degenerate elliptic problem ... 809

Shinya Nishibata, Masashi Ohnawa and Masahiro Suzuki
Nonlinear stability of boundary layer solutions to the Euler-Poisson equations in plasma physics ... 817

Sigrun Ortleb and Andreas Meister
A well-balanced DG scheme with unconditionally positive implicit time integration ... 823

Evgeniy Yu. Panov
On decay of periodic renormalized solutions to scalar conservation laws ... 831

Marica Pelanti and Keh-Ming Shyue
A mixture-energy-consistent numerical approximation of a two-phase flow model for fluids with interfaces and cavitation ... 839

Carlo Morosi, Mario Pernici and Livio Pizzocchero
A posteriori estimates for Euler and Navier-Stokes equations ... 847

Ramón G. Plaza
On the stability of degenerate viscous shock profiles ... 857

Giacomo Dimarco, Lorenzo Pareschi and Vittorio Rispoli
Asymptotically implicit schemes for the hyperbolic heat equation ... 865

Volker Elling and Joseph Roberts
Steady self-similar inviscid flow ... 873

Panters Rodríguez-Bermúdez and Dan Marchesin
Loss of strict hyperbolicity for vertical three-phase flow in porous media ... 881

Rinaldo M. Colombo, Paola Goatin and Massimiliano D. Rosini
On the management of vehicular and pedestrian flows ... 889

Olga Rozanova
Exact solutions with singularities to ideal hydrodynamics of inelastic gases ... 899

Ingenuin Gasser, Martin Rybicki and Winnifried Wollner
Modelling, simulation and optimization of gas dynamics in an exhaust pipe ... 907

Jochen Schütz, Michael Woopen and Georg May
A combined hybridized discontinuous Galerkin / hybrid mixed method for viscous conservation laws ... 915

Gabriella Puppo and Matteo Semplice
Finite volume schemes on 2D non-uniform grids ... 923
ANALYSIS OF OSCILLATIONS AND DEFECT MEASURES IN PLASMA PHYSICS

DONATELLA DONATELLI AND PIERANGELO MARCATI
Department of Information Engineering
Computer Science and Mathematics
University of L’Aquila
67100 L’Aquila, Italy

Abstract. We perform a rigorous analysis of the quasineutral limit for a hydrodynamical model of a viscous plasma represented by the Navier Stokes Poisson system in $3-D$ in the general setting of ill prepared initial data. In general the limit velocity field cannot be expected to satisfy the incompressible Navier Stokes equation, indeed the presence of high frequency oscillations strongly affects the quadratic nonlinearities and we have to take care of self interacting wave packets. We provide a detailed mathematical description of the convergence process by using microlocal defect measures and by developing an explicit correctors analysis. Moreover we identify an explicit pseudo parabolic pde satisfied by the leading correctors terms.

1. Introduction. In this paper we perform a rigorous analysis of the so called quasineutral limit for a hydrodynamical model of a viscous plasma represented by the Navier Stokes Poisson system in $3-D$, namely

$$\partial_t \rho^\lambda + \text{div}(\rho^\lambda u^\lambda) = 0, \quad (1)$$

$$\partial_t (\rho^\lambda u^\lambda) + \text{div}(\rho^\lambda u^\lambda \otimes u^\lambda) + \nabla (\rho^\lambda)^\gamma = \mu \Delta u^\lambda + (\nu + \mu) \nabla \text{div} u^\lambda + \rho^\lambda \nabla V^\lambda, \quad (2)$$

$$\lambda^2 \Delta V^\lambda = \rho^\lambda - 1. \quad (3)$$

Let us denote by $x \in \mathbb{R}^3$, $t \geq 0$, the space and time variable, $\rho(x,t)$ the negative charge density, $m(x,t) = \rho(x,t) u(x,t)$ the current density, $u(x,t)$ the velocity field, $V(x,t)$ the electrostatic potential, $\mu, \nu$ the shear viscosity and bulk viscosity respectively. The parameter $\lambda$ is the so called Debye length (up to a constant factor).

The system (1)-(3) is endowed with the following initial conditions,

$$\rho^{\lambda}_{t=0} = \rho^\lambda_0 \geq 0, \quad V^{\lambda}_{t=0} = V^\lambda_0, \quad (\text{ID})$$

$$\rho^\lambda u^\lambda |_{t=0} = m^\lambda, \quad m^\lambda = 0 \text{ on } \{ x \in \mathbb{R}^3 \mid \rho^\lambda_0(x) = 0 \},$$

$$\int_{\mathbb{R}^3} \left( \pi^\lambda |_{t=0} + \frac{|m^\lambda|^2}{2 \rho^\lambda_0} + \lambda^2 |V^\lambda_0|^2 \right) dx \leq C_0,$$

where

$$\pi^\lambda = \frac{(\rho^\lambda)^\gamma - 1 - \gamma (\rho^\lambda - 1)}{(\gamma - 1)}$$

2000 Mathematics Subject Classification. Primary: 35L65; Secondary: 35L40, 76R50.

Key words and phrases. Compressible and incompressible Navier Stokes equation, energy estimates, hyperbolic equations, Klein-Gordon equations, rescaling, dispersive estimates, microlocal defect measures.
and to simplify our notation from now we will take $\mu = \nu = 1$. In terms of physical variables the Debye length can be expressed as

$$\lambda = \frac{\lambda_D}{L}, \quad \lambda_D = \sqrt{\frac{\varepsilon_0 k_B T}{e^2 n_0}},$$

(4)

where $L$ is the macroscopic length scale, $\varepsilon_0$ is the vacuum permittivity, $k_B$ the Boltzmann constant, $T$ the average plasma temperature, $e$ the absolute electron charge and $n_0$ the average plasma density. In many cases the Debye length is very small compared to the macroscopic length $\lambda_D \ll L$ and so it makes sense to consider the quasineutral limit $\lambda \to 0$ of the system (1)-(3). In this situation the particle density is constrained to be close to the background density (equal to one in our case) of the oppositely charged particle. The limit $\lambda \to 0$ is called the quasineutral limit since the charge density almost vanishes identically. The velocity of the fluid then evolves according to the incompressible Navier Stokes flow. This type of limit has been studied by many authors. In the case of Euler Poisson system by Cordier and Grenier [4], Grenier [17], Cordier, Degond, Markowich and Schmeiser [3], Loeper [24], Peng, Wang and Yong [25], in the case of the Navier Stokes Poisson system by Wang [32] and Jiang and Wang [19], Feireisl and collaborators [6] and in the contest of a combined quasineutral and relaxation time limit by Gasser and Marcati in [11, 12, 13]. This paper is still a mathematical theoretical approach to this complicate physical problem which however removes many regularity and smallness assumptions of various papers in the literature see for instance Wang [32] and Jiang and Wang [19]. In fact Wang [32] studied the quasineutral limit for the smooth solution with well-prepared initial data. Wang and Jiang [19] studied the combined quasineutral and inviscid limit of the compressible Navier- Stokes-Poisson system for weak solution and obtained the convergence of Navier- Stokes-Poisson system to the incompressible Euler equations with general initial data. Moreover in [19] the vanishing of viscosity coefficient was required in order to take the quasineutral limit and no convergence rate was derived therein. The authors in [7] investigated the quasineutral limit of the isentropic Navier-Stokes-Poisson system in the whole space and obtained the convergence of weak solution of the Navier-Stokes-Poisson system to the weak solution of the incompressible Navier- Stokes equations by means of dispersive estimates of Strichartz’s type under the assumption that the Mach number is related to the Debye length. Ju, Li and Wang [20] studied the quasineutral limit of the isentropic Navier-Stokes-Poisson system both in the whole space and in the torus without the restriction on viscous coefficient with well prepared initial data. However there is no analysis for the quasineutral limit for the Navier Stokes Poisson system in the context of weak solutions and in the framework of general ill prepared initial data. The common feature of this kind of limits in the ill prepared data framework is the high plasma oscillations, namely the presence of high frequency time oscillations along the acoustic waves. In these phenomena there are different behaviors of the various vector fields acting in our system. Particularly relevant us to understand the relationship between high frequency interacting waves, dispersive behavior and the different role of time and space oscillations. In our analysis the velocity fields both disperse and oscillates however the dispersion behavior dominates on the high frequency time oscillations and Strichartz estimates are sufficient to pass into the limit of the convective term. The presence of quadratic terms on the electric field (e.g. $\rho^2 \nabla V^2$) cannot be analyzed in the same way since the dispersive behavior no longer dominates on time high frequency wave packets.
In the general case these quadratic terms will not vanish in the limit as $\lambda \to 0$, unless we have well prepared initial data.

2. Notations. If $F, G$ are functions we denote by $F \lesssim G$ the fact that there exists $c \in \mathbb{R}$ such that $F \leq G$. Then, we will denote by: $\mathcal{D}(\mathbb{R}^d \times \mathbb{R}_+)$ the space of test function $C_0^\infty(\mathbb{R}^d \times \mathbb{R}_+)$, by $\mathcal{D}'(\mathbb{R}^d \times \mathbb{R}_+)$ the space of Schwartz distributions and $\langle \cdot , \cdot \rangle$ the duality bracket between $\mathcal{D}'$ and $\mathcal{D}$. $W^{k,p}(\mathbb{R}^d) = (I-\Delta)^{-\frac{k}{2}}L^p(\mathbb{R}^d)$ and $H^k(\mathbb{R}^d) = W^{k,2}(\mathbb{R}^d)$ the nonhomogeneous Sobolev spaces, for any $1 \leq p \leq \infty$ and $k \in \mathbb{R}$. $W^{k,p}(\mathbb{R}^d) = (-\Delta)^{-\frac{k}{2}}L^p(\mathbb{R}^d)$ and $H^k(\mathbb{R}^d) = W^{k,2}(\mathbb{R}^d)$ denote the homogeneous Sobolev spaces. The notations $L_+^{k,2}$ and $L_+^{k,p}$ will abbreviate respectively the spaces $L^p([0,T];L^k(\mathbb{R}^d))$, and $L^p([0,T];W^{k,q}(\mathbb{R}^d))$, $L_+^p(\mathbb{R}^d)$ the Orlicz space defined as follows $L_+^p(\mathbb{R}^d) = \{ f \in L^1_{{\rm loc}}(\mathbb{R}^d) \mid |f|_{\chi\{x\mid |f|\leq \frac{1}{2}\}} \in L_+^2(\mathbb{R}^d), \ |f|_{\chi\{x\mid |f|> \frac{1}{2}\}} \in L^p(\mathbb{R}^d) \}$, see [1], [23] for more details. $\mathcal{L}(\mathbb{R}^d)$ the space of bounded operators, $\mathcal{K}(\mathbb{R}^d)$ the space of compact operators, if $X, Y$ are Banach spaces, $\mathcal{L}(X,Y)$ is the space of bounded operators, $Q$ and $P$ respectively the Leray’s projectors $Q$ on the space of gradients vector fields and $P$ on the space of divergence - free vector fields. Namely $Q = \nabla \Delta^{-1} \div, \quad P = I - Q$. It is well known that $Q$ and $P$ can be expressed in terms of Riesz multipliers, therefore they are bounded linear operators on every $W^{k,p}$ $(1 < p < \infty)$ space (see [27]).

Next we recall the basic notations concerning pseudo-differential operators and symbols to be used later on. We refer to [30], [31] for details. Assuming $\rho, \delta \in [0,1]$, $m \in \mathbb{R}$, we denote $S^m_{\rho,\delta}$ the set of $C^\infty$ symbols satisfying

$$\left| D^\rho_x D^\delta_\xi p(x,\xi) \right| \leq C_{\alpha,\beta} (\xi)^{m-|\alpha|+|\beta|}$$

for all $\alpha, \beta$, where $\langle \xi \rangle = (1 + |\xi|^2)^{1/2}$. In such case we say that the associated operator denoted by $OP(p(x,\xi))$ is given by

$$P(x,D)f(x) = \int p(x,\xi)\mathcal{F}f(\xi)e^{ix\xi}d\xi := OP(p(x,\xi))$$

(where $\mathcal{F}f(\xi) = (2\pi)^{-n} \int f(x)e^{-ix\xi}dx$ denotes the Fourier transform of the function $f$) belongs to $OPS^{m}_{\rho,\delta}$. If there are smooth symbols $p_{m-j}$, homogeneous in $\xi$ of degree $m - j$ for $|\xi| \geq 1$, i.e. $p_{m-j}(x,\xi) = r^{m-j}p_{m-j}(x,\xi)$ for $r > 0, |\xi| \geq 1$, and if

$$p(x,\xi) \sim \sum_{j \geq 0} p_{m-j}(x,\xi)$$

in the sense that

$$p(x,\xi) - \sum_{j \geq 0} p_{m-j}(x,\xi) \in S^{m-N}_{1,0}$$

for all $N$, then we say $p(x,\xi) \in S^m$ and $P(x,D)$ is polyhomogenous of order $m$. If $\Omega$ is an open set in $\mathbb{R}^d$, we denote by $\psi^m_{\rm comp}(\Omega, \mathcal{L}(H))$, respectively, $\psi^m_{\rm comp}(\Omega, \mathcal{K}(H))$ the space of polihomogenous pseudo-differential operators of order $m$ on $\Omega$, with values in $\mathcal{L}(H)$, respectively $\mathcal{K}(H)$ whose kernel is compactly supported in $\Omega \times \Omega$, moreover we recall that if $P \in \psi^m_{\rm comp}(\Omega, \mathcal{L}(H))$, then its symbol $p(x,\xi)$ is a linear application from $\psi^m_{\rm comp}(\Omega, \mathcal{L}(H))$ to $C_0^\infty(S^*\Omega, \mathcal{L}(H))$, where $S^*\Omega = S^{d-1}\times \Omega$.

Following P. Gérard we say that $\mu$ is the microlocal defect measure (or following L. Tartar the $H$-measure) for a bounded sequence $w_k$ in $L^2$ if for any
A ∈ ψ^0_{comp}(ω, K(H)) one has (up to subsequences)
\[ \lim_{k \to \infty} (A(w_k - w), (w_k - w)) = \int_{S^1} \text{tr}(a(x, \xi)\mu(dx\xi)). \]
where \( A = OP(a(x, \xi)) \).

3. Main results. The existence of global weak solutions for fixed \( \lambda > 0 \) for the system (1)-(3) with initial data satisfying (ID), has been proved in the case of a bounded domain in [5] and in the case of the whole domain in [9] and [10]. So we state here the main result concerning the convergence of solutions of the system (1)-(3) in the quasineutral regime.

**Theorem 3.1.** Let \((ρ^λ, u^λ, V^λ)\) be a sequence of weak solutions in \( \mathbb{R}^3 \) of the system (1)-(3), assume that the initial data satisfy (ID). Then

(i) \( ρ^λ \rightharpoonup 1 \) weakly in \( L^\infty([0, T]; L^2(\mathbb{R}^3)) \).

(ii) There exists \( u \in L^\infty([0, T]; L^2(\mathbb{R}^3)) \rightarrow L^2([0, T]; H^1(\mathbb{R}^3)) \) such that
\[
\lim_{\lambda \to 0} \int_{\mathbb{R}^3} |a(x, \xi)|^2 \mu(dx\xi) = 0.
\]

(iii) The gradient component \( Qu^λ \) of the vector field \( u^λ \) satisfies
\[
Qu^λ \rightharpoonup 0 \quad \text{strongly in } L^2([0, T]; L^p(\mathbb{R}^3)), \quad \text{for any } p \in [4, 6).
\]

(iv) The divergence free component \( Pu^λ \) of the vector field \( u^λ \) satisfies
\[
Pu^λ \rightharpoonup Pu = u \quad \text{strongly in } L^2([0, T]; L^2_{loc}(\mathbb{R}^3)).
\]

(v) There exist correctors \( E^+, E^- \) in \( L^\infty((0, T), L^2(\mathbb{R}^3)) \) and a positive microlocal defect measure \( ν^E \) on \( \mathbb{R}^3 \times S^2 \) depending measurably on \( t \), associated to the electric field \( E^λ = \nabla V^λ \), such that for all pseudodifferential operators \( A \in \psi^0_{comp}(\mathbb{R}^3, K(\mathbb{R}^3)) \), of symbol \( a(x, \xi) \) and \( \phi \in \mathcal{D}(0, t) \) one has
\[
\lim_{\lambda \to 0} \int dt\phi(t)λ^2(AE^λ, E^λ) = \int dt\phi(t)(AE^+, E^+) + \int dt\phi(t)(AE^−, E^−) + \int dt\phi(t)\int_{\mathbb{R}^3} \text{tr}\left( a(x, \xi)\frac{ξ \otimes ξ}{|ξ|^2} \right) dν^E.
\]

(vi) \( u = Pu \) satisfies the following equation
\[
P(∂_t u - Δ u + (u \cdot \nabla)u - \text{div}(E^+ \otimes E^+ + E^- \otimes E^-) - \text{div}(ν^E \frac{ξ \otimes ξ}{|ξ|^2})) = 0,
\]
in \( \mathcal{D}'([0, T] \times \mathbb{R}^3) \).

We omit here the proof of the theorem but we short describe the mains steps, for the details we refer to [8]. The main ingredients of our approach to this limiting process can be summarized as follows.

**Step 1:** Obtain a priori estimates independent of \( \lambda \), namely standard energy bounds and dispersive estimates on the density fluctuation. The main idea here is based on the observation that the density fluctuation \( ρ^λ - 1 \) satisfies a Klein-Gordon equation, so the acoustic waves analysis for the Navier Stokes Poisson system (1)-(3) follows by reading the system as a dispersive equation and we get uniform estimates in \( \lambda \) by the use of the \( L^p \)-type estimates due to Strichartz [15, 21, 28]. The particular type of Strichartz estimates for the
Klein Gordon equation that we use here can be recovered from the seminal paper by Strichartz [28] (where he studied the homogenous equation) and by Duhamel’s principle.

**Step 2:** In the previous step we get sufficient bounds in order to study the limiting behaviour of the velocity vector field. Therefore we can analyze separately the limiting behaviour of the divergence free part and the gradient part of \( u^{\lambda} \). Accordingly we obtain the strong convergence of the velocity field.

**Step 3:** The next stumbling block is to get enough compactness for the electric field in order to pass into the limit in the quadratic term \( \lambda \nabla V^{\lambda} \otimes \lambda \nabla V^{\lambda} \).

Since \( \lambda \nabla V^{\lambda} \) is bounded in \( L^{\infty}_{t}L^{2}_{x} \) we can define microlocal defect measure \( \nu^{E} \) introduced by P. Gérard in [14] and by L. Tartar (H-measure) in [29] with correctors \( E^{+} \) and \( E^{-} \) to handle time oscillations at frequency \( 1/\lambda \). An analogous use of the P. Gérard and L. Tartar ideas can be found in Y. Brenier and E. Grenier [2] and E. Grenier [16], regarding the Vlasov Poisson system.

4. Remark on microlocal defect measure and correctors. In the previous theorem we constructed a defect measure \( \nu^{E} \) and the correctors \( E^{\pm} \). They correspond to the physical phenomenon of the high frequency plasma oscillation. Notice that the correctors \( E^{\pm} \) remain important as \( \lambda \to 0 \) and are not vanishing, in fact we don’t have initial layer but on the contrary the effect of ill prepared initial data appears through \( E^{\pm} \) and remains important for all times. The construction of the defect measure \( \nu^{E} \) will be done by using the theory developed by P. Gérard in [14] and L. Tartar in [29]. Now two questions naturally arise: in which case the microlocal defect measure and the correctors are vanishing? Who are the correctors, are there any cases in which we can have an explicit construction of them? We will answer to these issues in the next two subsections.

4.1. Well prepared data: No time and space oscillations. We consider here the case of well prepared data, namely data that satisfy the following conditions,

\[
\int_{\mathbb{R}^{3}} |\rho^{\lambda}_{0} - 1|^{2} \chi_{(|\rho^{\lambda}_{0} - 1| \leq \delta)} dx + \int_{\mathbb{R}^{3}} |\rho^{\lambda}_{0} - 1|^{\gamma} \chi_{(|\rho^{\lambda}_{0} - 1| > \delta)} dx \leq M \lambda
\]

\[
\text{div } u_{0} = 0
\]

\[
\| \sqrt{\rho^{\lambda}_{0}} u^{\lambda}_{0} - u_{0} \|_{L^{2}} \leq M \lambda \quad \| \lambda \nabla V^{\lambda}_{0} \|_{L^{2}} \leq M \lambda,
\]

where \( u^{\lambda}_{0} = u^{\lambda}(x, 0) \) and \( u_{0} \) is the initial data of the limiting system. Then by standard energy estimates on the solutions of the system (1)-(3) we obtain the following bounds for \( \rho^{\lambda}, V^{\lambda}, u^{\lambda} \)

\[
\int_{\mathbb{R}^{3}} |\rho^{\lambda} - 1|^{2} \chi_{(|\rho^{\lambda} - 1| \leq \delta)} dx + \int_{\mathbb{R}^{3}} |\rho^{\lambda} - 1|^{\gamma} \chi_{(|\rho^{\lambda} - 1| > \delta)} dx \leq M \lambda
\]

\[
\| \sqrt{\rho^{\lambda}} u^{\lambda} - u \|_{L^{\infty}(0, T; L^{2})} + \| \lambda \nabla V^{\lambda} \|_{L^{\infty}(0, T; L^{2})} \leq M(T) \lambda^{\min\{1/2, 1/\gamma\}}
\]

By the estimates (7) and (8) it follows easily that there are no space and time oscillations so there are no microlocal defect measures and correctors and this entails compactness and strong convergence for the density, the velocity and electric field.
4.2. Regular initial data: No space oscillations, only time oscillations. An other relevant situation is if we consider regular initial data, namely if we work in the framework of $H^s$, $s \geq d/2$ solutions for the system (1)-(2) (see [18], [22]). In that case we can have an explicit construction of the correctors and we can prove that they satisfy a pseudoparabolic differential equation. We are able to prove the following theorem.

**Theorem 4.1.** Let be $(\rho^\lambda, u^\lambda, V^\lambda)$ be a sequence of the Navier Stokes Poisson system, satisfying for $s \geq 4$

$$\|\rho^\lambda - 1\|_{L^\infty(0,T;H^s(\mathbb{R}^3))} \leq C \quad \|\lambda E^\lambda\|_{L^\infty(0,T;H^s(\mathbb{R}^3))} \leq C$$

then, for all $s' < s - 2$

$$u^\lambda - \frac{1}{i} e^{-it/\lambda} E^+ - \frac{1}{i} e^{it/\lambda} E^- \rightarrow v \quad \text{strongly in } C^0(0,T,H^{s'-1}_{\text{loc}}(\mathbb{R}^3)).$$

$$\lambda(E^\lambda - e^{-it/\lambda} E^+ - e^{it/\lambda} E^-) \rightarrow 0 \quad \text{strongly in } C^0(0,T,H^{s'-1}_{\text{loc}}(\mathbb{R}^3)).$$

and $E^\pm$ satisfy

$$\partial_t E^\pm - \Delta E^\pm + Q \text{div}(v \otimes E^\pm) = 0, \quad PE^\pm = 0.$$  

From the conditions (9) we can deduce that we have strong compactness in time so the microlocal defect measures vanishes while time oscillations of order 1/\lambda still persists and the correctors $E^\pm$ are not vanishing. For rigorous details of the proof we refer to [8], for simplicity we sketch a formal argument.

**Proof.** First we rewrite (2) in terms of $E^\lambda$, namely

$$\lambda^2 \partial_t E^\lambda + E^\lambda = \text{div} \Delta^{-1} \nabla \text{div} \left( \rho^\lambda u^\lambda \otimes u^\lambda + (\rho^\lambda)^{\gamma \lambda} I - \lambda^2 E^\lambda \otimes E^\lambda \right) + \frac{\lambda^2}{2} \text{div} \left( |E^\lambda|^2 I \right) - 2 \nabla \text{div} u^\lambda,$$

In order to take into account the time oscillations we decompose the electric field and the velocity in the following way:

$$E^\lambda \sim \frac{E^+}{\lambda} e^{it/\lambda} + \frac{E^-}{\lambda} e^{-it/\lambda} \quad u^\lambda \sim v - ie^{it/\lambda} E^+ - ie^{-it/\lambda} E^-,$$

where $v$ is a divergence free vector field. Now if we replace this decomposition in the equation (13) we get

$$2i \partial_t E^+ e^{it/\lambda} - 2i \partial_t E^- e^{-it/\lambda} + \lambda \partial_t E^+ e^{it/\lambda} + \lambda \partial_t E^- e^{-it/\lambda} =$$

$$\text{div} \left[ \Delta^{-1} \nabla \text{div} \left( \rho^\lambda (v - ie^{it/\lambda} E^+ - ie^{-it/\lambda} E^-) \otimes (v - ie^{it/\lambda} E^+ - ie^{-it/\lambda} E^-) \right) + (\rho^\lambda)^{\gamma \lambda} I - \lambda^2 E^\lambda \otimes E^\lambda \right] + \frac{\lambda^2}{2} |E^\lambda|^2 I - 2 \nabla \text{div}(v - ie^{it/\lambda} E^+ - ie^{-it/\lambda} E^-).$$

If we consider only the oscillatory part of the electric field we get

$$\partial_t E^\pm + \text{div}(v \otimes E^\pm) - \nabla \text{div} E^\pm = 0, \quad PE^\pm = 0.$$

or equivalently

$$\partial_t E^\pm - \Delta E^\pm + Q \text{div}(v \otimes E^\pm) = 0.$$

$\square$
REFERENCES


E-mail address: donatell@univaq.it
E-mail address: marcati@univaq.it
ASYMPTOTIC STABILITY OF KINETIC PLASMAS FOR GENERAL COLLISION POTENTIALS

Renjun Duan
Department of Mathematics, The Chinese University of Hong Kong
Shatin, Hong Kong, China

Abstract. The motion of a fully ionized plasma consisting of electrons and ions is generally governed by the Vlasov-Maxwell-Landau system, where both the classical Landau collision with Coulomb potentials and the self-consistent Lorentz force coupled with the Maxwell equations are taken into account. This note presents a new weighted energy method, recently developed in [7, 8], to prove the asymptotic stability of the Cauchy problem on the corresponding system under the non-relativistic approximation when initial data is smooth and close to steady states. Some relative results on other collisional plasma models are also mentioned.

1. Background. Plasma is recognized as the fourth state of matter besides the solid, liquid and gas states, cf. [24, 31]. 99.9% of the universe exists in a plasma state. For a fully ionized plasma, it is a gas consisting of charged particles, e.g. electrons and ions. The motion of plasmas strongly responds to the self-consistent electromagnetic field \((E, B)\) through the Maxwell equations

\[
\frac{1}{c} \partial_t E - \nabla \times B = -\frac{4\pi}{c} J, \quad \frac{1}{c} \partial_t B + \nabla \times E = 0, \tag{1}
\]

\[
\nabla \cdot E = 4\pi \rho, \quad \nabla \cdot B = 0. \tag{2}
\]

Here \(\rho\) is the total charge, \(J\) the electric current, and \(c\) the light speed. Plasma physics involves the physics of classical mechanics, electromagnetism, and non relativistic statistical mechanics. Challenge in the study of plasma physics lies in the long-range coulomb interaction [40].

Let us recall three types of models for describing the motion of plasmas [24]. First of all, microscopic motion equations governing the time evolution of all plasma particles \(1 \leq i \leq N_0\) of \(s\)-species with position and velocity \((x_i(t), \xi_i(t))\) at time \(t\) take the form of

\[
\frac{dx_i}{dt} = \xi_i, \quad \frac{d\xi_i}{dt} = \frac{q_s}{m_s} [E(t, x_i) + \frac{\xi_i}{c} \times B(t, x_i)],
\]

with \((E, B)\) coupled with the Maxwell equations (1)-(2), and \(\rho, J\) given by

\[
\rho = \sum_s q_s \int_{\mathbb{R}^3} N_s(t, x, \xi) d\xi, \quad J = \sum_s q_s \int_{\mathbb{R}^3} \xi N_s(t, x, \xi) d\xi,
\]

2000 Mathematics Subject Classification. Primary: 82D10, 76P05; Secondary: 35B40.

Key words and phrases. Plasma, Boltzmann collision, Landau collision, asymptotic stability, time-weighted energy method.

The author is supported by the GRF (Project No. 400511) from RGC of Hong Kong.
with
\[ N_s(t, x, \xi) = \sum_{i=1}^{N_s} \delta(x - x_i(t)) \delta(\xi - \xi_i(t)). \]

Here and in the sequel \(m_s\) and \(q_s\) denote the mass and charge for each particle of \(s\)-species.

The goal for the kinetic plasma is to determine the velocity distribution functions \(f_s = f_s(t, x, \xi) \geq 0\), with \(s = i\) and \(e\), of gas particles with position \(x \in \mathbb{R}^3\) and velocity \(\xi \in \mathbb{R}^3\) at time \(t \geq 0\), which satisfy
\[ \partial_t f_s + \xi \cdot \nabla_x f_s + \frac{q_s}{m_s} (E + \frac{\xi}{c} \times B) \cdot \nabla_\xi f_s = \left( \frac{\partial f_s}{\partial t} \right)_c := \sum_{s'} Q(f_s, f_{s'}). \tag{3} \]

with \((E, B)\) coupled with the Maxwell equations (1)-(2), and \(\rho, J\) given by
\[ \rho = \sum_s q_s \int_{\mathbb{R}^3} f_s(t, x, \xi) \, d\xi, \quad J = \sum_s q_s \int_{\mathbb{R}^3} \xi f_s(t, x, \xi) \, d\xi. \tag{4} \]

Depending on the collisional mechanism [26, 38], the system is called the Vlasov-Maxwell-Boltzmann system or Vlasov-Maxwell-Landau system. In the case of the Boltzmann collisions, the collision term, cf. [1, 15], is given by
\[ Q(f_1, f_2)(\xi) = \int_{\mathbb{R}^3 \times S^2} B(\xi - \xi_*, \omega) \{ f_1(\xi') f_2(\xi'_*) - f_1(\xi) f_2(\xi_*) \} \, d\xi_* d\omega, \]

where the pre-collisional and post-collisional velocity pairs \((\xi, \xi_*)\), \((\xi', \xi'_*)\) satisfy
\[ \xi' = \xi - \frac{2m_2}{m_1 + m_2} (\xi - \xi_*) \cdot \omega, \quad \xi'_* = \xi_* + \frac{2m_1}{m_1 + m_2} (\xi - \xi_*) \cdot \omega, \]

and the Boltzmann collision kernel takes the form of
\[ B(\xi - \xi_*, \omega) = \Phi(|\xi - \xi_*|) b(\frac{\xi - \xi_*}{|\xi - \xi_*|} \cdot \omega), \tag{5} \]
\[ \Phi(|z|) \sim |z|^\gamma (-3 < \gamma \leq 1), \quad \sin \theta b(\cos \theta) \sim \frac{1}{\theta^{1+2s}} (0 < s < 1). \tag{6} \]

An example comes from the inverse power law \(U(r) = r^{-(p-1)} (p > 2)\) for which one has \(\gamma = \frac{p-5}{p-1}, s = \frac{1}{p-1}\). In general, it is quite extensive to use the Grad’s angular cutoff assumption, that is to replace \(b(\cdot)\) in (5) by \(\tilde{b}(\cdot)\) such that
\[ \int_0^{\pi/2} \sin \theta \, \tilde{b}(\cos \theta) \, d\theta < \infty. \]

In the case of the Landau collisions [30], the collision term is given by
\[ Q(f_1, f_2) = \frac{1}{m_1} \nabla_\xi \cdot \int_{\mathbb{R}^3} \Phi(\xi - \xi') \{ \frac{1}{m_1} f_1(\xi) \nabla_\xi f_2(\xi') - \frac{1}{m_2} f_2(\xi) \nabla_\xi f_1(\xi') \} \, d\xi'. \tag{7} \]

with the Landau collision kernel
\[ \Phi(z) = |z|^{\gamma+2} (I - \frac{z \otimes z}{|z|^2}) (\gamma \geq -3). \tag{8} \]

The Coulomb potential corresponds to the case \(\gamma = -3\). There is a formal grazing limit to derive the Landau operator from the Boltzmann operator, cf. [38].
The third type of plasma equations are at the fluid level, with unknowns as number density $n_s$ and velocity $v_s$ of $s$-species, satisfying
\[
\begin{aligned}
\frac{\partial n_s}{\partial t} + \nabla \cdot (n_s v_s) &= 0, \\
m_s n_s (\partial_t v_s + v_s \cdot \nabla v_s) + \nabla P_s = q_s n_s (E + \frac{v_s}{c} \times B) \\
&\quad + \sum_{s'} \nu_{ss'} \frac{m_s m_{s'} n_s n_{s'}}{m_s n_s + m_{s'} n_{s'}} (v_s - v_{s'}),
\end{aligned}
\]
with $(E, B)$ coupled with the Maxwell equations (1)-(2), and $\rho, J$ given by
\[
\rho = \sum_s q_s n_s, \quad J = \sum_s q_s n_s v_s.
\]
They are called the Euler-Maxwell systems with/without collisions, cf. [3, 13]. Here, $P_s$ is the pressure depending only on $n_s$, and $\nu_{ss'}$ is the relaxation frequency between two species.

Due to the collision and particle-field interactive mechanism, a plasma usually relaxes to different kinds of profiles such as equilibrium states, periodic states, and wave patterns. Both physically and mathematically, it is an important task to understand the stability of those profiles. Stability theory in plasma physics addresses the following three questions: (i) Can the initial (even small) perturbation of a given profile imply the global-in-time existence of solutions? (ii) Will the solution converge to it? How fast for the rate of convergence? and (iii) If unstable, how to characterize the growth modes? In this paper we will not discuss the unstable case for which readers can refer to [24] for physical discussions, and to [25] and references therein for mathematical studies.

Finally we remark that problems without collisions are quite different due to the nonlinear effect and structure, see the recent survey [40], the gravitational Vlasov-Poisson system [25], Landau-damping [29], and the Euler-Maxwell system [13]. In addition, in the presence of the electromagnetic field, one may need to take into account some additional physical effects, such as the relativistic effect [20], quantum effect obeying the Fermi-Dirac/Bose-Einstein statistics [12], and the dynamical screening effect due to the plasma polarization [30, 31].

2. Known results on perturbation theory of kinetic plasmas. From now on we only focus on the perturbation theory for the plasma motion equations at the kinetic level; see the pioneering work [36] on the pure Boltzmann equation. The so-called Boltzmann’s H-theorem for the spatially homogeneous Boltzmann equation tells that the physical entropy $\int (-f \log f) \, d\xi$ must be nondecreasing and attain its maximum at global Maxwellians, which interprets the second law of thermodynamics. Thus, it is expected that the global Maxwellian is time asymptotically stable. However, in the spatially non-homogenous case, the total entropy $\int \int (-f \log f) \, dx d\xi$ vanishes at any local Maxwellian
\[
M_{[\rho(t,x),u(t,x),T(t,x)]}(\xi) = \frac{\rho(t,x)}{[2\pi T(t,x)]^{3/2}} e^{-\frac{|\xi - u(t,x)|^2}{2T(t,x)}},
\]
with macroscopic quantities depending on time-space variables. In other words, this kind of degeneration of H-theorem implies that the linearized operator $L$ of the nonlinear collision term $Q(f, f)$ around the global Maxwellian is degenerate in a velocity space of finite dimensions. If the spatial domain is a torus, the dissipation of those macroscopic degenerate quantities can be recovered by the Poincare inequality.
whereas in the case of the whole space, it is not obvious to figure out their dissipative feature. The macro-micro decomposition of the solution in the form of \( f = Pf + (I - P)f \) is an extensively used tool to deal with the issue, cf. [17, 27, 28]. In fact, through the Grad’s moment method [14], it can be verified that the macro part \( Pf \) is coupled with the micro dissipative \( (I - P)f \) in a way that system of thirteen moment equations with higher moments ignored is of the first-order hyperbolic conservation laws with symmetric relaxations which satisfy the Shizuta-Kawashima stability condition [23]. Therefore, the macro dissipation could be induced by the micro one [10, 11]. Here, we point out that it is more difficult to treat the soft (large-velocity degenerate) collision kernel compared to the hard one with the positive lower bound, cf. [8, 33].

Another difficulty related to the dissipative structure of the coupled system for kinetic plasmas lies in the degeneration of Maxwell equations which preserve the total energy if it was assumed that the electric and magnetic fields propagate in vacuum without particles, cf. [11]. However, the coupling can produce the weak dissipation of \((E, B)\). Indeed, if system of the thirteen moment equations coupled with the Maxwell equations is again investigated, it becomes still of the first-order hyperbolic conservation laws but with anti-symmetric relaxations [35]. Though the classical Shizuta-Kawashima condition fails in this case, one can develop some new idea to analyze the complex dissipative structure.

After one has a better understanding of the dissipative structure of the coupled system, in order to consider the nonlinear stability, the key point is to control all the nonlinear terms in terms of the obtained dissipation in a proper way to have the global-in-time a priori estimates. Unfortunately, a few additional difficulties appear in this step; see [18] for a complete presentation. We here mention only one of them saying that for the non-hard sphere model, particularly for the soft potential case, the weak micro dissipation makes it impossible to control those nonlinear terms with the velocity growth effect in a direct way. For example, let us consider the Cauchy problem on the Vlasov-Maxwell-Boltzmann system near Maxwellians. The global existence was firstly proved in [19] on torus for the hard-sphere model. It was later extent in [32] to the whole space. The large-time behaviour of solutions was given in [22] for the torus and [11] for the whole space. However, all the corresponding results have remained unknown in the case of the non-hard-sphere model for the Vlasov-Maxwell-Boltzmann system either with or without angular cutoff.

The same situation occurs to the Vlasov-Maxwell-Landau system in the case \( \gamma + 2 < 1 \) for the Landau kernel (8). Recently, the time-asymptotic stability of solutions near Maxwellians was proved in [18] for the relatively simple Vlasov-Poisson-Landau system with the Coulomb potential on the torus, where the proof is based on the choice of an algebraic velocity weight which captures the anisotropic dissipative property of the linearized Landau operator, and also on the potential form of the self-consistent force. The result of [18] was immediately generalized to the whole space in [34] by following the same approach, [39] by using the pure energy method without the analysis of the linearized solution operator, and also [9] by applying a new weight function that we will mention later on. Though these results exist, the problem on the Vlasov-Maxwell-Landau system for the Coulomb potential still stands because of the weak dissipation and the non-potential force form.
3. Our main results. Our investigation on the asymptotic stability of solutions for general collision potentials starts from the study of the Vlasov-Poisson-Boltzmann system in the whole space \([7, 8]\), where \(-2 \leq \gamma \leq 1\) for the angular cutoff Boltzmann collision kernel is supposed. The main idea is to introduce into the velocity weight function a time-velocity-dependent exponential factor \(\exp\{\lambda(\xi)^q/(1+t)^\vartheta\}\) with \(\lambda > 0\), \(1 \leq q \leq 2\) and \(\vartheta > 0\) properly chosen. In fact, an extra weak dissipative mechanism can be recovered from the time derivative of this exponential factor in the way of

\[
\partial_t \exp\{\lambda(\xi)^q/(1+t)^\vartheta\} = -\lambda \vartheta (1+t)^{-\vartheta -1} \langle \xi \rangle^q \exp\{\lambda(\xi)^q/(1+t)^\vartheta\}.
\]  

(9)

Thus one can deal with the estimate on a triple inner product even with the velocity-growth rate at most \(\|\xi\|^2\) as long as one of those triple terms decays faster than \((1+t)^{\gamma}\). The approach of \([7, 8]\) is easily applied to the Vlasov-Poisson-Landau system for all the Landau collision kernels \((8)\), as mentioned before, without using the diffusion dissipative property of the Landau operator, cf. \([9]\).

However, due to the technique of the approach, the problem on the angular cutoff Vlasov-Poisson-Boltzmann system still remains unknown for the strongly soft potentials \(-3 < \gamma < -2\). But, this has been solved in the case of the angular non cutoff Vlasov-Poisson-Boltzmann system in \([5]\) by letting the angular singularity index \(s\) in the Boltzmann collision kernel \((5)-(6)\) sufficiently close to 1, where the approach of the proof is inspired by \([18]\) to use the fractional order diffusion dissipation property of the linearized angular non cutoff Boltzmann operator along with the functional framework founded by \([15]\). Very recently, the result of \([5]\) is further generalized in \([6]\) to the Vlasov-Maxwell-Boltzmann system with angular non-cutoff in the whole space, where \(\gamma\) and \(s\) in \((6)\) is assumed to satisfy \(\max\{-3, -\frac{3}{2} - 2s\} < \gamma < -2s\), \(1/2 \leq s < 1\). Here, different from \([18]\) for the case of the potential force, it seems necessary in \([6]\) to incorporate the exponential weight into the energy functional to treat the non-potential Lorentz force. Moreover, the functional framework of \([1]\) on the basis of the triple norm and commutators is employed in \([6]\).

The extension of results in \([9]\) to the complex Vlasov-Maxwell-Landau system with Coulomb potentials is not immediate, due to the regularity-loss property of the whole system \([11]\). To resolve the difficulty, a new time-weighted energy method is developed in \([4]\) by using an idea from \([21]\) to deduce the time-weighted estimate with the time rate of negative power. In fact, in general the highest order derivative of the electric field \(E\) is not included in the usual energy dissipation rate so that it is a problem to make weighted estimates on the linear term involving the electric field. Time weight of negative power is designed to produce the dissipation of the highest order derivative of \(E\) up to a time-decay factor.

It should be pointed out that solution spaces that are used in \([4, 6]\) involve very high smoothness and velocity integrability, and also initial data belongs to \(L^1\) in spatial variable. It is still an interesting problem to remove these restrictions for proving the global stability of the Cauchy problem under small perturbations.

4. Vlasov-Maxwell-Landau system. In what follows we present in a precise way the main results obtained in \([4]\). Consider the Cauchy problem of the Vlasov-Maxwell-Landau system (3), (4), (7), (8) and (1), (2) of two-species particles in the whole space. For brevity, write \(f_\pm = f_\pm, e\) and set all the physical parameters to be unit. By expecting \(f_\pm \to M = M_{[1,0,1]}(\xi)\) and \([E, B] \to 0\), we set the perturbation \(u = M^{-1/2}(f - M)\) with \(u = [u_+, u_-], f = [f_+, f_-]\). Let \(q_0 = \text{diag}(1, -1)\). The
Theorem 4.1. Let \([u, E, B]\) be the solution to the Cauchy problem of the linearized homogeneous Vlasov-Maxwell-Landau system (10) with \(S = 0\). Assume \(-3 \leq \gamma < -2\) in (8). Define the velocity weight function \(w = w(\xi) = \langle \xi \rangle^{-\frac{\gamma+2}{2}}\). Then, for \(\ell \geq 0\) and \(m \geq 0\) with \(m = |\alpha|\),
\[
\|w^\ell \partial^\alpha u\| + \|\partial^\alpha (E, B)\| \leq C(1 + t)^{-\sigma_m}(\|w^{\ell + \ell_{low}} u_0\|_{Z_1} + \|\langle E_0, B_0\rangle\|_{L^1_x}) + (1 + t)^{-j}(\|w^{\ell + \ell_{high}} \nabla^{j+1}_x \partial^\alpha u_0\| + \|\nabla^{j+1}_x \langle E_0, B_0\rangle\|),
\]
(11)
where \(\partial^\alpha = \partial_{x}^\alpha, Z_1 = L^2_x(L^1_x), \) and
\[
\sigma_m = \frac{3}{4} + \frac{m}{2}, \ell_{low}^* > 2\sigma_m, \ell_{high}^* > 0, \quad 0 \leq j < \ell_{high}^*.
\]

We remark that the estimate (11) exhibits the compensation property of two kinds of degenerations of the system. The additional velocity weights \(w^{\ell_{low}}\) and \(w^{\ell_{high}}\) are used to compensate the large-velocity degeneration of the collision kernel for soft potentials, and the additional space differentiation \(\nabla^{j+1}_x\) to obtain the time-decay rate \((1 + t)^{-j}\) is used to compensate the regularity-loss of the system. Notice that regularity-loss is not a deficiency of our approach but an essential feature; it can be seen from the eigenvalue analysis of the damped Euler-Maxwell system (cf. [3]) from which one of eigenvalues behaves at the large-frequency domain as
\[
\lambda(ik) \sim -\frac{1}{|k|^2} \pm ik |\langle k\rangle \to \infty|.
\]

Before stating the stability result for the nonlinear problem, we list some notations as follows. Fix \(\theta\) with \(0 < \theta < 1/4\). Define
\[
w_{\tau, \lambda} = w_{\tau, \lambda}(t, \xi) = \langle \xi \rangle^{(\gamma+2)\tau} \exp\left\{\frac{\lambda}{(1 + t)^\theta} (\xi)^2\right\},
\]
where \(\tau \in \mathbb{R}\) and \(\lambda \geq 0\). For \(u = u(t, x, \xi)\), define
\[
|u(x)|^2_{\tau, \lambda} = \int_{\mathbb{R}^3} w_{\tau, \lambda}^2(t, \xi) |u|^2 d\xi,
\]
\[
|u(x)|^2_{D, \tau, \lambda} = \sum_{i,j=1}^3 \int_{\mathbb{R}^3} w_{\tau, \lambda}^2(t, \xi) \left\{\sigma^{ij} \partial_i u \partial_j u + \sigma^{ij} \frac{\xi_i \xi_j}{2} |u|^2\right\} d\xi,
\]
and their further space integrations are denoted by \(\|u\|_{\tau, \lambda}^2\) and \(\|u\|_{D, \tau, \lambda}^2\), respectively.

Here \(\partial_i = \partial_{\xi_i}\), and \(\sigma^{ij} = \phi^{ij} * M\) is the Landau collision frequency. Write \(\partial^\beta = \partial_{x}^\beta \partial_{\xi}^\beta\). The energy functional and its dissipation rate are given by
\[
E_{N, \ell, \lambda}(t) \sim \sum_{|\alpha| + |\beta| \leq N} \|\partial^\beta u(t)\|_{\beta |\beta| - \ell, \lambda}^2 + \|\langle E, B\rangle\|_{H^N}^2.
\]
and
\[ D_{N,t,\lambda}(t) = \sum_{|\alpha|+|\beta| \leq N} \| \partial_\beta^\alpha \{ I - P \} u(t) \|^2_{D_{N,|\beta|-\ell,\lambda}} + \sum_{|\alpha| \leq N-1} \| \nabla_x \partial_\alpha^\alpha (a_{\pm}, b, c) \|^2 \\
+ \| a_+ - a_- \|^2 + \| E \|^2_{H^{N-1}} + \| \nabla_x B \|^2_{H^{N-2}} \\
+ \frac{\lambda}{(1 + t)^{1 + \sigma}} \sum_{|\alpha|+|\beta| \leq N} \| \langle \xi \rangle \partial_\beta^\alpha \{ I - P \} u(t) \|^2_{|\beta|-\ell,\lambda}. \]

The last term of \( D_{N,t,\lambda}(t) \) is generated by (9) as mentioned before. Let \( N_0 \) and \( \ell_0 \) be fixed properly large with \( \ell_0 - 3N_0 \) being also properly large, and let \( \lambda_0 > 0 \) and \( \epsilon_0 > 0 \) be fixed properly small. Set \( N_1 = \frac{3}{2}N_0, \ell_1 = \frac{1}{2}\ell_0 \). Define
\[ X(t) = \sup_{0 \leq s \leq t} \{ \mathcal{E}_{N_{t}}(s) + (1 + s)^{\frac{2}{3}} \mathcal{E}_{N_{t-2}}(s) \} \]
\[ + \sup_{0 \leq s \leq t} \{ (1 + s)^{-1 + \sigma} \mathcal{E}_{N_{t},\ell_{1},\lambda_{0}}(s) + \mathcal{E}_{N_{t-1},\ell_{1},\lambda_{0}}(s) + (1 + s)^{\frac{2}{3}} \mathcal{E}_{N_{t-3},\ell_{1-1},\lambda_{0}}(s) \} \]
\[ + \sup_{0 \leq s \leq t} \{ \mathcal{E}_{N_{0},\ell_{0},\lambda_{0}}(s) + (1 + s)^{2} \mathcal{E}_{N_{0-1},\ell_{0},\lambda_{0}}(s) + (1 + s)^{2(1+\sigma)} \| \nabla_x (E, B)(s) \|^2_{H^{N_{0}-1}} \}. \]

Theorem 4.2. Assume \(-3 \leq \gamma < -2\). Take \( \vartheta, N_0, \ell_0, N_1, \ell_1, \lambda_0 > 0, \epsilon_0 > 0 \) as mentioned before. Fix \( \ell_2 > \frac{3}{4} + \frac{N_0}{2} \). Let \( u_0 = [u_{0+}, u_{0-}] \) satisfy \( f_{\pm,0,0}(x, \xi) = M(\xi) + M^{1/2}(\xi) u_{0+}(x, \xi) \geq 0 \). Then, there are \( Y_0 \) and \( X(t) \), where \( Y_0 \) depends only on initial data \( (u_0, E_0, B_0) \), such that the Cauchy problem of the Vlasov-Maxwell-Landau system (10) admits a unique global solution satisfying \( f_{\pm}(t, x, \xi) = M(\xi) + M^{1/2}(\xi) u_{\pm}(t, x, \xi) \geq 0 \) and \( X(t) \leq CY^2 \), for all time \( t \geq 0 \).

The complete proof of Theorem 4.1 and Theorem 4.2 can be found in [4]. We finally point out, inspired by [16] and [37], that it would be an interesting and challenging problem to generalize all the current results in the case of either the whole space or torus to the case of the bounded domain for the kinetic plasma under the influence of the self-consistent force.

REFERENCES


NUMERICAL DISCRETIZATIONS FOR SHALLOW WATER EQUATIONS WITH SOURCE TERMS ON UNSTRUCTURED MESHES

ARNAUD DURAN AND FABIEN MARCHE
Institut de Mathématiques et de Modélisation de Montpellier (I3M) Université Montpellier 2, CC 051, 34090 Montpellier, France

CHRISTOPHE BERTHON AND RODOLPHE TURPAULT
Laboratoire de Mathématiques Jean Leray Université de Nantes, France

Abstract. In the following lines we introduce two frictional schemes for the discretization of the 2D Shallow Water system, on unstructured meshes. The starting point consists in writing both of them as convex combinations of 1D schemes. Then, we propose to include the resistance effects proceeding to a slight adaptation of the gathered convex components, using the frictional approach recently developed in [2]. This method turns out to provide an excellent behavior for vanishing water heights, and does not require a modification of the CFL. Numerical experiments will be performed in order to assess the capacity of the two schemes in dealing with wetting and drying, complex geometry and topography.

Introduction. In this work we consider discretizations of the 2D Non linear Shallow Water equations (NSW). As the name suggests, the NSW model consists in an hyperbolic set of non linear equations, and is used to describe the motion of shallow flows, as flood waves, flooding and drying, dam breaks, or more generally any kind of hydrodynamic processes near coasts or in bed rivers. Denoting $h$ the water height, $q = (q_x, q_y)$ the discharge and $z$ a parametrization of the bed slope, we write the set of NSW equations under its conservative form:

$$w_t + \nabla \mathcal{H}(w) = -B(w, z) - F(w),$$

with

$$w = \begin{pmatrix} h \\ q_x \\ q_y \end{pmatrix}, \quad \mathcal{H}(w) = \begin{pmatrix} \frac{q_x^2}{h} + \frac{1}{2} gh^2 & \frac{q_y}{h} \\ \frac{q_x r_y}{h} & \frac{q_y^2}{h} + \frac{1}{2} gh^2 \end{pmatrix},$$

and the following expressions for bathymetry and friction :

$$B(w, z) = \begin{pmatrix} 0 \\ gh z_x \\ gh z_y \end{pmatrix}, \quad F(w) = \begin{pmatrix} 0 \\ \kappa ||q||_h q_x \\ \kappa ||q||_h q_y \end{pmatrix}.$$

2000 Mathematics Subject Classification. Primary: 58F15, 58F17; Secondary: 53C35.

Key words and phrases. Nonlinear Shallow Water, Finite Volume, Well Balanced schemes, friction, unstructured mesh.
The behavior of the resistance $F$ is partly governed by the positive constants $\kappa$ and $\gamma$, and may be watched closely as the water elevation tends to zero. A suitable construction of a numerical scheme for the discretization of the Shallow Water system is generally subject to some numerical requirements, discussed in the major part of the published studies on the theme. The first obligation is the well known well balancing property, that is the preservation of the following motionless steady state: $\eta = cte$ and $u = 0$. The second unavoidable property we want to achieve is the robustness, namely the ability to preserve the convex set of admissible states $\Omega = \{(h,q_x,q_y), h \geq 0\}$. Finally, we direct our discussion on the capacity in handling dry cells or low values of the water height, which turns out to be a major concern when the physical model deals with a friction term under the form (3). One of the main objectives of this work consists in proposing numerical schemes offering a stable and accurate treatment of such configurations.

The paper is split up in four parts: in Sections 1 and 2 we propose a detailed presentation of the first scheme, called BF-scheme in the sequel. Then, we briefly give some words regarding a second order extension, and discuss about a twin numerical approach. The last part is devoted to numerical tests expected to validate and compare the abilities of the two schemes resulting from these developments.

1. Frictionless scheme. The following approach has been developed by Berthon, first in a 1D framework, leading to a well balanced and robust scheme, with an excellent behavior in the neighborhood of dry cells. Before caring about the inclusion of friction, we propose to set up the main results and notations related to its extension in the context of unstructured triangulations (see also [3]).

The choice of $W = t(\eta, \eta u, \eta v)$ instead of $w$ as conservative vector variable during the evaluation of the numerical fluxes leads to the following alternative form of homogeneous NSW:

$$w_t + \nabla \cdot \left( \xi \mathcal{H}(W) - \begin{pmatrix} 0 & 0 \\ \frac{gh_z}{2} & 0 \end{pmatrix} \right) = 0,$$  \hfill (4)

where we introduced the new variable $\xi = h/\eta$. In this first part we seek a discretization of (4) in an unstructured context. To this purpose, let’s introduce some notations concerning the geometry: we consider the dual mesh related to a triangulation $\mathcal{T}$ of the computational domain. The nodes of $\mathcal{T}$ will be denoted $(S_i)$, the elements $(T_i)$, and the vertex centered cells $(C_i)$. In what follows, $K_i$ will refer to the set of subscripts $j$ for which $C_j$ is adjacent to $C_i$, and $l_{ij}$ the length of the associated interface $\Gamma_{ij}$; a basic example of geometry is available on Fig. 1, where other useful notations are given. Denoting $w^n_i = (h^n_i, q_{ix,i}^n, q_{iy,i}^n)$ a constant interpolation of the solution at the cell $C_i$ at time $t^n$, we chose to evolve this approximate state gathering the components of a 3 points 1D scheme (Fig. 1 - 2):

$$w^n_i + C_i \left[ \sum_{j \in K_i} \left| T_{ij} \right| \right] w_{ij}^{n+1} = 0,$$  \hfill (5)

$$w_{ij}^{n+1} = w^n_i - \frac{\Delta t}{\left| T_{ij} \right|} l_{ij} \left( \phi(w^n_i, w^n_j, \bar{n}_{ij}) - \phi(w^n_i, w^n_i, \bar{n}_{ij}) \right).$$  \hfill (6)

Herein, referring to [3], the numerical fluxes $\phi$ are defined by:

$$\phi(w^n_i, w^n_j, \bar{n}_{ij}) = X_{ij} H(W^n_i, W^n_j, \bar{n}_{ij}),$$  \hfill (7)
where $H = H(U, V, \vec{n})$ denotes a flux function for $\mathcal{H}, \vec{n}$, and:

$$X_{ij} = \begin{cases} \xi_n^i & \text{if } H^h(W^n_i, W^n_j, \vec{n}_{ij}) \geq 0, \\ \xi_n^j & \text{otherwise} \end{cases}.$$  \hspace{1cm} (8)

**Remark 1.** In what follows, the function $H : (U, V, \vec{n}) \mapsto H(U, V, \vec{n})$ would be supposed to verify these two classical properties:

- **Consistency:** $H(W, W, \vec{n}) = \mathcal{H}(W), \vec{n}$ $\forall W \in \Omega.$ \hspace{1cm} (9)
- **Conservativity:** $H(U, V, \vec{n}) = -H(V, U, -\vec{n})$ $\forall U, V \in \Omega.$ \hspace{1cm} (10)

From a practical point of view, computations will be performed using the HLLC or VFRoe - relaxation solver, which associated fluxes are known to fulfill the numerical requirements mentioned above, and handling with contact discontinuities.

Hence, according to (10), the “left” fluxes in (6) just reads: $\phi(w^n_i, w^n_j, \vec{n}_{ij}) = \frac{h_n^i}{n_i} (\mathcal{H}(w^n_i), \vec{n}_{ij}),$ and calling Green’s formula, we obtain:

$$w^{n+1}_i = w^n_i - \frac{\Delta t}{|C_i|} \sum_{j \in K_i} l_{ij} \phi(w^n_i, w^n_j, \vec{n}_{ij}).$$ \hspace{1cm} (11)

We finally complete this discretization of NSW considering the treatment of the bathymetry $\mathcal{B}(w, z)$. In accordance with [3], we denote:

$$H_{ij} = \begin{cases} \eta_n^i & \text{if } H^h(W^n_i, W^n_j, \vec{n}_{ij}) \geq 0, \\ \eta_n^j & \text{otherwise} \end{cases},$$ \hspace{1cm} (12)

and chose the following discretization for the bed slope source term:

$$B_{ij} = \frac{g}{2} \left( H_{ij} \eta_n^i (X_{ij} - \xi_n^i) \vec{n}_{ij} \right),$$ \hspace{1cm} (13)

to write:

$$w^{n+1}_i = w^n_i - \frac{\Delta t}{|C_i|} \sum_{j \in K_i} l_{ij} (\phi(w^n_i, w^n_j, \vec{n}_{ij}) - B_{ij}).$$ \hspace{1cm} (14)
We emphasize that the resulting scheme has been shown to be well balanced and robust under the following CFL condition:

$$\Delta t \max_{i \in Z, j \in K_i} \frac{l_{ij}}{|T_{ij}|} |\lambda^\pm_{ij}| < \frac{1}{2},$$  \hspace{1cm} (15)$$
supplemented by the following CFL restriction:

$$\Delta t \max_{i \in Z, j \in K_i} \left[ \frac{l_{ij}}{|T_{ij}|} \left( \max(0, H^h(W_i, W_j, \bar{a}_{ij})) - \min(0, H^h(W_i, \bar{a}_{ij})) \right) \right] < \eta^i_\alpha,$$  \hspace{1cm} (16)$$
where the velocity waves $\lambda^\pm_{ij}$ in (15) refers to the upper and lower lower extremities of the Riemann solver’s dependency cone. Actually, the proofs are straightforward, getting back to the formulation (5,6) and invoking the properties of the 1D approach.

2. Friction scheme. As shown in [2], in a 1D context, we can achieve a robust numerical treatment of the resistance terms considering a suitable adaptation of the Riemann states involved in the fluxes. More precisely, the proposed scheme is based on the use of the following modified HLL Riemann solver:

$$\tilde{\lambda}_R(x_t, w_L, w_R) = \begin{cases} 
    w_L \text{ if } \frac{x}{\Delta t} \leq a^- \\
    w^* + (1 - \alpha)(w_R^* - w^*) - \frac{h^2}{2\gamma} F(w_L) \text{ if } a^- \leq \frac{x}{\Delta t} \leq 0 \\
    w^* + (1 - \alpha)(w_R^* - w^*) - \frac{h^2}{2\gamma} F(w_R) \text{ if } 0 \leq \frac{x}{\Delta t} \leq a^+ \\
    w_R \text{ if } \frac{x}{\Delta t} \geq a^+
\end{cases}, \hspace{1cm} (17)$$

where $w^*$ stands for the HLL intermediate state:

$$w^*(x_t, w_L, w_R) = \frac{a^+ w_R - a^- w_L}{a^+ - a^-} (H(w_R) - H(w_L)), \hspace{1cm} \text{and (18)}$$

$$w_L(x_t, w_L, w_R) = \left( \frac{h^*(x_t, w_L, w_R)}{(hu)_L} \right), \hspace{1cm} w_R(x_t, w_L, w_R) = \left( \frac{h^*(x_t, w_L, w_R)}{(hu)_R} \right). \hspace{1cm} (19)$$

Herein, $H$ denotes the monodimensional exact flux function of the NSW equations, and $a^\pm$ are the maximum and minimum velocity waves involved in the Riemann solver. Obviously, there is a close dependence between the leading friction parameter $\alpha$ and the friction source term $F(w)$. A discretization will be proposed later on, covering Darcy and Manning friction laws (see Section 4). Actually, such choice is also submitted to some specific requirements, and we refer to [2] for more details.

In a 1D context, using consistency relations and an appropriate analysis of the wave speeds involved in the HLL solver, we obtain the following scheme:

$$h_{i+1}^{n+1} = h_{i}^{n} - \frac{\Delta t}{\Delta x} (H_{i+\frac{1}{2}}^{h} - H_{i-\frac{1}{2}}^{h}),$$

$$\left( hu \right)_{i+1/2}^{n+1} = \left( hu \right)_{i}^{n} - \frac{\Delta t}{\Delta x} \left[ \alpha_{i+\frac{1}{2}} H_{i+\frac{1}{2}}^{hu} - \alpha_{i-\frac{1}{2}} H_{i-\frac{1}{2}}^{hu} - \left( 1 - \alpha_{i-\frac{1}{2}} \right) s_{i-\frac{1}{2}}^{-,hu} + \left( 1 - \alpha_{i+\frac{1}{2}} \right) s_{i+\frac{1}{2}}^{-,hu} \right], \hspace{1cm} (20)$$

where $H_{i+\frac{1}{2}}^{hu}$ stands for an interpolation of $H(w)$ at the node $i+1/2$. The friction terms are defined by:

- $\alpha_{i+1/2} = \frac{(a_{i+1/2}^+ - a_{i+1/2}^-)}{(a_{i+1/2}^+ - a_{i+1/2}^-) + \mathcal{F}_{i+1/2}^{hu} \Delta x}$.
- $\mathcal{F}_{i+1/2}^{hu} = \frac{\nu q_{i+1/2}}{h_{i+1/2}}, \hspace{0.5cm} h_{i+1/2} = \left( hu \right)_{i+1/2}^{n+1} = \left( hu \right)_{i}^{n} + \frac{(hu)_i^n - \mathcal{F}_{i+1/2}^{hu} \frac{\Delta t}{\Delta x}}{2}, \hspace{0.5cm} q_{i+1/2} = \frac{\|a^+_i\| + \|a^-_i\|}{2}.$
- $s_{i+1/2}^{-,hu} = \min(0, a_{i+1/2}^-)/(hu)_{i+1/2}^{n} - \min(0, a_{i+1/2}^+)/(hu)_{i+1/2}^{n} - H^{hu}(w_i^n).$
• $s_{i-1/2}^+ = \max(0, a_{i-1/2}^- ((hu)_i^n - (hu)_{i-1}^n))$ and $s_{i-1/2}^- = \max(0, a_{i-1/2}^+ ((hu)_i^n + H^{nu}(w_i^n))$.

Remark 2. In the pioneer paper, the evaluation of the numerical fluxes $H_{i+1/2}$ is performed with the HLL relaxation solver. In practise, computations with any other consistent and conservative flux function in the (1D) sense of (9) and (10) (as those provided by the HLLC and VFRoe solver) remains valid.

Now, in order to perform an extension to the unstructured case, we chose to extrapolate formulas (20), defining the convex components as :

$$
\tilde{h}_{ij}^n = h_i^n - \frac{\Delta t}{\Delta x_i^2} \left( \phi^h(w_i^n, w_j^n, \bar{n}_{ij}) - \phi^h(w_i^n, w_i^n, \bar{n}_{ij}) \right),
$$

$$
(hu)_{ij}^n = (hu)_i^n - \frac{\Delta t}{\Delta x_i} \left[ \alpha_{ij} \phi^{hu}(w_i^n, w_j^n, \bar{n}_{ij}) - \alpha_{ii} \phi^{hu}(w_i^n, w_i^n, \bar{n}_{ij}) \right] - \left(1 - \alpha_{ii}\right) s_{ii}^{+,hu} + \left(1 - \alpha_{ij}\right) s_{ij}^{-,hu},
$$

$$
(hv)_{ij}^n = (hv)_i^n - \frac{\Delta t}{\Delta x_i} \left[ \alpha_{ij} \phi^{hv}(w_i^n, w_j^n, \bar{n}_{ij}) - \alpha_{ii} \phi^{hv}(w_i^n, w_i^n, \bar{n}_{ij}) \right] - \left(1 - \alpha_{ii}\right) s_{ii}^{+,hv} + \left(1 - \alpha_{ij}\right) s_{ij}^{-,hv}.
$$

These 1D updates are then gathered using (5) ; introducing the terms (13) related to the topography, we obtain the following scheme :

$$
\begin{align*}
\Delta x_i \left( \phi^h(w_i^n, w_j^n, \bar{n}_{ij}) - \phi^h(w_i^n, w_i^n, \bar{n}_{ij}) \right),
\end{align*}
$$

$$
\begin{align*}
\Delta x_i \left[ \alpha_{ij} \phi^{hu}(w_i^n, w_j^n, \bar{n}_{ij}) - \alpha_{ii} \phi^{hu}(w_i^n, w_i^n, \bar{n}_{ij}) \right] - \left(1 - \alpha_{ii}\right) s_{ii}^{+,hu} + \left(1 - \alpha_{ij}\right) s_{ij}^{-,hu},
\end{align*}
$$

$$
\begin{align*}
\Delta x_i \left[ \alpha_{ij} \phi^{hv}(w_i^n, w_j^n, \bar{n}_{ij}) - \alpha_{ii} \phi^{hv}(w_i^n, w_i^n, \bar{n}_{ij}) \right] - \left(1 - \alpha_{ii}\right) s_{ii}^{+,hv} + \left(1 - \alpha_{ij}\right) s_{ij}^{-,hv}.
\end{align*}
$$

Remark 3. When $\kappa = 0$, the leading friction parameter $\alpha_{ij}$ is obviously equal to 1, and we recover the condensed formulation established in the frictionless case (14).

Remark 4. One of the main features of the current frictional scheme is that we do not have to use a more restrictive CFL to ensure the robustness property. Indeed, we can easily verify that formulas (22-23) and (14) are equivalent for the water height.

3. Additional investigations. We first discuss about a second order extension of the BF-scheme (22-24). In order to improve the accuracy, we chose to adopt the reconstruction technique introduced in [4], following closely the lines of [7]. Avoiding redundancies, we do not give details on this method here, and just propose a brief overview of the final results. Actually, we simply reach an analogous formulation of (22-24), where the original values have been replaced by the second order ones. After straightforward computations, the proof of the well balancing property appears to be the same as in the first order case, and the preservation of the water height positivity is ensured provided a slight adaptation of the additional CFL restriction (16), involving the second order values $W_{ij}$ and $W_{ji}$ :

$$
\Delta t \max_{i \in I, j \in K} \left[ \frac{l_{ij}}{|T_{ij}|} \left( \max(0, H^h(W_{ij}, W_{ji}, \bar{n}_{ij})) - \min(0, H^h(W_{ij}, \bar{n}_{ij})) \right) \right] < \eta_i^h. 
$$
In a second hand, we point out that the main strategy of this work consists in the creation of a 2D discretization of NSW using a 3 point monodimensional scheme (formulation 5 - 6). In reality, this technique is not restricted to the only use of Berthon’s scheme, and can also be run considering another 1D approach. Indeed, we mention that the lines of Sections 1 and 2 can be followed taking the 1D pre balanced scheme developed by Liang et al. in [9] as starting point. Considering the similarities with the two previous sections, we neither go further in the developments. We just note that the resulting scheme is very close than the one described in [7] for the frictionless case, with the advantage of ensuring the preservation of the admissible states under a less restrictive CFL. As for the BF case, the inclusion of the resistance and the second order extension previously discussed can also be performed, preserving the well balancing property ; however we refer to [7] for the proof of the robustness at order 2, which requires additional work. We mention that deeper investigations are currently in progress on the subject [8] ; the scheme obtained in this way will be called the LF-scheme in the next section.


4.1. Dam break with friction. We consider a classical one - dimensional dam break problem with friction on a flat bottom. One can build an approximation of the exact solution by splitting the domain in two regions behind the wave front location ; in the first area, frictional effects can be neglected ; the exact solution is provided considering an ideal fluid flow model for Shallow Water equations. In the wave tip region bed, the variation of the wave speed is small, and the resistance phenomena becomes dominant under acceleration and inerty effects. Physical and mathematical issues of such construction are detailed in [6]. In a 2D framework, we will consider a rectangular channel with dimensions \([-10, 10] \times [0, 4]\). The initial water depth is set to 1m at the left of the dam \((x \leq 0)\), and 0 elsewhere. For this test, we consider a Darcy friction law : \( F(w) = (0, (f/8) \|u\|u, (f/8) \|u\|v) \), and recover the formulation (3) with \( \gamma = 2 \) and \( \kappa = f/8 \). The Darcy coefficient is set to \( f = 0.05 \), and we use a 8241 nodes splitted cartesian grid, with \( \Delta x = \Delta y = 0.1m \) (Fig. 3). Fig. 4 shows some snaphots of the water depth along the middle section, until \( t = 2.5s \). We can observe a good agreement with the analytical solution ; the wave front location seems to be accurately computed, which tends to validate the ability of the BF and LF schemes to deal with wetting and drying. We can also point out the good concordance with the predictions provided by the work of Céa et al in [5].

4.2. Moving boundary over a parabolic bottom. The following test is also issuing from a 1D benchmark simulation for NSW models (see [9]), derived itself from a test case proposed by Thacker in 1981 (see [10]). Initially, the domain is assumed to be a 8640m length channel, and we set its width to 1000m for a 2D extrapolation. Computations are run on a regular triangulation (Fig. 3) with \( \Delta x = 27m \), \( \Delta y = 25m \), and the parametrization of the bathymetry is given by:
This time, we consider a linear friction term: $F(w) = \tau(0, \kappa hu, \kappa hv)$. As we do not enter in the formalism (3), the simulation will be run with the following adaptation of the friction parameter (see (20) and below): $F_{i+1/2} = \kappa$. This test involves an oscillatory planar flow, which motion is described by the following equations:

$$
z(x, y) = h_0\left(\left(\frac{x}{a}\right)^2 - 1\right).
$$

$$
h(t, x, y) = \frac{a^2 B e^{-\kappa t}}{8g^2 h_0}\left(-s\kappa \sin(2st) + \left(\frac{\kappa^2}{4} - s^2\right)\cos(2st)\right)
- \frac{B^2 e^{-\kappa t}}{g} \frac{e^{-\kappa t/2}}{g} \left(Bs \cos(st) + \frac{\kappa B}{2} \sin(st)\right) x,
$$

$$
u(t, x, y) = Be^{-\kappa t/2} \sin(st),
$$

where $s = \sqrt{(8 gh_0/a^2 - \kappa^2)/2}$. For the simulation, the governing parameter for the bed friction term is fixed to $\kappa = 0.001$, and we also set $h_0 = 10m$, $a = 3000m$ and $B = 5m/s$. The periodic motion being submitted to non negligible frictional effects, the amplitude of the oscillations is expected to decrease with respect to time, until the apparition of a motionless steady state in the center of the channel. Such phenomena should be highlighted by the time history of the predicted shoreline location, which is compared to the exact one (see Fig. 5) until time $T = 10000s$:

$$
x = \frac{a^2 e^{-\kappa t/2}}{2gh_0} \left(-Bs \cos(st) - \frac{\kappa B}{2} \sin(st)\right) \pm a.
$$

We now discuss about the behavior of the first and second order schemes toward the numerical error. To do so, we chose to gradually increase the refinement of the mesh, taking respectively $\Delta x = 192, 96, 48$ and $24m$ as discretization steps. On Fig. 6 we can see the evolution of the $L^1$-error with respect to $\Delta x$, in a log-log scale. Concerning the error on the water height (left), we reach a convergence rate of 0.95 for both first order LF and BF schemes, and 1.53 for the second order reconstruction. Similar results are obtained for the normal discharge (right). As in the first test, the two schemes seems to offer similar results in terms of accuracy, although the second order reconstruction seems to be slightly more efficient for the LF-scheme in this case.

4.3. Malpasset Dam Break. We finally validate the capacity of the two approaches in handling with complex geometry and wetting and drying in a real-world 2D case, proceeding to the Malpasset Dam Break benchmark test. This dam was located in the Deparment of Var, in the south of France, and collapsed in December.
1959. Thanks to the bathymetric and geometric data available, we run the computations in a vertex centered mesh issuing from a triangulation of 13541 nodes, in which we enforce the value of $z$. Among the many possibilities for the definition of the bed friction, we chose to consider the Manning-Chezy formulation:

$$\mathcal{F}(w) = \begin{pmatrix} \frac{0}{n^2 \frac{|q|}{h^{1/3}}} q_x \\ n^2 \frac{|q|}{h^{1/3}} q_y \end{pmatrix},$$

and set up the Manning coefficient to $n = 0.03$. We follow the time evolution of the location of the flood wave, for which we have a reference at several gauges along the river, provided by a physical experimentation on a reduced model. The predicted arrival time of the water front at gauges 6 to 14 are plotted on Fig. 7, and compared to the reference. Numerical and experimental data only differs from 20s to the maximum : reaching such level of accuracy in this difficult context highlights the real efficiency of the BF and LF schemes.
Figure 7. Malpasset Dam Break: comparison of predicted wave front propagation time with experimental data at gauges 6 to 14.

REFERENCES


E-mail address: Arnaud.Duran@math.univ-montp2.fr
SCHEMES WITH WELL-CONTROLLED DISSIPATION (WCD)
FOR SCALAR CONSERVATION LAWS WITH
PSEUDO-PARABOLIC REGULARIZATION

JAN ERNEST
Seminar of Applied Mathematics
ETH Zürich, Rämistrasse 101
8092 Zürich, Switzerland

Abstract. We study non-classical solutions of hyperbolic conservation laws arising as limits of small-scale mechanisms in the underlying physical model. Here we focus on scalar conservation laws with pseudo-parabolic regularizations. For this type of problems, standard numerical schemes may not converge to the correct (physically relevant) limit solutions. We explain this lack of convergence by means of the equivalent equation of the schemes and design a new class of schemes called schemes with well-controlled dissipation (WCD). They are constructed in such a way that the leading-order terms of the equivalent equation exactly reproduce the regularizing effects in the model while the influence of the higher-order residual terms is bounded appropriately. This ensures that the non-classical effects that are observed in the numerical solutions are mainly based on the physically relevant small-scale effects. The newly developed numerical schemes are validated for the example of the Buckley-Leverett equation with pseudo-parabolic regularization.

1. Introduction. We consider standard scalar conservation laws of the form

\[ u_t + f(u)_x = 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+. \]

(1)

It is well-known that solutions of (1) may develop discontinuities in finite time even for smooth initial data and therefore have to be considered in a distributional sense. However, these weak solutions need not be unique and suitable admissibility criteria have to be defined. One possibility is to impose pointwise conditions such as the Lax or Oleinik entropy conditions at shocks. A more general framework is based on a convex entropy function \( S(u) \) and a corresponding entropy flux \( Q(u) \) satisfying \( Q'(u) = S'(u)f'(u) \). A solution of (1) is termed entropy solution if it satisfies

\[ S(u)_t + Q(u)_x \leq 0, \]

(2)
in a weak (distributional) sense. A solution satisfying (2) and the pointwise entropy conditions due to Lax and Oleinik is called classical. Standard numerical schemes such as Lax-Friedrichs or Rusanov scheme may be applied to approximate them.

For many physical applications however, the solutions depend on small-scale effects such as capillarity or viscosity. These effects can be modeled by including high-order regularization terms on the right-hand side of (1), i.e.

\[ u_t^\varepsilon + f(u^\varepsilon)_x = R^\varepsilon(u^\varepsilon), \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+. \]

(3)

2000 Mathematics Subject Classification. Primary: 65M06, 65M12.
Key words and phrases. Non-classical shocks, Conservation laws, Pseudo-parabolic approximations, WCD schemes, Equivalent equation.

551
Typical examples are the diffusive-dispersive regularization \( R^\varepsilon(u^\varepsilon) = \varepsilon u_{xx}^\varepsilon + \delta \varepsilon^2 u_{xxx}^\varepsilon \) and the pseudo-parabolic regularization \( R^\varepsilon(u^\varepsilon) = \varepsilon u_{xx}^\varepsilon + \delta \varepsilon^2 u_{xxt}^\varepsilon \). We are interested in the limit of solutions as \( \varepsilon \to 0 \) for a fixed \( \delta \), as it is well-known that for this type of models classical and non-classical solutions may arise for \( \varepsilon \to 0 \) depending on the value of \( \delta \) [4, 5, 7, 9]. By non-classical solutions we mean weak solutions satisfying (2) for a single entropy function \( \eta \), but not the pointwise entropy conditions.

A standard example is given by the scalar cubic conservation law with diffusive-dispersive regularization

\[
 u_t + (u^3)_x = \varepsilon u_{xx} + \delta \varepsilon^2 u_{xxx}.
\]

Jacobs, McKinney and Shearer characterize the limiting solutions for \( \varepsilon \to 0 \) for all values of \( \delta \) and prove that the solutions are classical for \( \delta \leq 0 \) and non-classical for \( \delta > 0 \) [5].

The fact that the limiting solution depends on the value of \( \delta \) is challenging from a numerical perspective and standard schemes fail to converge to the correct solution. To be able to capture the small-scale terms in (4) correctly, Kissling and Rohde use a heterogeneous multiscale method considering the conservation law as the macroscale formulation and the regularization as the microscale model [6]. LeFloch and Mohammadian study another approach called schemes with controlled dissipation [8]. These schemes consist of high-order finite difference approximations of the flux function and each regularizing term, which are constructed in such a way that the leading-order terms in the equivalent equation match the terms in the physical model. The main drawback of these schemes is the fact that one \textit{a priori} has to select a value of \( \varepsilon \). A fixed \( \varepsilon \) however leads to a failure in approximating shocks of arbitrary strength. This may be explained by the equivalent equation where the residual high-order terms start to interfere with the physically relevant leading-order terms and pollute the numerical approximation.

In order to avoid these unwanted effects Ernest, LeFloch and Mishra design schemes with well-controlled dissipation (WCD) providing a quadratic inequality which adaptively chooses \( \varepsilon \) in such a way that the influence of the high-order terms is bounded by a small constant times the leading-order terms [1]. Hence, no \textit{a priori} choice of \( \varepsilon \) is necessary as the numerical algorithm adaptively sets \( \varepsilon \). This ensures that the numerical effects are mainly based on the correct physical terms at every time step and yields good approximations even for shocks with very large amplitude.

The goal of this paper is to extend the construction of the WCD schemes to the pseudo-parabolic regularization. The main difficulty is to properly control the additional time derivative in the dispersive term. In section 2 we start by deriving the quadratic condition on \( \varepsilon \) for a general scalar conservation law with pseudo-parabolic regularization. In section 3 we test the newly developed WCD scheme for the example of the Buckley-Leverett flux function. Section 4 briefly recapitulates the main advantages of the novel WCD schemes.

2. Mathematical Formulation. Let us consider a general scalar conservation law with pseudo-parabolic regularization

\[
 u_t + f(u)_x = \varepsilon u_{xx} + \delta \varepsilon^2 u_{xxt}, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+.
\]

It can be shown that under some additional conditions the regularization is conservative and entropy dissipative and hence the limiting solution is an entropy solution.
of (1) [9]. Following the derivation of the WCD schemes for the scalar cubic conservation law in [1] we derive the WCD schemes for (5). Consider the discretized domain \{x_i\}_{i=1}^n with uniform mesh width \Delta x and let \( u_i = u(x_i, t) \) and \( f_i = f(u_i(t)) \). Then, we use the following 2\( p \)-th order semi-discrete scheme

\[
\frac{d u_i}{d t} = -\frac{1}{\Delta x} \left( \sum_{j=-p}^{j=p} \alpha_j f_{i+j} \right) + \frac{c}{\Delta x} \left( \sum_{j=-p}^{j=p} \beta_j u_{i+j} \right) + \delta c^2 \left( \sum_{j=-p}^{j=p} \beta_j (u_{i+j})_t \right),
\]

(6)

where \( \alpha_j \) and \( \beta_j \) are given by the relations

\[
\sum_{j=-p}^{j=p} j \alpha_j = 1 \quad \text{and} \quad \sum_{j=-p}^{j=p} j^l \alpha_j = 0 \quad \forall l \neq 1, \quad 0 \leq l \leq 2p
\]

(7)

and

\[
\sum_{j=-p}^{j=p} j^2 \beta_j = 2 \quad \text{and} \quad \sum_{j=-p}^{j=p} j^l \beta_j = 0 \quad \forall l \neq 2, \quad 0 \leq l \leq 2p.
\]

(8)

Note that (7) and (8) define the coefficients \( \alpha_j \) and \( \beta_j \) in such a way that the resulting scheme is of order 2\( p \) and the leading-order terms in the equivalent equation of (6) match the pseudo-parabolic regularization in (5). The key concept in the construction of the WCD schemes is a proper bound on the residual high-order terms in the equivalent equation. Hence, the full equivalent equation including all high-order terms needs to be derived. By Taylor expansions one finds that

\[
\sum_{j=-p}^{j=p} \alpha_j f_{i+j} \approx \Delta x f_x + \sum_{k=2p+1}^{\infty} \frac{\Delta x^k}{k!} \sum_{j=-p}^{j=p} \alpha_j j^k f^{[k]},
\]

and

\[
\sum_{j=-p}^{j=p} \beta_j u_{i+j} \approx \Delta x^2 u_{xx} + \sum_{k=2p}^{\infty} \frac{\Delta x^k}{k!} \sum_{j=-p}^{j=p} \beta_j j^k u^{[k]}.
\]

Introducing the notation

\[
A^p_k = \sum_{j=-p}^{j=p} \alpha_j j^k, \quad B^p_k = \sum_{j=-p}^{j=p} \beta_j j^k,
\]

the full equivalent equation reads

\[
u_t = -f(u)_x + c\Delta x u_{xx} + \delta c^2 \Delta x^2 u_{xxt} - \sum_{k=2p+1}^{\infty} \frac{\Delta x^{k-1}}{k!} A^p_k f^{[k]} + c \sum_{k=2p+1}^{\infty} \frac{\Delta x^{k-1}}{k!} B^p_k u^{[k]} + \delta c^2 \sum_{k=2p+1}^{\infty} \frac{\Delta x^k}{k!} B^p_k u^{[k]}.
\]

(9)

Note that the leading-order terms in (9) match the regularizing terms in (5) exactly by setting \( c = \epsilon / \Delta x \). Let us now assume that there is an isolated shock wave connecting two states \( u_L \) and \( u_R \) and set \( \overline{u} = u_L - u_R > 0 \). The opposite case may be handled similarly. At the discontinuity it formally holds that

\[
\overline{u}^{[k]} = \frac{\overline{u}}{\Delta x^k}, \quad f^{[k]} = \frac{f}{\Delta x^k}, \quad u_t^{[k]} = \frac{u_t}{\Delta x^k}.
\]
The shock speed can be precomputed at the beginning of any simulation. The only variable term is the limiting solution as the parameter still depend on time. As we want to obtain a condition that constrains the scheme, unlike for the diffusive-dispersive regularization, the two estimates (11) and (12) respectively we obtain

\[ |h.o.t.| \leq \left( \hat{S}_p^f |f| \right) + \left| \hat{S}_p^D c |u| \right| + \hat{S}_p^D |\delta c^2| |u| \]

as a consequence, the estimates (11) and (12) simplify to

\[ |h.o.t.| \leq \left( \hat{S}_p^f |f| \right) + \left| \hat{S}_p^D c |u| \right| + \hat{S}_p^D |\delta c^2| |u| \]

where

\[ \hat{S}_p^f = \sum_{k=2p+1}^{\infty} \sum_{j=-p}^{p} \frac{\alpha_j j^k}{k!} \]

\[ \hat{S}_p^D = \sum_{k=2p+1}^{\infty} \sum_{j=-p}^{p} \frac{\beta_j j^k}{k!} \]

Unlike for the diffusive-dispersive regularization, the two estimates (11) and (12) still depend on time. As we want to obtain a condition that constrains the scheme parameter in each time step, the time dependence has to be removed. Noting that the limiting solution as \( \varepsilon \to 0 \) satisfies the conservation law (1) we assume that

\[ ||u_t|| \approx |f_x| = |f| \Delta x. \]

As a consequence, the estimates (11) and (12) simplify to

\[ |h.o.t.| \leq \left( \hat{S}_p^f |f| \sigma + \hat{S}_p^D c + \hat{S}_p^D |\delta c^2| \right) \frac{|u|}{\Delta x} \]

\[ |L.o.t.| \geq \left| -\sigma + c + |\delta c^2| \right| \frac{|u|}{\Delta x} \]

where \( \sigma \) denotes the shock speed which is defined as

\[ \sigma = \frac{|f|}{|u|}. \]

The two estimates on the leading-order and higher-order terms imply that the balance (10) is fulfilled if the scheme parameter \( c \) satisfies

\[ \left( |\delta c^2| - \frac{\hat{S}_p^D |\delta c|}{\tau} \right)^2 + \left( 1 - \frac{\hat{S}_p^D}{\tau} \right) c - \left( 1 + \frac{\hat{S}_p^f}{\tau} \right) \sigma > 0, \]

at every time step. This quadratic condition will be referred to as WCD condition. Note that the coefficients \( \hat{S}_p^D \) and \( \hat{S}_p^f \) only depend on the order \( p \) of the scheme and can be precomputed at the beginning of any simulation. The only variable term is the shock speed \( \sigma \), which needs to be updated at every time step. In order that the
scheme parameter $c$ is real we require the coefficient of the quadratic term to be positive. This can be enforced by requiring

$$1 - \frac{\hat{S}_D}{\tau} > 0 \quad \text{or equivalently} \quad \hat{S}_P < \tau, \quad (14)$$

which we call order condition. As $\hat{S}_P$ is decreasing in the order $p$, a smaller value of the balancing factor $\tau$ implies a higher order of scheme. This leads to a trade-off between restricting the influence of the higher-order terms in the equivalent equation and using a reasonably small order of scheme.

3. Numerical Examples. The regularized Buckley-Leverett equation is given by

$$u_t + f(u)x = \varepsilon u_{xx} + \delta \varepsilon^2 u_{xxt}$$

with flux function

$$f(u) = \frac{u^2}{u^2 + \kappa (1 - u)^2}, \quad \text{if } 0 \leq u \leq 1,$$

$$f(u) = 1 \quad \text{for } u > 1 \quad \text{and} \quad f(u) = 0 \quad \text{for } u < 0.$$ 

Let $\kappa > 0$ be fixed. The function $u$ represents the water saturation, $f$ the water fractional flow function and $\kappa$ the viscosity ratio of water to oil. We refer to [9] for a detailed motivation of the regularizing terms. We are interested in limiting solutions for $\varepsilon \to 0$ for a fixed $\delta > 0$. Van Duijn, Peletier and Pop show that non-classical shock waves appear for choices of $\delta$ above a certain threshold and they provide numerical experiments. We consider them as reference solutions for testing the WCD schemes [9]. Let us therefore study Riemann problems of the form

$$u(x,t) = \begin{cases} u_L & x \leq x_0 \\ 0 & x > x_0 \end{cases} \quad (15)$$

for varying left states $u_L$. The flux parameter $\kappa$ is set to $\kappa = 2$. The integration of the semi-discrete formulation (6) in time is done using a strong stability-preserving third-order Runge Kutta scheme described in [2]. Let $u_L = 0.75$ for $x \leq x_0 = 0.4$. We compare the classical ($\delta = 0$) and a non-classical ($\delta = 5$) solution at a final time $T = 0.8$. We use a 4th-order WCD scheme for $\tau = 0.1$ with a CFL number of 0.45. Note that for this choice of $\tau$ and order of scheme, the order condition (14) is satisfied. In (6) we use the scheme parameter $c = c_{WCD}$, where $c_{WCD}$ denotes the parameter computed by the WCD condition (13). Numerical approximations on two different meshes are displayed in Figure 1. As expected, the limiting solution for $\delta = 0$ is classical whereas for $\delta = 5$ we see an additional non-classical state. The WCD scheme approximates both cases well and converges as the mesh is refined. As a second example we let $u_L = 0.9$ for $x \leq x_0 = 0.4$ and compute the limiting solution for $\delta = 5$ at a final time $T = 0.6$ on a grid with 4000 mesh points. The numerical approximation with the 4th-order WCD scheme for $\tau = 0.1$ is shown in Figure 2. The WCD scheme resolves the shocks sharply and there are no oscillations.

We now study the Riemann problem (15) for the cases $u_L = 0.68$ and $u_L = 0.55$ computed up to a final time $T = 0.6$ on a grid with 2000 mesh points. The authors show that $u_L = 0.68$ is the smallest possible choice of left state such that the non-classical intermediate state still is clearly visible. For $u_L = 0.55$ they only observe oscillations before the solution jumps from $u_L$ to $u_R$, but no non-classical state. The numerical results computed with the WCD scheme completely agree with this observation as shown in Figure 3.
Figure 1. Classical (left) and non-classical (right) solution of the regularized Buckley-Leverett equation for the Riemann problem (15) with $u_L = 0.75$ and $x_0 = 0.4$.

Figure 2. Non-classical limiting solution of the regularized Buckley-Leverett equation for the Riemann problem (15) with $u_L = 0.9$ and $x_0 = 0.4$.

As a final numerical example we consider a Riemann problem of the form

$$u(x,t) = \begin{cases} 
0.25 & x \leq 0.4 \\
0.85 & x > 0.4,
\end{cases} \quad (16)$$

which is solved for $\delta = 0$ and $\delta = 5$ on a grid with 2000 mesh points up to a final time $T = 0.6$. Again, for $\delta = 0$ the regularization only consists of a diffusive term and therefore only classical solutions are expected. For $\delta = 5$ the additional dispersive term is expected to cause a non-classical state. Figure 4 shows that the WCD scheme complies with the theoretical expectations.

4. Conclusion. We have extended the construction of the schemes with well-controlled dissipation (WCD) proposed by Ernest, LeFloch and Mishra [1] to the case of a scalar conservation law with pseudo-parabolic regularization. The WCD
schemes are constructed in such a way that the leading-order terms of the equivalent equation exactly match the pseudo-parabolic regularization terms in the physical model, whereas the higher-order residual terms are bounded appropriately. The additional time derivative in the dispersive term is controlled through a limit argument.

The main advantage of the WCD schemes lies in the fact that one does not need to \textit{a priori} fix the scheme parameter $\varepsilon$ as the WCD condition (13) provides a useful criterion that adaptively sets a reasonable value of $\varepsilon$ at every time step. This ensures that the non-classical effects are mainly based on the correct small-scale mechanisms.

Finally, the WCD schemes have been tested by numerical examples for the regularized Buckley-Leverett equation, which serves as a prototype of regularized scalar
conservation laws with convex-concave flux function. Classical and non-classical shock waves are approximated well only driven by the dynamics of the equation without a priori knowledge of the solution.

REFERENCES


E-mail address: ernest@stat.math.ethz.ch
CONVERGENCE OF A SL SCHEME FOR THE EIKONAL EQUATION WITH DISCONTINUOUS COEFFICIENTS

ADRIANO FESTA
Imperial College of London
EEE Department
London SW7 2AZ, UK

MAURIZIO FALCONE
Sapienza, Università di Roma
Dipartimento di Matematica
P.za A. Moro 2, 00185 Roma, Italy

Abstract. We consider the stationary eikonal equation where the coefficients are allowed to be discontinuous. The discontinuities must belong to a special class for which the notion of viscosity solutions in the sense of Ishii is suitable. We present a semi–Lagrangian scheme for the approximation of the viscosity solution also studying its properties. The main result is an a-priori error estimate in the $L^1$-norm. In the last section, we illustrate some tests and applications where the scheme is able to compute the right solution.

1. Introduction. In this paper we study the following boundary value problem. Let $\Omega \subset \mathbb{R}^N$ be an open bounded domain with a Lipschitz boundary $\partial \Omega$, we consider the Dirichlet problem

$$
\begin{cases}
    c(x)|Du(x)| = f(x) & x \in \Omega, \\
    u(x) = g(x) & x \in \partial \Omega,
\end{cases}
$$

(1)

where $f$, $c$ and $g$ are given real functions defined on $\Omega$. We focus our attention on the case where $f$ is positive, Borel measurable and possibly discontinuous.

In the most classical case, where $c(x)$ is constantly equal to one, $f(x) \equiv 1$ and $g(x) \equiv 0$, we get the eikonal equation giving the characterization of the distance from $\partial \Omega$. In other applications, e.g. in geometrical optics, computer vision, control theory and robotic navigation, $c$ and $f$ can vary but have typically a constant sign (e.g. positive). It is worth to note that in the study of many problems motivated by real world applications a discontinuous $f$ and/or a degenerate $c$ can appear in a natural way. In fact, one can easily imagine that the velocity of a front in a medium is affected by the physical properties of the medium and can be discontinuous if the medium is stratified by different materials. In the famous Shape-from-Shading problem the right-hand side is $f(x) = [(1 - I^2(x))/I^2(x)]^{1/2}$ where $I$ is the brightness of the image. Depending on the shape of the object represented in the image $I$ can be discontinuous.

Another motivation to deal with discontinuous Hamiltonians comes directly from control theory. In the control framework discontinuous functions can be used to

2000 Mathematics Subject Classification. Primary: 35F30, 35R05; Secondary: 65N15.

Key words and phrases. Hamilton-Jacobi equation, discontinuous Hamiltonian, viscosity solutions, semi–Lagrangian schemes, a-priori error estimates.
represent targets (using \( f \) as a characteristic function) and/or state constraints (using \( f \) as an indicator function). It is interesting to point out that, when the Hamiltonian is discontinuous, the knowledge of \( f \) at every point will not guarantee the well-posedness of the problem, even in the framework of viscosity solution. To deal with this problem we will adopt the notion of discontinuous viscosity solutions via semicontinuous envelopes of \( f \) introduced by Ishii in [13]. Other results of well-posedness of Hamilton-Jacobi equations in presence of discontinuous coefficients have been presented by various authors in several works (see [4, 11, 2]) and in the specific case of the eikonal equation [19, 16].

Our primary goal is to prove convergence for a semi–Lagrangian scheme. The typical convergence result, given in the \( L^\infty \) norm, is natural when dealing with classical viscosity solutions (see e.g. Crandall and Lions [6], Barles and Souganidis [3] and Falcone and Ferretti [10]). It is clear that, dealing with discontinuous coefficients and/or discontinuous solutions, the classical assumptions for convergence in the uniform norm are not satisfied. Then, it seems more natural to look for convergence in the \( L^1 \) norm as it happens in the analysis of approximation schemes for conservation laws. However, the list of contributions on this topic is rather short. At our knowledge, the only two convergence results in \( L^1 \) has been proved by Lin and Tadmor [18, 15] for a central finite difference scheme and by Bokanowski et al. [5] in dimension one. Deckelnick and Elliott [8] studied a problem where the solution is still Lipschitz continuous although the Hamiltonian is discontinuous. In particular, they proposed a finite difference scheme and an \( a\)-\textit{priori} error estimate. Although our work has been also inspired by their results, we use different techniques and our analysis is devoted to a scheme of semi–Lagrangian type (\( SL - scheme \)). The benefits of a SL-scheme with respect to a finite difference scheme are a better ability to follow the informations driven by the characteristics and the fact that they do not require a structured grid. These peculiarities give us a faster and more accurate approximation in many cases as it has been reported in the literature (see e.g. [9, 7] or appendix A of [1]). It is also important to note that we prove an \( a\)-\textit{priori} error estimate in the general case where also discontinuous viscosity solutions may appear.

2. The model problem and previous results. Let us start introducing the definition of discontinuous viscosity solution and summarize for readers convenience some well-posedness results.

\textbf{Assumption A0.} The boundary data
\[ g : \partial \Omega \to [0, +\infty] \text{ is continuous}, \]  
\[ c : \Omega \to \mathbb{R} \text{ is a non negative and continuous function such that } c(x) \leq M_c \text{ for all } x \in \Omega. \]  
Additional hypotheses will be added on the set where \( c \) vanishes later in this paper. Moreover, the function \( f : \mathbb{R}^N \to [\rho, +\infty], \rho > 0 \) is Borel measurable and possibly discontinuous.

Let us remind the definition of \textit{discontinuous viscosity solution} for (1) introduced by Ishii in [13]. Let \( f \) be bounded in \( \Omega \), we define \( f_* \) and \( f^* \) which are respectively the lower semicontinuous and the upper semicontinuous envelope of \( f \) as
\[ f_*(x) = \lim_{r \to 0^+} \inf \{ f(y) : |y - x| \leq r \}, \quad f^*(x) = \lim_{r \to 0^+} \sup \{ f(y) : |y - x| \leq r \}. \]  

\textbf{Definition 2.1.} A lower (resp. upper) semicontinuous function \( u : \Omega \to \mathbb{R} \cup \{ +\infty \} \) (resp. \( u : \Omega \to \mathbb{R} \)) is a viscosity super- (resp. sub-) solution of the equation (1) if for
every \( \phi \in C^1(\Omega) \), \( u(x) < +\infty \), and \( x \in \arg\min_{x \in \Omega} (u - \phi) \), (resp. \( x \in \arg\max_{x \in \Omega} (u - \phi) \)), we have

\[
c(x)|D\phi(x)| \geq f_*(x), \quad (\text{resp. } c(x)|D\phi(x)| \leq f^*(x)).
\]

A function \( u \) is a discontinuous viscosity solution of (1) if \( u^* \) is a subsolution and \( u_* \) is a supersolution.

Note that also Dirichlet boundary conditions must be interpreted in the following weak sense.

**Definition 2.2.** An upper semicontinuous function \( u : \overline{\Omega} \to \mathbb{R} \), subsolution of (1), satisfies the Dirichlet type boundary condition in the viscosity sense if for all \( \phi \in C^1(\Omega) \) and \( x \in \partial \Omega \), \( x \in \arg\max_{x \in \overline{\Omega}} (u - \phi) \) such that \( u(x) > g(x) \), we have

\[
c(x)|D\phi(x)| \leq f^*(x).
\]

Lower semicontinuous functions that satisfy a Dirichlet type boundary condition are defined accordingly.

In order to guarantee uniqueness we add the following assumption on \( f \).

**Assumption A1.** Let us assume that there exist \( \eta > 0 \) and \( K \geq 0 \) such that for every \( x \in \Omega \) there is a direction \( n = n_x \in \mathcal{S}^N \) (\( \mathcal{S}^N \) is the unit ball of dimension \( N \) centered at 0) such that

\[
f(y + rd) - f(y) \leq Kr,
\]

for every \( y \in \Omega \), \( d \in \mathcal{S}^N \), \( r > 0 \) with \( |y - x| < \eta \), \( |d - n| < \eta \) and \( y + rd \in \Omega \).

Under Assumptions A0–A1, a comparison theorem between sub- and supersolution holds [8] (a more general result can be found in [17]). It is important to highlight that adopting the concept of discontinuous viscosity solution, that comparison result is not enough to prove uniqueness. Multiple discontinuous solutions may exist without any contradiction. In that case, an important role is played by two special elements of the class of solutions, the minimal supersolution and the maximal subsolution, defined respectively (see [17] for details) as

\[
V_m = \inf_{a \in A} \int_0^{\tau_x(a)} f_*(y(t, a))dt + g(y(\tau_x(a)), a),
\]

\[
V_M = \inf_{a \in A} \int_0^{\tau_x(a)} f^*(y(t, a))dt + g(y(\tau_x(a)), a);
\]

where \( \tau_x(a) \) is the first exit time from the domain of a trajectory starting from \( x \) and subject to the controlled dynamics \( \dot{y}(t) = a(t) \), with control \( a \) in \( \mathcal{S}^N \). It is also important to remark that in the case of existence of a continuous viscosity solution automatically \( V_m \equiv V_M \) and the family of solutions reduces to just one solution. For an example of this case let us consider (1) in \( \Omega = (-1, 1) \), \( f(x) = 0 \), for \( x < 0 \), and \( f(x) = x \) for \( x \geq 0 \) and \( c(x) = x \). Let us fix the boundary condition \( u(-1) = u(1) = 0 \). It is easy to verify that the piecewise continuous function

\[
u(x) = \begin{cases} 
0 & x < 0 \\
1 - x & x \geq 0 
\end{cases}
\]

is a viscosity solution of the problem. Indeed, we can change at \( x = 0 \) the value for the solution in \([0, 1]\) obtaining a family of discontinuous viscosity solutions whose upper semicontinuous envelope is always \( V_M \) whereas the lower semicontinuous envelope coincides with \( V_m \).
In order to state a more precise result, let us restrict ourselves to the special case \( N = 2 \) where \( c \) vanishes on an interface \( \Sigma_0 \) splitting the domain in two parts. This choice is made to simplify the presentation, more general situations can be treated in a similar way.

Let us denote by \( \ell(C) \) the length of a curve \( C \) and assume the existence of a regular curve \( \Sigma_0 \) which splits the domain \( \Omega \) into two subset \( \Omega_j \), \( j = 1, 2 \), where \( c \) does not vanish.

**Assumption A2.** Let \( \Sigma_0 := \{ x \in \Omega | c(x) = 0 \} \), we assume that \( \Omega = \Omega_1 \cup \Omega_2 \cup \Sigma_0 \) \( \ell(\Sigma_0) < +\infty \) and \( \Omega_j \cap \partial \Omega \neq \emptyset \) for \( j = \{ 1, 2 \} \).

We conclude this section with a regularity result, which can be derived by adapting the classical proof by Ishii [14]:

**Theorem 2.3.** Let \( \Omega \) be an open domain with Lipschitz boundary. Assume A0, A1, A2. Let \( u : \overline{\Omega} \to \mathbb{R} \) be a bounded viscosity solution of the problem (1), then \( u \) is Lipschitz continuous in every set \( \Omega_1 \) and \( \Omega_2 \).

3. The semi–Lagrangian approximation scheme and its properties. We construct a semi–Lagrangian approximation scheme for the equation (1) following the approach illustrated in [9].

Using the Kruzkov’s change of variable, \( v(x) = 1 - e^{-u(x)} \), problem (1) becomes

\[
\begin{cases}
\max_{a \in S^2} \{ c(x) a \cdot Dv(x) \} = f(x)(1 - v(x)) & x \in \Omega, \\
v(x) = 1 - e^{-g(x)} & x \in \partial \Omega.
\end{cases}
\]  
(7)

There exists a clear interpretation of this equation as the value function of an optimization problem with constant running cost and discount factor equal to one, and the dynamics given by \( \frac{c(x)}{f(x)} a \) (see [1] for more details).

We discretize the left-hand side term of the first equation in (7) as a directional derivative getting the following discrete problem:

\[
\begin{cases}
\max_{a \in S^2} \{ c(x) a \cdot Dv_h(x) \} = f(x)(1 - v_h(x)) & x \in \Omega, \\
v_h(x) = 1 - e^{-g(x)} & x \in \partial \Omega.
\end{cases}
\]  
(8)

where \( h \) is a positive real number and we will assume (to simplify the presentation) that \( x - \frac{h}{f(x)} c(x) a \in \overline{\Omega} \) for every \( a \in S^2 \).

Let introduce a space discretization of (8) yielding a fully discrete scheme. We construct a regular triangulation of \( \Omega \) made by a family of simplices \( S_j \), such that \( \Omega = \bigcup_j S_j \), denoting \( x_m, m = 1, ..., L \), the nodes of the triangulation, by \( \Delta x := \max_j \text{diam}(S_j) \) the size of the mesh (\( \text{diam}(B) \) denotes the diameter of the set \( B \)) and by \( G \) the set of the knots of the grid. We look for a solution of

\[
\begin{cases}
W(x_m) = \frac{1}{1+h} \min_{a \in B(0,1)} I[W](x_m - \frac{h}{f(x_m)} c(x_m) a) + \frac{h}{1+h} x_m \in G, \\
W(x_m) = 1 - e^{-g(x_m)} x_m \in G \cap \partial \Omega,
\end{cases}
\]  
(9)

where \( I[W](x) \) is a linear interpolation of \( W \) at the point \( x \). Therefore, we look for the solution of equation (9) in the space of piecewise linear functions

\( \mathcal{W}^{\Delta x} := \{ w : \overline{\Omega} \to \mathbb{R} | w \in C(\Omega) \ \text{and} \ Dw(x) = \text{cost}_j \ \text{for any} \ x \in S_j \} \),

the existence and uniqueness of a solution in such space is an easy application of the Contraction Mapping Theorem.
\( \Delta x = h \) & \( V_m \text{error} \| \cdot \|_1 \) & \( \text{Ord}(L^1) \) & \( V_M \text{error} \| \cdot \|_1 \) & \( \text{Ord}(L^1) \) \\ \hline \\ 0.2 & 0.1729 & 0.2040 & & \\ 0.1 & 0.1166 & 0.5684 & 0.1054 & 1.0642 \\ 0.05 & 0.0765 & 0.6080 & 0.0517 & 1.0276 \\ 0.025 & 0.0495 & 0.6280 & 0.0257 & 1.0084 \\ 0.0125 & 0.0349 & 0.5042 & 0.0130 & 0.9833 \\

Table 1. Test 1: experimental errors in \( L^1(\Omega) \) for the approximation of \( V_m \) and \( V_M \).

**Proposition 1.** Let \( x_m = \frac{h}{f(x_m)}v(x_m) a \in \Omega \), for every \( x_m \in G \) and for any \( a \in B(0, 1) \), then there exists a unique solution \( W \) of (9) in \( W^\Delta x \).

We can also prove that this scheme is monotone and consistent with the equation, and the following a-priori error estimate in \( L^1(\Omega) \) holds true.

**Theorem 3.1.** Let \( A0, A1 \) and \( A2 \) hold true. Let \( v(x) \) be a viscosity solution of (7) and \( W(x) \in W^\Delta x \) be a piecewise linear function satisfying (9). Then, there exist two positive constants \( C, C' \) (independent from \( h \) and \( \Delta x \)) such that for \( h \) and \( \Delta x \) satisfying \( h \Delta x \leq \rho M c \), (\( \rho \) and \( M c \) appear in Assumption \( A0 \)) we have:

\[
\|v(x) - W(x)\|_{L^1(\Omega)} \leq C\sqrt{h} + C'\Delta x. \tag{10}
\]

**Proof.** We just sketch the main steps of the proof (which can be found in [12]). We start introducing the set \( \Sigma_{\Delta x} \) defined as

\[
\Sigma_{\Delta x} := \{ x \in \Omega | S^2_{x,\Delta x} \cap \Sigma_0 \neq \emptyset \},
\]

where \( S^2_{x,\Delta x} \) denotes the ball of radius \( \Delta x \) centred at \( x \). We observe that

\[
\|v(x) - W(x)\|_{L^1(\Omega)} \leq \sum_{j=1,2^*} \int_{\Omega_j} |v(x) - W(x)| dx + \int_{\Sigma_{\Delta x}} |v(x) - W(x)| dx, \tag{11}
\]

where \( \Omega := \cup_j \Omega_j \) is the partition of \( \Omega \) generated from \( \Sigma_0 \).

By the definition of Kruzkov’s transform, we know that \( |v(x) - W(x)| \leq 2 \) for every \( x \in \Omega \) and adding the assumptions on the set \( \Sigma_0 \) we get, for a fixed \( C' > 0 \),

\[
\int_{\Sigma_{\Delta x}} |v(x) - W(x)| dx \leq 2 \int_{\Sigma_{\Delta x}} dx \leq 2\ell(\Sigma_0) \Delta x \leq C' \Delta x. \tag{12}
\]

To prove the statement, we need to prove an estimate for the term \( \int_{\Omega_j} |v(x) - W(x)| dx \) for every \( j \). To this end, we remind that for Theorem 2.3, both \( v(x) \) and \( W(x) \) are Lipschitz continuous, so we can use a modification of the classical argument based on the variable doubling.

4. Numerical experiments and applications. In this section we present some test problems pointing out the main features of our numerical scheme.

4.1. Test 1: a 1D example. We want to solve the following equation on the interval \([-1, 1]\)

\[
| xu' | = f(x), \tag{13}
\]
Figure 1. Test 1 (left): two approximations of the value function, Test 2 (right): level sets of the approximated value function.

with \( u(\pm 1) = 0 \) and

\[
f(x) := \begin{cases} 
1 & x > 0, \\
0 & x < 0.
\end{cases}
\]  

(14)

We denote by \( f^* \) the function defined above where \( f(0) = 1 \) and by \( f_* \) the same function where \( f(0) = 0 \). As we explained in the previous section, to have a bounded solution we use the Kruzkov transform to pass to the equation

\[
|xv'| = f(x)(1 - v(x)).
\]  

(15)

In this case we can get easily a correct solution (which is not unique) and we can explicitly calculate \( V_m \) and \( V_M \). In this particular case the solutions of the dynamics are very simple, they are \( y(t) = xe^{-at} \). Then we have

\[
V_m(x) = \begin{cases} 
\int_0^{\ln x} 0 \ dt = 0 & x \leq 0 \\
\int_0^{\ln x} e^{-t} dt = [-e^{-t}]_0^{\ln x} = 1 - x & x > 0
\end{cases}
\]  

(16)

It is simple to show with the same reckoning, that \( V_m(x) = V_M(x) \) for \( x \in \Omega \setminus \{0\} \). The functions will diverge only at \( x = 0 \) where \( V_m(0) = 0 \neq 1 = V_M(0) \) accordingly to the fact that \( V_m \) is l.s.c. and \( V_M \) is u.s.c.

Therefore, in this case we do not have a unique viscosity solutions, however

\[
V_m(x) = (V_M)_*(x) = \begin{cases} 
0 & x \leq 0, \\
1 - x & x > 0,
\end{cases}
\]  

(17)

is our unique lower semicontinuous solution. We want to verify if the numerical approximation introduced in the previous section, converges to all the elements of the solutions class. Let make a test obtaining the results contained in Figure 1 and in Table 1 where we perform the scheme using \( f_* \) (in some sense, we are computing \( V_m \)) and \( f^* \) (computing \( V_M \)). We can see in both cases a good convergence in \( L^1 \)-norm although there is no convergence in the \( L^\infty \)-norm.

4.2. Test 2: a 2D example. Let us consider, the equation \(|Du(x)| = f(x)\) in \( \Omega = (-1, 1)^2 \), where

\[
f(x_1, x_2) := \begin{cases} 
2, & (x_1 - \frac{1}{2})^2 + x_2^2 \leq \frac{1}{8} \text{ and } x_2 \geq x_1 - \frac{1}{2}, \\
3, & (x_1 - \frac{1}{2})^2 + x_2^2 \leq \frac{1}{8} \text{ and } x_2 < x_1 - \frac{1}{2}, \\
1, & \text{otherwise},
\end{cases}
\]
and we apply homogeneous boundary conditions $u(x) = 0$ on $\partial \Omega$. Note that in this case, discontinuities of $f$ occur both along curved lines and straight lines. This shows the good capability of the semi–Lagrangian scheme to approximate the solution (Figure 1), this is due to the fact that the directions used in the scheme are not aligned with the geometry of the grid. Note that in this case we have existence of a continuous solution (therefore uniqueness of the viscosity solution) and that the error estimate in $L^1$–norm applies. We can see in Table 2 that in this case we have also convergence in the $L^\infty$–norm. Due to the fact that in this case an analytic solution is not available we have used as exact solution a numerical approximation obtained on a very fine grid ($\Delta x = 0.005$).

4.3. Test 3: finding the exit from a labyrinth. We apply our scheme to find the exit path from a labyrinth $Q$. We can write this problem as a minimum time problem with state constraints (the walls can not be crossed). The geometry of the labyrinth is shown in Figure 2 where the gray square is the exit (the target $T$ for the minimum time problem). We have computed the solution of

$$|Du(x)| = f(x) \quad x \in Q \setminus T,$$

with Dirichlet boundary conditions $u(x) = 0$ on $\partial T$ and discontinuous running cost

$$f(x) = \begin{cases} M & \text{if } x \text{ is on the walls}, \\ 1 & \text{otherwise}. \end{cases}$$

In the test we have chosen $\Delta x = h = 0.0078, M = 10^{10}$. In Figure 2 we can see the plot of the value function obtained.

Acknowledgments. This work was supported by the European Union under the 7th Framework Programme FP7-PEOPLE-2010-ITN SADCO, Sensitivity Analysis for Deterministic Controller Design.

REFERENCES
Figure 2. Test 3: The labyrinth (left) and the value function of the minimum time problem (right).


E-mail address: a.festa@imperial.ac.uk
E-mail address: falcone@mat.uniroma1.it
A BRIEF NOTE ON RECENT DEVELOPMENTS ON DIVERGENCE-MEASURE FIELDS

HERMANO FRID

Instituto de Matemática Pura e Aplicada - IMPA
Estrada Dona Castorina, 110
Rio de Janeiro, RJ 22460-320, Brazil

Abstract. In this work we are particularly interested in analyzing some consequences of the additional assumption that the domain has a Lipschitz boundary, in the investigation of the properties of the divergence-measure fields, specially the ones so denominated extended, that is, those which are vector-valued (Radon) measures whose divergence is a signed (Radon) measure.

1. Introduction. The purpose of this paper is to give a short overview on recent developments of the theory of the divergence-measure fields introduced by Chen and Frid [2, 3, 4], which was further developed by Silhavý [14, 15]. We recall basic definitions, results, and include some discussion for the case when one makes the additional assumption that the underlying domain has a Lipschitz boundary. We begin by making a brief review of the basic theory, and then we make the assumption that the domain possesses a Lipschitz deformable boundary and pass to analyze some consequences of this assumption. We refer to [13] for a more detailed review of the theory of the divergence-measure fields up to this date. We also refer [8] and the papers already mentioned for a more complete bibliography on the theory of divergence-measure fields, as well as many others of its possible applications.

2. Divergence-Measure Fields. We begin by recalling the definition of the divergence-measure fields.

Definition 2.1. Let $U \subset \mathbb{R}^N$ be open. For $F \in L^p(U; \mathbb{R}^N)$, $1 \leq p \leq \infty$, or $F \in \mathcal{M}(U; \mathbb{R}^N)$, set

$$|\text{div} \, F|(U) := \sup \left\{ \int_U \nabla \varphi \cdot F : \varphi \in C^1_0(U), \ |\varphi(x)| \leq 1, \ x \in U \right\}.$$  \hfill (1)

For $1 \leq p < \infty$, we say that $F$ is an $L^p$-divergence-measure field over $U$, i.e., $F \in \mathcal{DM}^p(U)$, if $F \in L^p(U; \mathbb{R}^N)$ and

$$\|F\|_{\mathcal{DM}^p(U)} := \|F\|_{L^p(U; \mathbb{R}^N)} + |\text{div} \, F|(U) < \infty.$$  \hfill (2)

2000 Mathematics Subject Classification. Primary: 26B20, 28C05, 35L65, 35B35; Secondary: 26B35, 26B12, 35L67.

Key words and phrases. Divergence-measure fields, normal traces, Gauss-Green theorem, product rule.

The author thanks CNPq grant proc. 303950/2009-9, and FAPERJ, through grant E-26/103.019/2011, for partially supporting to this research.
We say that \( F \) is an extended divergence-measure field over \( D \), i.e., \( F \in \mathcal{DM}^{ext}(U) \), if \( F \in \mathcal{M}(U; \mathbb{R}^N) \) and
\[
\|F\|_{\mathcal{DM}^{ext}(U)} := |F|(U) + |\text{div} F|(U) < \infty. \tag{3}
\]
If \( F \in \mathcal{DM}^*(U) \) for any open set \( U \subseteq \mathbb{R}^N \), then we say \( F \in \mathcal{DM}^*_{\text{loc}}(\mathbb{R}^N) \).

We first recall the following product rule proved in [2], which is a sort of dual of the product rule in the next Theorem 2.3, and its proof is almost entirely transposed to prove that product rule, which is the key for the establishment of the Gauss-Green formula (see Theorem 2.4 below).

**Theorem 2.2** (Chen and Frid [2]). Given \( F \in \mathcal{DM}^{\infty}(U) \) and \( g \in BV(U) \cap L^{\infty}(U) \), then \( gF \in \mathcal{DM}^{\infty}(U) \) and
\[
\text{div}(gF) = \tilde{g}\text{div} F + F \cdot \nabla g, \tag{4}
\]
in the sense of Radon measures in \( U \), where \( \tilde{g} \) equals \( g \) a.e. is the limit of a mollified sequence for \( g \) through a symmetric mollifier, and \( F \cdot \nabla g \) is a Radon measure absolutely continuous with respect to \( |\nabla g| \), whose absolutely continuous part with respect to the Lebesgue measure in \( U \) satisfies
\[
(F \cdot \nabla g)_{ac} = F \cdot (\nabla g)_{ac}, \quad \text{a.e. in } U. \tag{5}
\]
Moreover, and \( |F \cdot \nabla g|(U) \leq \|F\|_{\infty} |\nabla g|(U) \).

We now recall a result by Silhavý in [14] that is in some sense a dual formulation for the previous result, in the sense that it compensates a relaxation on the regularity of the vector field \( F \), which now may be just a vector measure, by imposing more regularity on the function \( g \), which now is assumed to be in \( W^{1,\infty}(U) \). The proof follows exactly by the same lines as that of Theorem 2.2 just recalled, and so we omit it.

**Theorem 2.3** (Silhavý [14]). Given \( F \in \mathcal{DM}^{ext}(U) \) and \( g \in W^{1,\infty}(U) \), then \( gF \in \mathcal{DM}^{ext}(U) \) and
\[
\text{div}(gF) = g\text{div} F + \nabla g \cdot F, \tag{6}
\]
in the sense of Radon measures in \( U \), where \( \nabla g \cdot F \) is a Radon measure absolutely continuous with respect to \( |F| \). Moreover,
\[
\begin{align*}
(i) & \quad |\nabla g \cdot F|(U) \leq \|\nabla g\|_{\infty} |F|(U), \\
(ii) & \quad \text{If } h \in W^{1,\infty}(U), \nabla (gh) \cdot F = h \nabla g \cdot F + g \nabla h \cdot F = \nabla g \cdot hF + \nabla h \cdot gF. \\
(iii) & \quad \text{If } V \subset U \text{ is an open set, then } (\nabla g \cdot F)[V] = \nabla g \cdot F|_{U}|V|. \\
(iv) & \quad (\nabla g \cdot F)_{ac} = \nabla g \cdot (F)_{ac}.
\end{align*}
\]

We recall the Gauss-Green formula for general divergence-measure fields, first proved in [3, 4] and extended by Silhavy in [14].

**Theorem 2.4** (Chen & Frid [3, 4], Silhavy [14]). If \( F \in \mathcal{DM}^{ext}(U) \) then there exists a linear functional \( F \cdot \nu : \text{Lip}(\partial U) \to \mathbb{R} \) such that
\[
F \cdot \nu(g|_{\partial U}) = \int_U \nabla g \cdot F + \int_U g \text{ div} F, \tag{7}
\]
for every \( g \in \text{Lip}(\mathbb{R}^N) \cap L^{\infty}(\mathbb{R}^N) \). Moreover,
\[
|F \cdot \nu(h)| \leq |F|_{\mathcal{DM}(U)} |h|_{\text{Lip}(\partial U)}, \tag{8}
\]
for all \( h \in \text{Lip}(\partial U) \), where we use the notation
\[
|g|_{\text{Lip}(C)} := \sup_{x \in C} |g(x)| + \text{Lip}_C(g).
\]
We now discuss a direct way of defining the normal trace functional $F \cdot \nu : \text{Lip} (\partial U) \to \mathbb{R}$. The formula was first obtained in [3, 4], under regularity restrictions on the boundary, and in [14], for general boundaries. Before stating the corresponding result, we recall the following lemma, which is a slight modification of Lemma 3.3 of [14].

**Lemma 2.5** (Silhavy [14]). If $F \in \mathcal{DM}^{c\mathbb{F}}(U)$, $m \in \text{Lip}(U)$, $t \in \mathbb{R}$ and if $T \subset m^{-1}(t)$ is a compact subset of $U$, then the restriction $\nabla m \cdot F|_T$, of $\nabla m \cdot F$ to $T$, satisfies

$$
\nabla m \cdot F|_T = 0.
$$

(9)

The following result gives a simple formula to compute the normal trace of $\mathcal{DM}$-fields. This formula, displayed in (i) of the statement below, was first obtained in [3, 4] under some regularity restrictions on the boundary, and later was extended to general domains in [14]. The item (ii) gives an useful necessary condition for the normal trace to be a measure over $\partial U$ established by Silhavy [14].

**Theorem 2.6.** Let $F \in \mathcal{DM}^{c\mathbb{F}}(U)$ and $m : \mathbb{R}^N \to \mathbb{R}$ be a nonnegative Lipschitz function with $\text{supp} m \subset U$ which is strictly positive on $U$, and for each $\varepsilon > 0$ let $L_\varepsilon = \{x \in U : 0 < m(x) < \varepsilon\}$. Then:

(i) (cf. [3, 4] and [14]) If $g \in \text{Lip}(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$, we have

$$
F \cdot \nu(g)\partial U = - \lim_{\varepsilon \to 0} \varepsilon^{-1} \int_{L_\varepsilon} g \, d(\nabla m \cdot F);
$$

(10)

(ii) (cf. [14]) If

$$
\liminf_{\varepsilon \to 0} \varepsilon^{-1} |\nabla m \cdot F|(L_\varepsilon) < \infty,
$$

(11)

then $F \cdot \nu$ is a measure over $\partial U$.

**Remark 1.** A typical example of $m$ in the statement of Theorem 2.6 is provided by $m(x) = \text{dist}(x, \partial U)$, for $x \in U$, and $m(x) = 0$, for $x \in \mathbb{R}^N \setminus U$.

**Remark 2.** The following interesting example from [14] shows cases where $F \cdot \nu$ is a measure over $\partial U$ and cases where $F \cdot \nu$ fails to be a measure. Namely, for $1 \leq \alpha < 3$, let $F : \mathbb{R}^2 \setminus \{0\} \to \mathbb{R}^2$ be defined by

$$
F(x) = \frac{1}{|x|^{\alpha}}(x_2, -x_1),
$$

and let $U = \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1, \; x_2 < 0\}$. Clearly, $\text{div} F = 0$, in $\mathbb{R}^2 \setminus \{0\}$, and we easily verify that

$F \in \mathcal{DM}^{c\mathbb{F}}(U; \mathbb{R}^2)$ with $1 \leq p < 2/(\alpha - 1)$, for $1 < \alpha < 3$, and $p = \infty$, for $\alpha = 1$. Now, if $g \in \text{Lip}(\partial U)$ and $\text{supp} \; g \subset \{(x_1, x_2) \in \mathbb{R}^2 : x_2 = 0, \; |x_1| < 1\}$, we may use (11) with $m$ satisfying $m(x) = -x_2$, for $-\varepsilon_0 < x_2 \leq 0$, and $|x_1| < 1 - \varepsilon_0$, with $\varepsilon_0 > 0$ small enough so that $g(x) = 0$, if $|x_1| \geq 1 - \varepsilon_0$. Also, we may consider an extension of $g$ to $\mathbb{R}^2$ such that $g(x_1, x_2) = g(x_1, 0)$, for $|x_2| < \varepsilon_0$. Applying (11) with $m$ and the extension of $g$ so defined, we get

$$
F \cdot \nu(g) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^\varepsilon \int_{-1}^1 g(t, s) \frac{t}{(t^2 + s^2)^{1/2}} \, dt \, ds,
$$

which gives

$$
F \cdot \nu(g) = \int_{-1}^1 g(t, 0) \text{sgn}(t) |t|^{1-\alpha} \, dt,
$$

for $1 \leq \alpha < 2$, for
and
\[ F \cdot \nu(g) = \lim_{\varepsilon \to 0} \int_{|t| > \varepsilon} g(t,0) \text{sgn}(t)|t|^{1-\alpha} \, dt, \quad \text{for } 2 \leq \alpha < 3. \]

This shows that, for $1 \leq \alpha < 2$, $F \cdot \nu$ is a measure, while, for $2 \leq \alpha < 3$, $F \cdot \nu$ is not a measure on $\partial U$.

**Remark 3.** For $\varepsilon_0 > 0$ sufficiently small and $0 < s < \varepsilon_0$, we may consider the open set $U_s := \{ x \in U : m(x) > s \}$, for $m$ as in Theorem 2.6. By Theorem 2.6, for the normal trace $F \cdot \nu|\partial U_s$, we have the following formula similar to (10),
\[ F \cdot \nu(g|\partial U_s) = -\lim_{\varepsilon \to 0} \varepsilon^{-1} \int_{\{s < m(x) > s+\varepsilon\}} g \, d(\nabla m \cdot F), \quad (12) \]
and, again, we have that the condition
\[ \liminf_{\varepsilon \to 0} \varepsilon^{-1}|\nabla m \cdot F|(\{s < m(x) < s+\varepsilon\}) < \infty, \quad (13) \]
implies that $F \cdot \nu|\partial U_s$ is a measure on $\partial U_s$. If we consider the monotone function $V(s) = |\nabla m \cdot F|(\{0 < m(x) < s\})$, for $s \in (0,\varepsilon_0)$, we see that the left-hand side of (13) is the right-derivative of $V$ at $s$, except possibly for a countable subset of $(0,\varepsilon_0)$, and we know that it exists for a.e. $s \in (0,\varepsilon_0)$. Therefore, $F \cdot \nu|\partial U_s$ is a measure for a.e. $s \in (0,\varepsilon_0)$, and in this sense we may assert that “for almost all boundaries” $\partial U$ the normal trace $F \cdot \nu|\partial U$ is a measure.

3. Domains with a Lipschitz deformable boundary. We now enter into the main subject of the present paper, beginning with the definition of deformable Lipschitz boundary.

**Definition 3.1.** Let $\Omega \subset \mathbb{R}^N$ be an open set. We say that $\partial \Omega$ is a **deformable Lipschitz boundary** if the following hold:

(i) For each $x \in \partial \Omega$, there exist $r > 0$ and a Lipschitz mapping $\gamma : \mathbb{R}^{N-1} \to \mathbb{R}$ such that, upon rotating and relabeling the coordinate axis if necessary,
\[ \Omega \cap Q(x,r) = \{ y \in \mathbb{R}^N : \gamma(y_1,\cdots,y_{N-1}) < y_N \} \cap Q(x,r), \]
where $Q(x,r) = \{ y \in \mathbb{R}^N : |y_i - x_i| \leq r, \ i = 1,\cdots,N \}$. We denote by $\tilde{\gamma}$ the map $\tilde{y} \mapsto (\tilde{y},\gamma(\tilde{y}))$, $\tilde{y} = (y_1,\cdots,y_{N-1})$.

(ii) There exists a map $\Psi : \partial \Omega \times [0,1] \to \bar{\Omega}$ such that $\Psi$ is a bi-Lipschitz homeomorphism over its image and $\Psi(x,0) = x$, for all $x \in \partial \Omega$. For $s \in [0,1]$, we denote by $\Psi_s$, the mapping from $\partial \Omega$ to $\bar{\Omega}$ given by $\Psi_s(x) = \Psi(x,s)$, and set $\partial \Omega_s := \Psi_s(\partial \Omega)$.

We say that the Lipschitz deformation $\Psi : \partial \Omega \times [0,1] \to \bar{\Omega}$ is **regular**, and that $\Omega$ has a **regular Lipschitz deformable boundary**, if, besides (i) and (ii), we have

(iii) $J[\nabla \Psi_s \circ \tilde{\gamma}] \to J[\nabla \tilde{\gamma}]$, as $s \to 0$, in the weak star topology of $L^\infty(B)$ for any bounded open set $B \subset \mathbb{R}^{N-1}$ such that $\tilde{\gamma}(B) \subset \partial \Omega$, with $\tilde{\gamma}$ as in (i), where $J[\nabla g]$ denotes the Jacobian of $\nabla g$ (see, e.g., [10]).

We start our discussion by introducing the level set function $h : \mathbb{R}^N \to \mathbb{R}$, defined by
\[ h(x) = \begin{cases} 0, & \text{for } x \in \mathbb{R}^N \setminus \bar{\Omega}, \\ s, & \text{for } x \in \partial \Omega_s, \\ 1, & \text{for } x \in \Omega \setminus \Psi(\partial \Omega \times [0,1]). \end{cases} \]
By the formula (10) we have

\[ F \cdot \nu(g | \partial U) = - \lim_{\varepsilon \to 0} \varepsilon^{-1} \int_{L_\varepsilon} g d(\nabla h \cdot F), \tag{14} \]

for any \( F \in DM^{ext}(\Omega) \), and any \( g \in \text{Lip} (\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N) \), with \( L_\varepsilon = \{ x \in \Omega : 0 < h(x) < \varepsilon \} \).

**Remark 4.** The following standard example of a domain with a regular deformable Lipschitz boundary shows that, for the sake of studying local properties of the normal trace operator, any domain with a Lipschitz boundary may be viewed as a domain with a regular deformable Lipschitz boundary. So, let

\[ U := \{ x \in \mathbb{R}^N : \gamma(x_1, \ldots, x_{N-1}) < x_N \}, \tag{15} \]

with \( \gamma : \mathbb{R}^{N-1} \to \mathbb{R} \) is a Lipschitz function. \( U \) is then an unbounded open set, \( \partial U \) is the graph of \( \gamma \), \( \partial U = \{ (\hat{x}, x_N) \in \mathbb{R}^N : \hat{x} \in \mathbb{R}^{N-1}, x_N = \gamma(\hat{x}) \} \), and it is very easy to define a regular Lipschitz deformation for \( \partial U \) by simply setting \( \Psi((\hat{x}, \gamma(\hat{x})), s) = (\hat{x}, \gamma(\hat{x}) + s \delta), \ \hat{x} \in \mathbb{R}^{N-1}, s \in [0, 1], \) where \( \delta > 0 \) is arbitrary.

It turns out that, by property (i) in Definition 3.1, for test functions \( g \), as in (7), with support contained in a sufficiently small neighborhood, say, a neighborhood like those appearing in Definition 3.1(i), the normal trace operator given by (7), may be defined using (14) where \( h \) is the level set function associated to this trivial standard deformation. More specifically, in this case, the level set function is simply defined by

\[ h(x) = \begin{cases} 0, & \text{if } x_N < \gamma(\hat{x}), \\ s, & \text{if } x_N = \gamma(\hat{x}) + s \delta, \text{ for } s \in [0, 1], \\ 1, & \text{if } x_N \geq \gamma(\hat{x}) + \delta. \end{cases} \]

Therefore, considered as distributions in \( \mathbb{R}^N \) with support contained in \( \partial \Omega \), the normal trace operators associated to \( DM^{ext} \)-fields can always be split in a countable sum of distributions, whose supports possess the finite intersection property, each of which may be defined like the normal trace operator for a standard domain as just described. Indeed, it suffices to employ a partition of unity subordinated to a suitable covering of \( \partial \Omega \).

**Theorem 3.2.** Let \( \Omega \) be a bounded open set with a deformable Lipschitz boundary and \( F \in DM^1(\Omega) \). Let \( \Psi : \partial \Omega \times [0, 1] \to \Omega \) be a Lipschitz deformation of \( \partial \Omega \). Then, for almost all \( s \in [0, 1] \), and all \( \phi \in C_0^\infty(\mathbb{R}^N) \),

\[ \int_{\Omega_s} \phi \text{div} F = \int_{\partial \Omega_s} \phi(\omega) F(\omega) \cdot \nu_s(\omega) dH^{N-1}(\omega) - \int_{\Omega_s} F(x) \cdot \nabla \phi(x) dx, \tag{16} \]

where \( \nu_s \) is the unit outward normal field defined \( H^{N-1} \)-almost everywhere in \( \partial \Omega_s \), and \( \Omega_s \) is the open subset of \( \Omega \) bounded by \( \partial \Omega_s \).

We have the following formula for the normal trace for a \( DM^1 \)-field.

**Theorem 3.3.** Let \( F \in DM^1(\Omega) \), where \( \Omega \) is a bounded open set with a Lipschitz boundary admitting a regular deformation \( \Psi : \partial \Omega \times [0, 1] \to \Omega \). Denoting by \( F \cdot \nu|_{\partial \Omega} \) the continuous linear functional \( \text{Lip}(\partial \Omega) \to \mathbb{R} \) given by the normal trace of \( F \) at \( \partial \Omega \), we have the formula

\[ F \cdot \nu|_{\partial \Omega} = \text{ess. lim}_{s \to 0} F \circ \Psi_s(\cdot) \cdot \nu_s(\Psi_s(\cdot)), \tag{17} \]

with equality in the sense of \( (\text{Lip}(\partial \Omega))^* \), where on the right-hand side the functionals are given by ordinary functions in \( L^1(\partial \Omega) \).
We now turn to the case where $F \in \mathcal{DM}^{ext}(\Omega)$. Let us again consider the level set function $h$ associated to the regular Lipschitz deformation $\Psi : \partial \Omega \times [0, 1] \to \Omega$. Let us consider the measure $\mu$ over $\Psi(\partial \Omega \times [0, 1])$ given by

$$\mu := |\nabla h - F| \big| \Psi(\partial \Omega \times [0, 1]).$$

We consider the pull back of $\mu$ by $\Psi$, $\Psi^* \mu$, which is the measure on $\partial \Omega \times [0, 1]$ defined by

$$\langle \Psi^* \mu, \varphi \rangle = \langle \mu, \varphi \circ \Psi^{-1} \rangle, \quad \forall \varphi \in C(\partial \Omega \times [0, 1]).$$

We may apply the disintegration process to $\Psi^* \mu$ (see, e.g., theorem 2.28, p.57 in [1]) to write $\Psi^* \mu = \sigma \otimes \tilde{\mu}_s$, for the Radon measure $\sigma$ on $[0, 1]$ given by the projection of $\Psi^* \mu$ onto $[0, 1]$, and so $\sigma(E) = \int E \Psi^* \mu(\partial \Omega \times E)$ for any Borelian $E \subset [0, 1]$, and Radon measures $\tilde{\mu}_s$ such that $s \mapsto \tilde{\mu}_s$ is $\sigma$-measurable, $\tilde{\mu}_s(\partial \Omega) = 1$, $\sigma$-a.e. in $[0, 1]$, so that we have

$$\int_{\partial \Omega \times [0, 1]} \varphi(\omega, s) d\Psi^* \mu = \int_{[0, 1]} \left( \int_{\partial \Omega} \varphi(\omega, s) d\tilde{\mu}_s(\omega) \right) d\sigma(s), \quad \forall \varphi \in C(\partial \Omega \times [0, 1]).$$

Therefore, by pushing forward the equation $\Psi^* \mu = \sigma \otimes \tilde{\mu}_s$ by $\Psi$, we obtain

$$\mu = \sigma \otimes \mu_s, \quad \mu_s := (\Psi_s)^* \tilde{\mu}_s,$$

where, for any $\zeta \in C(\partial \Omega_s)$,

$$\langle (\Psi_s)_* \tilde{\mu}_s, \zeta \rangle = (\tilde{\mu}_s, \zeta \circ \Psi_s).$$

In particular, for any $\phi \in C_0^\infty(\mathbb{R}^N)$, with supp $\phi \cap \Omega \subset \Psi(\partial \Omega \times [0, 1])$, we have

$$\int_{\Omega} \phi(x) \theta \cdot d\mu = \int_{[0, 1]} \left( \int_{\partial \Omega_s} \phi(x) \theta(x) d\mu_s \right) d\sigma(s), \quad \forall \theta \in C(\partial \Omega \times [0, 1]).$$

where $\theta$ is the $\mu$-measurable function, with $|\theta| = 1$, $\mu$-a.e., such that

$$\theta \mu = |\nabla h - F| \big| \Psi(\partial \Omega \times [0, 1]).$$

Now, we have the decomposition $\sigma = H(s) d\sigma + \sigma_{\text{sing}}$, for some non-negative $H \in L^1([0, 1])$, and $\sigma_{\text{sing}} = \sigma(\mathcal{N})$, for some Borelian $\mathcal{N} \subset [0, 1]$ of one-dimensional Lebesgue measure zero, by the Lebesgue decomposition theorem (see, e.g., [10], p.42). We then denote

$$\overline{(\nabla h - F)} := \theta H(s) \mu_s.$$

We have the following analogue of Theorem 3.2 when $F \in \mathcal{DM}^{ext}(\Omega)$.

**Theorem 3.4.** Let $\Omega$ be a bounded open set with a deformable Lipschitz boundary and $F \in \mathcal{DM}^{ext}(\Omega)$. Let $\Psi : \partial \Omega \times [0, 1] \to \Omega$ be a Lipschitz deformation of $\partial \Omega$. Then, for almost all $s \in [0, 1]$, and all $\phi \in C_0^\infty(\mathbb{R}^N)$,

$$\int_{\Omega_s} \phi \text{div} F = \int_{\partial \Omega_s} \phi(\omega) \overline{(\nabla h - F)} - \int_{\Omega_s} \nabla \phi(x) \cdot F.$$

Similarly to what was done for $\mathcal{DM}^1$-fields, from Theorem 3.4 we get the following result.

**Theorem 3.5.** Let $F \in \mathcal{DM}^{ext}(\Omega)$, where $\Omega$ is a bounded open set with a Lipschitz boundary admitting a regular deformation $\Psi : \partial \Omega \times [0, 1] \to \Omega$. Denoting by $F \cdot \nu|_{\partial \Omega}$ the continuous linear functional $\text{Lip}(\partial \Omega) \to \mathbb{R}$ given by the normal trace of $F$ at $\partial \Omega$, we have the formula

$$F \cdot \nu|_{\partial \Omega} = \text{ess, lim}_{s \to 0} \Psi_s^* d\overline{(\nabla h - F)}_s,$$
with equality in the sense of \((\text{Lip}(\partial \Omega))^*\), where on the right-hand side the functionals are given by the pull back by \(\Psi_s\) of the measures \(d(\nabla h \cdot F)_s\), resulting from the disintegration of \(d(\nabla h \cdot F)\).

4. Application to time-regularity of entropy solutions to hyperbolic conservation laws. Let \(n, d \in \mathbb{N}, \mathbb{R}^{d+1}_+ = \mathbb{R}^d \times (0, \infty)\), and \(\mathcal{U} \subset \mathbb{R}^n\) be an open and convex set. We consider the \(N\)-dimensional system of \(n\) conservation laws

\[
\partial_t U + \sum_{\alpha=1}^d \partial_\alpha F^\alpha(U) = 0, \quad \text{in } \mathbb{R}^{N+1}_+,
\]

with \(U(x,t) \in \mathcal{U}\) and \(F^\alpha : \mathcal{U} \to \mathbb{R}^n\), where \(\partial_\alpha\) denotes the partial derivative with respect to \(x_\alpha\).

Together with (23), we consider the initial data

\[
U(x,0) = U_0(x).
\]

The following result provides a time-regularity information about entropy solutions of the problem (23)-(24). It extends a result established in [8] (theorem 4.5.1), which follows from the theory for \(L^\infty\) divergence-measure fields.

**Theorem 4.1.** Let \(U_0 \in L^{1}_{\text{loc}}(\mathbb{R}^d)\), and let \(U \in L^{1}_{\text{loc}}(\mathbb{R}^d \times [0, \infty))\) be a weak solution of (23)-(24), in the sense that, for any \(\phi \in C_c(\mathbb{R}^{d+1})\), we have

\[
\int_{\mathbb{R}^{d+1}_+} U(x,t) \partial_t \phi + \sum_{\alpha=1}^d F^\alpha(U) \partial_\alpha \phi \, dx \, dt + \int_{\mathbb{R}^d} U_0(x) \phi(x,0) \, dx = 0.
\]

Let \(\eta : \mathcal{U} \to [0, \infty)\) be a strictly convex function, with \(\eta(U) \geq c_1 |U| + c_2\), for some \(c_1 > 0, c_2 \in \mathbb{R}\), such that \(\eta(U(x,t)) \in L^p(K \cap \mathbb{R}^{d+1}_+):K\) for any compact set \(K \subset \mathbb{R}^{d+1}_+, \) for some \(p > 1\). Suppose that there exists a vector measure \(Q \in \mathcal{M}(K \cap \mathbb{R}^{d+1}_+; \mathbb{R}^d)\), for any compact set \(K \subset \mathbb{R}^{d+1}_+, \) such that \(\eta(U(x,t))\) satisfies

\[
\partial_t \eta(U) + \text{div}_x Q \leq 0, \quad \text{in } \mathbb{R}^{d+1}_+,
\]

in the sense of the distributions, where \(\mathcal{M}(\Omega; \mathbb{R}^d)\) denotes the \(\mathbb{R}^d\)-valued Radon measures with finite total variation on \(\Omega\). Then,

\[
U \in C((0, \infty) \setminus S; L^1_{\text{loc}}(\mathbb{R}^d)),
\]

for some at most countable set \(S \subset (0, \infty)\). Moreover, if we have, for all nonnegative \(\psi \in C_c(\mathbb{R}^{d+1})\),

\[
\int_{\mathbb{R}^{d+1}_+} [\eta(U(x,t)) \partial_t \psi \, dx \, dt + \nabla_x \psi \cdot dQ] + \int_{\mathbb{R}^d} \eta(U_0(x)) \psi(x,0) \, dx \geq 0,
\]

then the above strong continuity holds on the right for \(t = 0\).

5. Acknowledgments. The author gratefully acknowledges the support from CNPQ, through grant proc. 303950/2009-9, and FAPERJ, through grant E-26/103.019/2011.
REFERENCES


E-mail address: hermano@impa.br
MIXED SYSTEMS WITH BOUNDARIES

RAUL BORSCHE
Fachbereich Mathematik
Technische Universität Kaiserslautern, Germany

RINALDO M. COLOMBO
INDAM Unity, University of Brescia, Italy

MAURO GARAVELLO
Department of Mathematics and Applications
University of Milano-Bicocca, Italy

Abstract. We consider a mixed problem, consisting of a system of balance laws with two fixed external boundaries and two further internal moving free boundaries, coupled with an ordinary differential equation related to the movement of the free boundaries. A well posedness result is presented. The motivating application of a pneumatic gas spring is considered in detail through numerical integrations.

1. Introduction. This paper deals with mixed problems consisting of 1D systems of hyperbolic balance laws on a fixed space interval $[0, L]$ coupled with ordinary differential equations. More precisely, we consider the following system

\[
\begin{align*}
\partial_t u^- + \partial_x f^-(u^-) &= g^-(u^-) \quad &0 < x < \gamma^-(t) \\
\partial_t u^+ + \partial_x f^+(u^+) &= g^+(u^+) \quad &\gamma^+(t) < x < L \\
b^-(u^-(t, \gamma^-(t)\cdot), u^+(t, \gamma^+(t)\cdot)) &= B^-(t, w(t)) \\
b^+(u^-(t, \gamma^-(t)\cdot), u^+(t, \gamma^+(t)\cdot)) &= B^+(t, w(t)) \\
b_0^-(w(0)) &= 0 \\
b_L^+(u^+(t, L)) &= 0 \\
\dot{w} &= F(t, u^-(t, \gamma^-(t)\cdot), u^+(t, \gamma^+(t)\cdot), w(t)) \\
\dot{\gamma}^- &= \Pi^-(w(t)) \\
\dot{\gamma}^+ &= \Pi^+(w(t)) .
\end{align*}
\]

Here, $\gamma^\pm$ are free boundaries in the sense that their positions are not known a priori but they have to be determined while solving (1). Precise regularity assumptions on the functions appearing in (1) are presented in Section 2. For completeness, we specify here the function spaces:

- $f^\pm : \mathbb{R}^{n\pm} \to \mathbb{R}^{n\pm}$
- $g^\pm : \mathbf{BV}([0, L]; \mathbb{R}^{n\pm}) \to L^1([0, L]; \mathbb{R}^{n\pm})$
- $B^\pm : \mathbb{R}^+ \times \mathbb{R}^m \to \mathbb{R}^{n_\pm - \ell_\pm}$
- $F : \mathbb{R}^+ \times \mathbb{R}^{n^-} \times \mathbb{R}^{n^+} \times \mathbb{R}^m \to \mathbb{R}^m$
- $\Pi^\pm : \mathbb{R}^m \to \mathbb{R}$
- $b^\pm : \mathbb{R}^{n\pm} \to \mathbb{R}^{n_\pm - \ell_\pm}$
- $b_0 : \mathbb{R}^{n^-} \to \mathbb{R}^\ell$,
- $b_L : \mathbb{R}^{n^+} \to \mathbb{R}^\ell$

Key words and phrases. Conservation Laws, Mixed Systems.
where \( u^\pm \in \mathbb{R}^{n^\pm} \) are the vectors of the unknown conserved densities, \( w \in \mathbb{R}^m \) is the unknown governed by the ordinary differential equation, \( t^\pm \in \{1, \ldots, n^\pm - 1\} \) are the numbers of characteristics entering the domain in \( \gamma^\pm \) and \( t_0 \in \{1, \ldots, n^-\} \), respectively \( t_L \in \{1, \ldots, n^+\} \), are the numbers of characteristics entering the domain in \( x = 0 \) and \( x = L \) respectively.

Below, we state the well posedness of (1), extending the results in [1], where no boundaries at \( x = 0 \) and \( x = L \) were considered. The present setting comprehends the description of a pneumatic gas spring, composed of two cylindrical chambers, filled with a compressible fluid and separated by a piston with a valve. This application is treated in detail in Section 3.

2. Analytical Results. Throughout, we denote \( \mathbb{R}^+ = [0, +\infty[ \) and \( \mathbb{R}^+_0 = ]0, +\infty[ \).

Let \( \Omega^\pm \subseteq \mathbb{R}^{n^\pm} \) be open sets and let \( B_r(w) \) be the open ball centered at \( w \) with radius \( r \).

Fix the reference states \( \hat{u}^\pm \in \Omega^\pm, \hat{w} \in \mathbb{R}^m \) and \( \hat{x}^\pm \in ]0, L[ \). We refer to [3, 6] for the classical vocabulary about conservation and balance laws. For \( u^\pm \in \Omega^\pm \) and \( i = 1, \ldots, n^\pm \), call \( \lambda_i^\pm(u^\pm) \) the \( i \)-th eigenvalue of \( Df_i^\pm(u^\pm) \) and \( \gamma_i^\pm(u^\pm) \) the corresponding right eigenvector. Moreover define

\[
\mathcal{U}^\pm = \left\{ u^\pm \in \text{BV}([0, L]; \mathbb{R}^{n^\pm}): u^\pm([0, L]) \subset \Omega^\pm \right\}
\]

and \( \mathcal{U}_0^\pm = \{ u^\pm \in \mathcal{U}^\pm: TV(u^\pm) \leq \delta \} \) for all positive \( \delta \). As a first step, we formalize what we mean by solution to the Cauchy problem for (1).

**Definition 2.1.** Let \( T > 0 \) be fixed, \( u_0^\pm \in \text{BV}([0, L]; \Omega^\pm), w_o \in \mathbb{R}^m \) and \( x_0^\pm \in [0, L[ \). \( u^\pm \) is a solution to (1) with initial condition \( (u_0^\pm, w_o, x_0^\pm) \) on the time interval \( [0, T] \) if \( u^\pm \in C^0([0, T]; \text{BV}([0, L]; \Omega^\pm)), w \in W^{1,1}([0, T]; \mathbb{R}^m), \gamma^\pm \in W^{1,\infty}([0, T]; [0, L[) \) and

1. \( u^\pm \) is a weak entropy solution to \( \partial_t u^\pm + \partial_x f^\pm(u^\pm) = g^\pm(u^\pm) \) respectively in \( [0, T] \times ]0, \gamma^-(t)\) and in \( [0, T] \times ]\gamma^+(t), L[ \);
2. \( w \) solves \( \dot{w}(t) = F(t, u^-(t), \gamma^-(t)), u^+(t, \gamma^+(t)), w(t) \) in the sense of Carathéodory, for \( t \in [0, T]; \)
3. \( \gamma^\pm \) solve \( \dot{\gamma}^\pm(t) = \Pi^\pm (w(t)) \) for a.e. \( t \in [0, T]; \)
4. For a.e. \( t \in [0, T], \) the boundary conditions

\[
\begin{align*}
 b_0(u^-(t, 0)) &= 0, & b^-(u^-(t, \gamma^-(t) -), u^+(t, \gamma^+(t) +)) &= B^-(t, w(t)), \\
 b_L(u^+(t, L)) &= 0, & b^+(u^-(t, \gamma^-(t) -), u^+(t, \gamma^+(t) +)) &= B^+(t, w(t)),
\end{align*}
\]

are satisfied in the sense of the traces;
5. The initial condition is satisfied in the strong sense: \( u^\pm(0, x) = u_o(x) \) for a.e. \( x \in [0, L[ \), \( \delta^\pm(0) = x_0^\pm \).

We introduce the following assumptions.

**f:** \( f^\pm \in C^\infty(\Omega^\pm; \mathbb{R}^{n^\pm}) \) is smooth and such that, for all \( u^\pm \in \Omega^\pm, \) \( Df_i^\pm(u^\pm) \) is strictly hyperbolic and each characteristic field is either genuinely nonlinear or linearly degenerate. Hence we may assume that \( \lambda_i^{n-1}(u^\pm) < \lambda_i^{n+1}(u^\pm) \) for all \( u^\pm \in \Omega^\pm \) and \( i = 2, \ldots, n^\pm \).

**g:** For \( \delta_o > 0, g^\pm: U_{\delta_o} \to L^1([0, L]; \mathbb{R}^{n^\pm}) \) is such that for suitable \( L_1, L_2 > 0, \forall u^\pm, \tilde{u}^\pm \in U_{\delta_o}, \)

\[
\left\| g^\pm(u^\pm) - g^\pm(\tilde{u}^\pm) \right\|_{L^1} \leq L_1 \left\| u^\pm - \tilde{u}^\pm \right\|_{L^1} \quad \text{and} \quad TV\left(g^\pm(u^\pm)\right) \leq L_2.
\]
The following result about the well-posedness of (1) holds.

(II): \( \Pi^+ \in C^{0,1}(\mathbb{R}^m; \mathbb{R}) \).

(NC): There exists \( c > 0 \) such that

\[
\begin{align*}
\lambda_{n^- \to +}(\tilde{u}^-) &< -c \quad \text{and} \quad \lambda_{n^- \to +1}(\tilde{u}^-) > c \\
\lambda_{n^- \to \ell}(\tilde{u}^-) &< \Pi^- (\tilde{u}^-) - c \quad \text{and} \quad \lambda_{n^- \to \ell+1}(\tilde{u}^-) > \Pi^- (\tilde{u}^-) + c \\
\lambda_{\ell+}(\tilde{u}^+) &< \Pi^+ (\tilde{u}^+) - c \quad \text{and} \quad \lambda_{\ell+1}(\tilde{u}^+) > \Pi^+ (\tilde{u}^+) + c \\
\lambda_{\ell_L}(\tilde{u}^+) &< -c \quad \text{and} \quad \lambda_{\ell_L+1}(\tilde{u}^+) > c.
\end{align*}
\]

(b): \( b_0 \in C^1(\Omega^-; \mathbb{R}^m), b^+ \in C^1(\Omega^- \times \Omega^+; \mathbb{R}^{n^- \times \ell}) \) and \( b_L \in C^1(\Omega^+; \mathbb{R}^{\ell_L}) \) are such that

\[
\begin{align*}
\det \left( D b(\bar{u}^-) \begin{bmatrix} r_{n^- \to \ell+1}(\tilde{u}^-) & r_{n^- \to \ell+2}(\tilde{u}^-) & \cdots & r_{n^-}(\tilde{u}^-) \end{bmatrix} \right) &\neq 0 \\
\det \left( D u^- b^- (\tilde{u}^-, \tilde{u}^+) \begin{bmatrix} r_1^- (\tilde{u}^-) & r_2^- (\tilde{u}^-) & \cdots & r_{n^- - \ell^-}(\tilde{u}^-) \end{bmatrix} \right) &\neq 0 \\
\det \left( D u^+ b^+ (\tilde{u}^-, \tilde{u}^+) \begin{bmatrix} r_{\ell+1}(\tilde{u}^+) & r_{\ell+2}(\tilde{u}^+) & \cdots & r_{n^+}(\tilde{u}^+) \end{bmatrix} \right) &\neq 0 \\
\det \left( D b^+ (\tilde{u}^+) \begin{bmatrix} r_1^+ (\tilde{u}^+) & r_2^+ (\tilde{u}^+) & \cdots & r_{\ell_L}(\tilde{u}^+) \end{bmatrix} \right) &\neq 0.
\end{align*}
\]

(F): The map \( F: \mathbb{R}^+ \times \mathbb{R}^- \times \mathbb{R}^+ \times \mathbb{R}^m \to \mathbb{R}^m \) is such that

(F.1): For all \( u^\pm \in \Omega^\pm \) and \( w \in \mathbb{R}^m \), the function \( t \mapsto F(t, u^-, u^+, w) \) is Lebesgue measurable.

(F.2): For all compact subset \( K \) of \( \Omega^- \times \Omega^+ \times \mathbb{R}^m \), there exists \( C_K > 1 \) such that for all \( t \in \mathbb{R}^+ \) and \((u_1^- + u_1^+, w_1), (u_2^- + u_2^+, w_2) \) \( \in K \)

\[
\begin{align*}
\left\| F(t, u_1^-, u_1^+, w_1) - F(t, u_2^-, u_2^+, w_2) \right\|_{\mathbb{R}^m} &
\leq C_K \left( \| u_1^- - u_2^- \|_{\mathbb{R}^n^-} + \| u_1^+ - u_2^+ \|_{\mathbb{R}^{n^+}} + \| w_1 - w_2 \|_{\mathbb{R}^m} \right).
\end{align*}
\]

(F.3): There exists a function \( C \in L^1_{int}(\mathbb{R}^+; \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^m) \) such that for all \( t > 0, t^+ \in \Omega^\pm \) and \( w \in \mathbb{R}^m \) we have \( \| F(t, u^-, u^+, w) \|_{\mathbb{R}^m} \leq C(t) \left( 1 + \| w \|_{\mathbb{R}^m} \right) \).

(B): \( B^\pm \in C^1(\mathbb{R}^+ \times \mathbb{R}^m; \mathbb{R}^{n^- \times \ell^\pm}) \) are locally Lipschitz, i.e. for every compact subset \( K \) of \( \mathbb{R}^m \), there exists a constant \( C_K > 0 \) such that, for every \( t > 0 \) and \( w \in K \):

\[
\left\| \frac{\partial}{\partial t} B^\pm (t, w) \right\|_{\mathbb{R}^m} + \left\| \frac{\partial}{\partial w} B^\pm (t, w) \right\|_{\mathbb{R}^m} \leq C_K.
\]

The following result about the well-posedness of (1) holds.

Theorem 2.2. Let (f), (g), (II), (NC), (b), (F) and (B) hold. Assume moreover that \( b^\pm(\bar{u}^-, \bar{u}^+) = B^\pm(0, \bar{w}) \). Then, there exist positive \( \delta, \Delta, L, T \), domains \( \hat{D}_t \) (for \( t \in [0, T] \)) and maps \( \hat{P}(t, t_0): \hat{D}_{t_0} \to \hat{D}_{t_0 + t} \) (for \( t_0 + t \in [0, T] \)) such that if \( \| b_0(\bar{u}^-) \| + \| b_L(\bar{u}^+) \| < \delta \),

1. \[
\begin{align*}
\left( U_\delta \times U_\delta^+ \times B_\delta(\bar{w}) \right) \times \bar{\mathbb{R}} \subseteq & \hat{D}_t \\
\left( U_\Delta \times U_\Delta^+ \times B_\Delta(\bar{w}) \right) \times \mathbb{R} \subseteq & \hat{D}_t;
\end{align*}
\]

2. for all \( t_0, t_1, t_2 \) with \( t_0, t_1 \in [0, T] \), \( t_1 \in [0, T - t_0] \) and \( t_2 \in [0, T - t_0 - t_1] \), then \( \hat{P}(t_2, t_0 + t_1) \circ \hat{P}(t_1, t_0) = \hat{P}(t_1 + t_2, t_0) \) and \( \hat{P}(0, t_0) = Id \);

3. for \( t_0 \in [0, T] \), \( t \in [0, T - t_0] \), and \( (u^-, u^+, w, x), (\bar{u}^-, \bar{u}^+, \bar{w}, \bar{x}) \in \hat{D}_{t_0} \)

\[
\left\| \hat{P}(t, t_0)(u^-, u^+, w, x) - \hat{P}(t, t_0)(\bar{u}^-, \bar{u}^+, \bar{w}, \bar{x}) \right\|_{L^1 \times L^1 \times \mathbb{R}^m \times \mathbb{R}}
\]
\[ L \left( \| u^- - \bar{u}^\pm \|_{L^1} + \| u^+ - \bar{u}^\pm \|_{L^1} + \| w - \bar{w} \|_{R^m} + |x - x^\pm| \right); \]

4. for all \((u_0^-, u_0^+, w_0, x_0) \in \tilde{D}_0\), the map \(t \mapsto \tilde{P}(t,0)(u_0^-, u_0^+, w_0, x_0)\), defined for \(t \in [0,T_\delta]\), solves (1) in the sense of Definition 2.1.

The proof is a modification of that of [1, Theorem 2.6 and Theorem 2.7] and it is based on a well posedness result for solutions to balance laws with two boundaries. This consists just in a technical improvement of [4, Theorem 3.2] to the case of a system of balance laws with two boundaries. Remark that the choice of \(T_\delta\) in Theorem 2.2 is also subjected to the \(L^\infty\) norm of \(\Pi^\pm\) and to \(x^\pm_0\), since it is important to avoid that the moving boundaries \(\gamma^\pm\) reach \(x = 0\) or \(x = L\).

3. A model for a pneumatic gas spring and numerical simulations. We consider a model for a pneumatic gas spring, composed of two cylindrical chambers, filled with a compressible fluid and separated by a piston with a valve, see Figure 1. Here \(x\) denotes the coordinate along the tube, modeled by the real interval \([0,L]\).

![Figure 1. Schematic view of a gas spring. Note that due to the rod, above the piston the surface available to the fluid is less than below the piston.](image)

The function \(X(t) \in [0,L]\) gives the position of the piston inside the tube at time \(t\). The upper chamber has a cross sectional area \(A^+\), which is smaller than the lower one \(A^-\), since the rod of the piston has a not negligible size. This difference in the cross section plays a key role for the characteristics of the spring. Therefore, denoting by \(A = A(t,x)\) the section of the tube, we have

\[ A(t,x) = A^- \chi_{[0,X(t)]}(x) + A^+ \chi_{[X(t),L]}(x). \]
Recall the dynamic pressure, the total energy density and the energy flow related to (3), respectively given by:

\[ P(\rho, q) = \frac{q^2}{\rho} + p(\rho) \]
\[ E(\rho, q) = \frac{q^2}{2\rho} + \rho \int_{\rho_s}^\rho \frac{p(r)}{r^2} \, dr \]  
\[ F_E(\rho, q) = \frac{q}{\rho} \left( E(\rho, q) + p(\rho) \right), \]

where \( \rho_s \) is a fixed reference density. Introduce the total mass, the total linear momentum and the total energy

\[ M(t) = \int_0^L A(t, x) \rho(t, x) \, dx + m \]
\[ \mathcal{Q}(t) = \int_0^L A(t, x) q(t, x) \, dx + m \dot{X}(t) \]  
\[ E(t) = \int_0^L A(t, x) \left( E \left( \rho(t, x), q(t, x) \right) + \rho(t, x) g x \sin \theta \right) \, dx \]
\[ + m g \sin \theta \, X(t) + \frac{1}{2} m \dot{X}^2(t), \]

where \( m \) denotes the mass of the piston. Along smooth solutions, we prescribe the conservation of the total mass, the balance of the linear momentum with the gravitational impulse and the conservation of energy, namely:

\[ \dot{M}(t) = 0 \]
\[ \dot{\mathcal{Q}}(t) = -M g \sin \theta + A^- p(\rho(t, 0^+)) - A^+ p(\rho(t, L^-)) \]
\[ \dot{E}(t) = 0. \]

For a treatment of non smooth solutions see [2]. The following system then follows:

\[
\begin{cases}
\partial_t \rho + \partial_x q = 0 & t > 0, x \in [0, L], x \neq X(t) \\
\partial_t q + \partial_x \left( \frac{q^2}{\rho} + p(\rho) \right) = -\rho g \sin \theta - \nu_F(\rho, q) & t > 0, x \in [0, L], x \neq X(t) \\
q(t, 0) = 0 & t > 0 \\
q(t, L) = 0 & t > 0 \\
\dot{X} = -g \sin \theta - \frac{1}{m} \left( \Delta(A P) - \dot{X} \Delta(A q) \right) - \nu_S(\dot{X}) + \mathcal{F}(t) \\
\Delta(A q) - \dot{X} \Delta(A \rho) = 0 \\
\Delta(A F_E) - \dot{X} \Delta(A E) = \dot{X} \left( \Delta(A P) - \dot{X} \Delta(A q) \right) - \nu_I(q/\rho, \dot{X}),
\end{cases}
\]

where \( \nu_S \) and \( \nu_I \) represent respectively the friction between the piston and the pipe’s wall and the friction between the piston and the fluid. The load applied to the spring is \( \mathcal{F} \). Above, we used the following notation: for any quantity \( G = G(\rho, q) \) we set

\[ \Delta G(t) = G \left( \rho(t, X(t)^+), q(t, X(t)^+) \right) - G \left( \rho(t, X(t)^-), q(t, X(t)^-) \right). \]

Possible standard choices are, for instance,

\[ \nu_F(\rho, q) = \tilde{\nu}_F \frac{q \rho}{|q|}, \quad \nu_S(\dot{X}) = \tilde{\nu}_S \dot{X} \quad \text{and} \quad \nu_I(q/\rho, \dot{X}) = \tilde{\nu}_I \left( \dot{X} - \left\langle \frac{q}{\rho} \right\rangle \right)^2 \]
for suitable positive constants \( \nu_F, \nu_S \) and \( \nu_I \). Here, we denote with the symbol \( \langle q \rangle \) the quantity \( \frac{1}{2} \left( \frac{q(t,X(t)+)}{\rho(t,X(t)+)} + \frac{q(t,X(t)-)}{\rho(t,X(t)-)} \right) \).

Introduce the function \( \Phi: (\mathbb{R}^+ \times \mathbb{R}^2)^2 \times \mathbb{R} \to \mathbb{R}^2 \) whose components are defined by \( \Phi_1(u^-, u^+, V) = \Delta(A q) - \Delta(A \rho) V \) and \( \Phi_2(u^-, u^+, V) = \Delta(A F_E) - V \Delta(A E) - V (\Delta(A P) - V \Delta(A q)) + \nu_I(q/\rho, V) \) with \( P, E, F_E \) as in (5), so that the last two conditions in (7) can be rewritten as \( \Phi \left( u \left( t, X(t) - \right), u \left( t, X(t) + \right), X(t) \right) = 0. \)

**Proposition 1.** Let (\( \mathbf{p} \)) hold. Fix the positive parameters \( m, g \) and \( \theta \) and choose \( \nu_I \in \mathcal{C}^1(\mathbb{R} \times \mathbb{R}; \mathbb{R}) \). Assume there exist states \( \bar{u}^\pm \in \bar{\mathbb{R}}^+ \times \mathbb{R}, \) a position \( \bar{X} \in \bar{\mathbb{R}}^+ \) and a speed \( \bar{V} \in \mathbb{R} \) such that \( \Phi(\bar{u}^-, \bar{u}^+, \bar{V}) = 0. \) Assume moreover that

\[
\text{det} \left[ \begin{array}{cc}
D_u^- \Phi(\bar{u}^-, \bar{u}^+, \bar{V}) r_1(\bar{u}^-) & D_u^+ \Phi(\bar{u}^-, \bar{u}^+, \bar{V}) r_2(\bar{u}^+)
\end{array} \right] \neq 0, \tag{9}
\]

\[
\left( \begin{array}{c}
D_u^- \Phi(\bar{u}^-, \bar{u}^+, \bar{V}) r_1(\bar{u}^-), D_u^+ \Phi(\bar{u}^-, \bar{u}^+, \bar{V}) r_2(\bar{u}^+)
\end{array} \right) \bar{V} = 0. \tag{10}
\]

Then, there exists a \( \delta > 0 \) such that for all \( u^\pm \in B(\bar{u}^\pm, \delta) \) and \( V \in ]\bar{V} - \delta, \bar{V} + \delta[ \) the two sets of conditions

\[
\Phi(u^-, u^+, V) = 0 \quad \text{and} \quad \left\{ \begin{array}{c}
b^- (u^-, u^+) = V \\
b^+ (u^-, u^+) = V
\end{array} \right.
\]

are equivalent, for suitable smooth functions \( b^\pm \) that also satisfy

\[
\nabla_u b^-(u^-, u^+) r_1(\bar{u}^-) \neq 0, \quad \text{and} \quad \nabla_u b^+(u^-, u^+) r_2(\bar{u}^+) \neq 0.
\]

The proof is the same as that of [2, Proposition 2.5], hence we omit it.

**Proposition 2.** Let (\( \mathbf{p} \)). Fix the parameters \( m, g \in \bar{\mathbb{R}}^+ \) and \( \theta \in [0, \pi/2] \). Assume there exist states \( \bar{u}^\pm \in \bar{\mathbb{R}}^+ \times \mathbb{R}, \) a position \( \bar{X} \in \bar{\mathbb{R}}^+ \) and a speed \( \bar{V} \in \mathbb{R} \) such that \( \Phi(\bar{u}^-, \bar{u}^+, \bar{V}) = 0. \) Assume moreover that (9) and (10) hold. Then, system (7) can be rewritten in the form (1), where

\[
\begin{align*}
\nu_I & \geq 0 \\
\nu_I & > 0
\end{align*}
\]

and with \( b^\pm \) given by Proposition 1. Hence, under the subsonic conditions

\[
|\bar{V}| < \sqrt{p'(\bar{\rho}^-)} \quad \text{and} \quad |\bar{V}| < \sqrt{p'(\bar{\rho}^+)} , \tag{11}
\]

Theorem 2.2 can be applied.

To illustrate some qualitative features of the model (7), we choose the usual \( \gamma \)-pressure law

\[
p(\rho) = p_o \left( \frac{\rho}{\rho_o} \right)^\gamma , \tag{12}
\]
the friction terms as in (8), the parameters

\[ \begin{align*}
L &= 0.2 \text{m} \\
\rho_o &= 200 \text{ Kg/m}^2/\text{s}^2 \\
\rho_o &= 1 \text{ Kg/m} \\
\gamma &= 7/5 \\
\theta &= \frac{\pi}{2} \\
\bar{\nu}_F &= 10^{-6} \text{ m/s}^{-1} \\
\bar{\nu}_S &= 0.1 \text{ s}^{-1} \\
\bar{\nu}_I &= 100 \text{ m}^2\text{ Kg/s} \\
F &= -39.24 \text{ N}
\end{align*} \]  

and the initial data

\[ \begin{align*}
\rho_o(x) &= \begin{cases} 
1 \text{ Kg/m} & x < X_o \\
0.8^{5/7} \text{ Kg/m} & x > X_o 
\end{cases} \\
q_o(x) &= 0 \text{ Kg/s} \\
V_o &= 0 \text{ m/s}.
\end{align*} \]  

To obtain the figures below, the balance laws are solved by means of a wave–propagation method for a moving mesh, see [7]. The ordinary differential equation and the source term are integrated using a two stages Runge–Kutta method. The two systems are coupled using the Strang splitting [8, §17.4] in order to maintain the second order accuracy. Here, \( N = 2000 \) grid points are used, the CFL number is 0.9 and the time of integration runs from 0 to 0.025 s.

At time \( t = 0 \), the fluid is at rest, with the same hydrostatic pressure on the two sides of the piston. The external force \( F \) pushes the piston downwards, causing a 1-compression wave below it and a 2-rarefaction above it, see Figure 2. The former wave hits the bottom boundary at about \( t = 0.005 \text{s} \) and bounces upwards as a 2-compression wave. At about \( t = 0.011 \text{s} \), this 2-wave hits the piston, is partly reflected and partly refracted. This interaction also causes that the traces of the fluid speed at the piston become different from the piston speed, see Figure 3, right.

Due to the difference in the cross sections and according to mass conservation, the two traces are different from each other. Remarkably, at this point the solution to [1, Formula (1.2)] ceases to exist, while the present model (7) and the corresponding Definition 2.1 comprise the case of mass transfer between the two sides of the piston. Note that the traces of the fluid speed at the piston may differ, coherently with the conservation of mass, due to the difference in the surface sections \( A^+ \) and \( A^- \).
Later, the repeated interactions between the piston and the fluid cause the piston to accelerate upwards, see Figure 3, left.

At about $t = 0.019$ s, a 2-shock hits the piston coming from below, see Figure 2, and passes through it. As a result, both traces of the fluid speed at the piston suffer an upward jump, see Figure 3, right.

**REFERENCES**


E-mail address: borsche@mathematik.uni-kl.de
E-mail address: rinaldo@ing.unibs.it
E-mail address: mauro.garavello@unimib.it
DEGENERATE CONVECTION-DIFFUSION EQUATION WITH A
ROBIN BOUNDARY CONDITION

MOHAMED GAZIBO KARIMOU
Laboratoire de Mathématiques, CNRS: UMR 6623
Université de Franche-Comté, 16, route de Gray, 25030 Besançon, France

ABSTRACT. We study a Robin boundary problem for degenerate parabolic equation. We suggest a notion of entropy solution and propose a result of existence and uniqueness. Numerical simulations illustrate some aspects of solution behavior.

1. Introduction. Let Ω be an open bounded domain of \( \mathbb{R}^L \) with a Lipschitz boundary \( \partial \Omega \), and \( \eta \) the unit normal to \( \partial \Omega \) outward to \( \Omega \). The purpose of this paper is to discuss existence and uniqueness of entropy solution for the following initial boundary value problem

\[
(P) \begin{cases} 
  u_t + \text{div} f(u) - \Delta \phi(u) = 0 & \text{in } Q = [0,T] \times \Omega, \\
  u(0,x) = u_0(x) & \text{in } \Omega, \\
  b(u) - (f(u) - \nabla \phi(u)).\eta = 0 & \text{on } \Sigma = [0,T] \times \partial \Omega.
\end{cases}
\]

Here, \( u_0 \) is taking values on \([0,u_{\max}]\) for some \( u_{\max} > 0 \). Further, the function \( f \) is a Lipschitz continuous function. Moreover, we require that

\[ f(0) = 0 \text{ and } b(0) = 0. \quad (H1) \]

The diffusion term \( \phi \) is a continuous function. We consider that there exist a critical value \( u_c \) of the unknown \( u \) such that: \( \phi(.) \) is zero on \([0,u_c]\) with \( 0 \leq u_c \leq u_{\max} \) and \( \phi(.) \) is strictly increasing else. Then problem \((P)\) degenerates to hyperbolic when \( u \) takes values in the region \([0,u_c]\) where \( \phi \) is flat.

We suppose that the function \( b \) is a continuous non-decreasing function on \( \Sigma \). In some situation, \( b \) may be a maximal monotone graph on \( \mathbb{R} \) (see [4]). Here, we assume also that \( b \) satisfies the following hypotheses:

\[ b = \beta \circ \phi \text{ where } \beta \text{ is a non-decreasing Lipschitz continuous function.} \quad (H2) \]

\[ b(u_{\max}) \geq |f(u_{\max}).\eta|. \quad (H3) \]

For more than a few decades, the degenerate parabolic equation in bounded domain was studied by many authors mainly in the case of Dirichlet boundary conditions (see e.g. [10], [8]). The zero-flux boundary condition is studied in [1] for non-degenerate parabolic case, in [7] for fully degenerate hyperbolic equation and recently in [2] for the parabolic-hyperbolic problem. Remark, that the condition \( b(u) - (f(u) - \nabla \phi(u)).\eta = 0 \) on \( \Sigma \) includes in particular Neumann (zero-flux) condition on the boundary.

2000 Mathematics Subject Classification. Primary: 35F31; Secondary: 00A69.
Key words and phrases. Degenerate parabolic equation, Robin boundary condition, vanishing viscosity approximation, entropy solution, semigroup theory.

The author is supported by ANR CoToCoLa (Contemporary Topics on Conservation Laws).
We propose an adequate entropy formulation for problem \((P)\) which incorporates two boundary integrals. In [2], existence and uniqueness for the zero flux boundary condition were proved, under the assumption \((H3)\) that reads \(f(u_{\text{max}}) = 0\) in the zero-flux case \(b \equiv 0\). In contrast to the entropy formulation in [2], where the passage to the limit in the only boundary integral is straightforward, for our entropy inequality, we need the assumption \((H2)\), which permits to give a sense to the boundary integral with the term \(b(u)\). Indeed, we can deduce that \(b(u)\) has a trace on the boundary as a function in Sobolev space \(H^1(\Omega)\).

The proof of existence of our entropy solution for any space dimensions \(\ell \geq 1\) employs a vanishing viscosity approximation. We pass to limit in the interior of the domain \(Q\), by using the local compactness result of Panov [12], for this we suppose some relation between \(f\) and \(\phi\) (see Definition 3.5). One can refer to [2] for more details. We pay particular attention to the boundary term (here \((H2)\) is needed).

For the uniqueness result, we use nonlinear semigroup techniques (see, e.g., [6]) and Kruzhkov doubling of variables methods. The main goal is to compare two solutions of \((P)\), and it turns out that it is simpler to compare a solution of \((P)\) with a regular solution (in the sense that the total flux is continuous up to the boundary) of the stationary problem associated to \((P)\). Then we prove that entropy solution of \((P)\) is an integral solution, and we refer to the uniqueness of integral solutions granted by the general theory of nonlinear semigroup. Unfortunately, we are not able to obtain regular solution to the stationary problem for any space dimensions, but only in one space dimension. Then, we can deduce the uniqueness just now when \(\Omega\) is a bounded open interval of \(\mathbb{R}\). Notice that, for the same argument as for the zero-flux boundary condition [2], the problem of uniqueness is still open in multiple space dimensions.

The paper is organized as follows. In the next section, we give our definition of entropy solution and state some remarks useful for the well-posedness. In section 3, we prove existence result of entropy solution. In the section 4, we prove uniqueness in the case of one space dimension. The latter part is devoted to the numerical investigation of problem \((P)\). We adapt the approach of finite volumes in the spirit of Vovelle ([11]) to illustrate and interpret some observations in the case where the assumptions \((H2)\) and \((H3)\) are absent. Thereby, we justify the importance these assumptions in this paper.

2. Notion of entropy solution. Consider the following notion.

**Definition 2.1.** A measurable function \(u\) taking values on \([0, u_{\text{max}}]\) is called entropy solution of problem \((P)\) if \(\phi(u) \in L^2(0, T; H^1(\Omega))\), \(b(u) \in L^2(0, T; H^1(\Omega))\) and the following conditions hold:

\[
\forall k \in [0, u_{\text{max}}], \forall \xi \in C_0^\infty([0, T] \times \mathbb{R}^\ell), \text{ with } \xi \geq 0:
\]

\[
\int_0^T \int_\Omega \left\{ |u - k| \xi_t + \text{sign}(u - k) \left( f(u) - f(k) - \nabla \phi(u) \right) \cdot \nabla \xi \right\} dx dt + \int_\Omega |u_0 - k| \xi(0, x) dx + \int_0^T \int_{\partial \Omega} |f(k) \eta(x) - b(k)| \xi(t, x) d\mathcal{H}^{\ell-1} dt - \int_0^T \int_{\partial \Omega} |b(u) - b(k)| \xi(t, x) d\mathcal{H}^{\ell-1} dt \geq 0.
\]

(1)

Here \(\mathcal{H}\) represents the \((\ell - 1)-\) dimensional Hausdorff measure on \(\partial \Omega\).
Remark 1. 1. The entropy solution in the sense of Definition 2.1 is in particular a weak solution. Indeed, first take in inequality (1), $k = 0$ and use (H1). Next, take $k = u_{\max}$ and use (H3).

2. Let us stress that, in particular, the boundary condition $(f(u) - \nabla \phi(u)).\eta = b(u)$ is verified literally in the weak sense as in the case of zero flux boundary condition (see [2]). This contrasts with the properties of the Dirichlet problem (see [5]); we expect that the boundary condition should be relaxed if assumption (H3) is dropped (see [4, 3] and also numerical tests of section 5).

3. The integral in the boundary term is well defined due to the hypothesis (H2).

In general, uniqueness for evolution equation of kind (P) appear very difficult mainly for the initial boundary values problems. In this context, the use of nonlinear semigroup techniques offers many advantages. Let us present briefly another notion of solution coming from the theory of nonlinear semigroups (see, e.g., [6]).

Definition 2.2. Let $A$ be an $m$-accretive operator (see, e.g., [6]). Suppose that $h \in L^1(Q)$, $u_0 \in L^1(\Omega)$. A measurable function $v \in C([0, T]; L^1(\Omega; [0, u_{\max}]))$\footnote{Here, we will write $L^1(\Omega; [0, u_{\max}])$ for the set of all measurable functions from $\Omega$ to $[0, u_{\max}].$} is an integral solution of the abstract evolution problem

$$v_t + A(v) \ni h(t), \quad v(t = 0) = u_0,$$

if $v(0, .) = u_0(,)$ and for all $(u, z) \in A$

$$\frac{d}{dt} ||v(t) - u||_{L^1(\Omega)} \leq \int_{\Omega} sign_0(v(t) - u)(h(t) - z) + \int_{\{v = u\}} |h(t) - z| \text{ in } \mathcal{D}'(0, T).$$

We will see that entropy and integral solution coincide in the case $\Omega = (a, b)$ an interval of $\mathbb{R}$.

3. Existence of entropy solution. The main result of this part is the following:

Theorem 3.1. Let $\ell \geq 1$. Assume that (H1), (H2) and (H3) holds. Suppose that $(f, \phi)$ is non-degenerate (in the sense of Definition 3.5 below). Then, there exists an entropy solution $u$ for the problem (P).
To show the existence of entropy solutions, we approximate \( \phi(u) \) by \( \phi_{\varepsilon}(u') = \phi(u') + \varepsilon \text{Id}(u') \) for each \( \varepsilon > 0 \) and set \( b_{\varepsilon}(u') = \beta \circ \phi_{\varepsilon}(u') \). We obtain the following regularized strictly parabolic problem \( (P_{\varepsilon}) \) with unknown \( u' \):

\[
(P_{\varepsilon}) \begin{cases} 
   u'_t + \text{div } f(u') - \Delta \phi_{\varepsilon}(u') & = 0 \quad \text{in } Q = [0,T] \times \Omega, \\
   u'(0,x) & = u_0(x) \quad \text{in } \Omega, \\
   b_{\varepsilon}(u') - (f(u') - \nabla \phi_{\varepsilon}(u')).\eta & = 0 \quad \text{on } \Sigma = [0,T] \times \partial \Omega,
\end{cases}
\]

where \( (u'_0)_{\varepsilon} \) is a sequence of smooth functions that converges to \( u_0 \) a.e and respects the minimum/maximum values of \( u_0 \).

**Definition 3.2.** Let \( u_0 \) be a measurable \([0,u_{\text{max}}]\)-valued function. A measurable function \( u' \in L^2(0,T;H^1(\Omega)) \) taking values on \([0,u_{\text{max}}]\) is called weak solution of problem \( (P_{\varepsilon}) \) if: \( \forall \theta \in L^2(0,T;H^1(\Omega)) \cap L^\infty(Q) \) such that \( \theta_{\ell} \in L^2(\Omega) \) and \( \theta(T,.):=0 \), one has

\[
\int_0^T \int_{\Omega} \{ u' \theta_t + (f(u') - \nabla \phi_{\varepsilon}(u')).\nabla \theta \} \, dx \, dt + \int_{\Omega} u'_0 \theta(0,x) \, dx \\
- \int_0^T \int_{\partial \Omega} b_{\varepsilon}(u').\eta \partial H^{\ell-1} \, dt = 0. \tag{4}
\]

**Theorem 3.3.** For \( u_0 \in [0,u_{\text{max}}] \), assume \((H1), (H2) \) and \((H3) \) hold. Problem \( (P_{\varepsilon}) \) admits a weak solution \( u' \) which is also an entropy solution. In particular, we have \( 0 \leq u' \leq u_{\text{max}} \). In addition, there exists \( C \) independent on \( \varepsilon \) such that

\[
||\sqrt{\varepsilon} \nabla u'||_{L^2(\Omega)} \leq C; \tag{5}
\]

\[
||\phi_{\varepsilon}(u'||_{L^2(0,T;H^1(\Omega))} \leq C; \tag{6}
\]

\[
||b_{\varepsilon}(u_0)||_{L^1(\Sigma)} \leq C \quad \text{and} \quad \int_{\Omega} u' \, dx \leq C. \tag{7}
\]

This result can be proved, e.g., using Galerkin method (see e.g. [2]).

**Lemma 3.4.** Assume that the sequence \( (\Psi_j)_{j} \) is such that: \( ||\Psi_j||_{L^2(0,T;H^1(\Omega))} \leq C \) and \( \Psi_j \rightarrow \Psi \) in \( L^2(Q) \). Then \( \gamma \Psi_j \rightarrow \gamma \Psi \) in \( L^2(\Sigma) \), where \( \gamma \) is the trace operator.

The proof uses localization to a small neighborhood of \( \Sigma \).

To prove existence of entropy solution, we assume that the couple \((f(.),\phi(.))\) is non-degenerate in the sense of the following definition:

**Definition 3.5.** \((\text{Panov} \ [12])\). Let \( \phi \) be zero on \([0,u_c]\), strictly increasing on \([u_c,u_{\text{max}}]\) and a vector \( f = (f_1,\ldots,f_\ell) \). A couple \((f(.),\phi(.))\) is said to be non-degenerate if, for all \( \xi \in \mathbb{R}^\ell \setminus \{0\} \), the functions \( \lambda \mapsto \sum_{i=1}^\ell \xi_i f_i(\lambda) \) are not affine on the non-degenerate sub intervals of \([0,u_c]\).

**Theorem 3.6.** \((\text{Panov} \ [12])\). Assume that \((f,\phi)\) is non-degenerate in the sense of Definition 3.5. Suppose \( u' \), \( f(u') \), \( \phi_{\varepsilon}(u') \) are equi-integrable locally on \( Q \). Then, there exists subsequence \( (u'_{\varepsilon}) \) that converges in \( L^1_{\text{Loc}}(Q) \).
Proof of Theorem 3.1. (Sketched) The proof of existence of entropy solution uses Theorem 3.6 to justify the passage to the limit in $Q$ (for more details, see [2]) and Lemma 3.4 for boundary integral.

4. Uniqueness result of entropy solution in one space dimension. The main result of this section is the following theorem:

**Theorem 4.1.** Suppose that $\Omega = (a, b)$ is a bounded interval of $\mathbb{R}$, then (P) admits a unique entropy solution.

In order to study uniqueness in the framework of nonlinear semigroup theory, we consider for all bounded function $g$ taking values on $[0, u_{\text{max}}]$, the stationary problem $(S)$ associated to problem $(P)$:

$$
\begin{align*}
(S) \quad & \begin{cases}
    u + \text{div}(f(u) - \nabla \phi(u)) = g & \text{in } \Omega, \\
    b(u) - (f(u) - \nabla \phi(u)).\eta = 0 & \text{on } \partial \Omega.
\end{cases}
\end{align*}
$$

The notion of entropy solution of $(S)$ correspond to the time-independent entropy solution of $(P)$ with source term $g - u$. In the case where $\Omega = (a, b)$ is a bounded interval of $\mathbb{R}$, we have an important result, which states that, the total flux is regular at the points $a$ and $b$. This kind of regularity seem hard to obtain in multiple space dimensions for $(S)$, and even in dimension $\ell = 1$ for $(P)$.

**Proposition 2.** For all measurable function $g$ taking values in $[0, u_{\text{max}}]$ the problem $(S)$ admits a solution $u$ such that $(f(u) - \phi(u))_y$ is continuous up the boundary, i.e., $(f(u) - \phi(u))_y \in C([a, b])$. Moreover, $b(u) - (f(u) - \phi(u))_y.\eta(y)$ is zero at $y = a$ and $y = b$. (Here $\eta(a) = -1$ and $\eta(b) = +1$).

From now, let’s define the operator $A_{f, \phi, b}$ on $L^1$ associated with regular solutions of $(S)$ by its graph:

$$(u, z) \in A_{f, \phi, b} = \{ u \text{ such that } u \text{ is an entropy solution of } (S), \text{ with } g = u + z \}.$$

**Proposition 3.**
1. $A_{f, \phi, b}$ is accretive in $L^1(\Omega)$.
2. For all $\lambda$ sufficiently small, $R(I + \lambda A_{f, \phi, b})$ contains $L^1(\Omega; [0, u_{\text{max}}])$.
3. $\overline{D(A_{f, \phi, b})} = L^1(\Omega; [0, u_{\text{max}}])$.

For the proof of this proposition, we can refer to [2]. According to the general results of [6], it follows existence and uniqueness of integral solution in the sense of Definition 2.1:

**Corollary 1.** Let $\Omega = (a, b)$, $u_0, \hat{u}_0 \in L^1(\Omega)$ and $h, \hat{h} \in L^1(Q)$. Let $v, \hat{v}$ be integral solutions of $(3)$ (with operator $A_{f, \phi, b}$) associated with the data $(u_0, h)$ and $(\hat{u}_0, \hat{h})$, respectively. Then for a.e. $t \in [0, T]$.

$$
||v(t) - \hat{v}(t)||_{L^1} \leq ||u_0 - \hat{u}_0||_{L^1} + \int_0^t ||h\tau - \hat{h}\tau||_{L^1} dt.
$$

Adapted to our case, we have the following result

**Theorem 4.2.** Let $\Omega = (a, b)$. Let $v$ be an entropy solution of $(P)$ and $u$ be an entropy solution of $(S)$. Then

$$
\frac{d}{dt}||v(t) - u||_{L^1(\Omega)} \leq \int_\Omega \text{sign}(v - u)(u - g) dx \text{ in } D'(0, T). \tag{8}
$$

In particular, $v$ is an integral solution of $(3)$ with $h = 0$. 
Proof of Theorem 4.2 and Theorem 4.1. We consider \( v = v(t, x) \) an entropy solution of \((P)\) and \( u = u(y) \) an entropy solution of \((S)\). Consider nonnegative function \( \xi = \xi(t, x, y) \) having the property that \( \xi(x, y) \in C^\infty((0, T) \times \bar{\Omega}) \) for each \( y \in \bar{\Omega} \), \( \xi(t, x, .) \in C^\infty(\bar{\Omega}) \) for each \((t, x) \in [0, T) \times \bar{\Omega} \). Apply the doubling of variables [9] in the spirit of [2], we obtain this following inequality

\[
\int_0^T \int_{\Omega} |v - u|\xi dydxdt + \int_0^T \int_{\Omega} |v_0 - u|\xi(0, x, y)dxdy \\
+ \int_0^T \int_{\Omega} \int_{\Omega} sign(v - u)\left[(f(v) - \phi(v)x) - (f(u) + \phi(u)y)\right] \xi dx dy dt \\
+ \int_0^T \int_{\Omega} \int_{\Omega} |b(u) - (f(u) - \phi(u)y)\eta(x)|\xi dy dx dt \\
+ \int_0^T \int_{\Omega} \int_{\Omega} |(v - (f(v) - \phi(v)x)\eta(y)|\xi dx dy t \\
\geq \int_0^T \int_{x \in \Omega} \int_{y \in \partial \Omega} |b(u) - b(v)|\xi dx dy dt + \int_0^T \int_{y \in \Omega} \int_{x \in \partial \Omega} |b(u) - b(v)|\xi dx dy dt \\
+ \lim_{\sigma \to 0} \frac{1}{\sigma} \int_0^T \int_{\Omega} \int_{\Omega \times \Omega} \{ -\sigma < \phi(v) - \phi(u) \} \phi(v)x - \phi(u)y|^2\xi dx dy dt \geq 0. \tag{9}
\]

Next, following the idea of [1], we take the test function \( \xi(t, x, y) = \theta(t)\rho_n(x, y) \), where \( \theta \in C_0^\infty(0, T), \theta \geq 0, \rho_n(x, y) = \delta_n(\Delta) \) and \( \Delta = (1 - \frac{1}{n(b-a)})x - y + \frac{a+b}{2n(b-a)} \).

Then, \( \rho_n \in D(\bar{\Omega} \times \Omega) \) and \( \rho_{n\mid x\in\partial\Omega}(x, y) = 0 \). Due to this choice,

\[
\int_0^T \int_{x \in \Omega} \int_{y \in \partial \Omega} |b(v) - (f(v) - \phi(v)x)\eta(y)|\rho_n \theta dy dx dt = 0.
\]

By Proposition 2, \( b(u) - (f(u) - \phi(u)y)\eta(y) \in C_0([a, b]) \). Therefore we have \( |b(u) - (f(u) - \phi(u)y)\eta(x)| \to 0 \) when \( x \to y, i.e., as n \to \infty \). We conclude that

\[
\lim_{n \to \infty} \int_0^T \int_{x \in \partial \Omega} \int_{y \in \Omega} |b(u) - (f(u) - \phi(u)y)\eta(x)|\rho_n \theta dy dx dt = 0.
\]

with the calculation detailed in [2], we deduce that

\[
\int_0^T \int_{\Omega} \theta sign(v - u)\left[(f(v) - \phi(v)x) - (f(u) - \phi(u)y)\right] \xi dx dy dt \to 0.
\]

Hence, we get (8) by passing to the limit in (9) with the above choice of \( \xi \). Thus, the entropy solution \( v \) of the problem \((P)\) is an integral solution of \((3)\). This proves that \( v \) is a unique entropy solution due to Corollary 1. \(\square\)

5. Role of hypotheses (H2), (H3) and some numerical illustrations. The numerical analysis of \((P)\) is not the aim of this paper, although we consider this alternative in a future work. We assume (H1) holds, \( u_c = 0.6 \) and \( u_{\text{max}} = 1 \). We present briefly the importance of the hypotheses (H2), (H3). We apply now the ideas developed e.g., in the work of Vovelle ([11]) to construct a monotone finite
volume scheme which take into account the boundary condition. The interval \([0, 1]\) is divided into \(I\) cells. We initialize the scheme by:

\[
\forall i \in \{1, ..., I\} : u_i^0 = \frac{1}{\delta x} \int_{(i-1)\delta x}^{i\delta x} u_0(x) \, dx,
\]

(10)

the numerical approximation solution at \(t = n\delta t\) in the cell number \(i \in \{2, ..., I - 1\}\) is:

\[
u^{n+1}_i = u^n_i - \frac{\delta t}{\delta x} \left( F(u^n_i, u^n_{i+1}) - F(u^n_{i-1}, u^n_i) - \frac{\phi(u^n_{i+1}) - \phi(u^n_i) + \phi(u^n_{i-1})}{\delta x} \right),
\]

(11)

with the boundary conditions taken into account via

\[
u_{1}^{n+1} = u_{1}^{n} - \frac{\delta t}{\delta x} \left( F(u^n_{1}, u^n_{2}) - \phi(u^n_{2}) + \phi(u^n_{1}) \right) - b(u^n_{1})\]

(12)

\[
u_{I}^{n+1} = u_{I}^{n} - \frac{\delta t}{\delta x} \left( b(u^n_{I}) - F(u^n_{I-1}, u^n_{I}) + \phi(u^n_{I}) - \phi(u^n_{I-1}) \right).
\]

(13)

Here, \(F\) is a numerical flux which we assume monotone, consistent, Lipschitz continuous (see [11]). In the sequel, we take \(u_0(x) = 0.7\) if \(x \in \left[\frac{1}{2}, 1\right]\) and \(u_0(x) = 0\) if \(x \in [0, \frac{1}{2}]\). We take \(\delta x = 0.01\), \(\delta t = \frac{\delta x^2}{5}\), \(\phi(u) = (u - 0.6)^+\) and consider a numerical solution at time \(t = 0.12\). Initially, we remove the hypothesis \((H3)\), by taking \(f(u) = \frac{1}{2} u^2\) and \(b(u) = \phi(u)\), we observe numerically the loss of maximum principle (see Figure 1), this mean that the the solution \(u\) can be greater than \(u_{\text{max}}\). Our entropy formulation requires to choose \(b(u)\) in the functional space that permit to define the trace of \(b(u)\) on the boundary. In the context where assumption \((H2)\) is not taken into account, \(b(u) = u\) and \(f(u) = u(1 - u)1_{[0,1]}\): numerically, we observe a boundary layer (see Figure 2) and this is confirmed by theoretical results of [4]. Now, taking into account assumptions \((H3)\), \((H2)\), with data \(f(u) = u(1 - u)1_{[0,1]}\); \(b(u) = \phi(u)\) the numerical observation shows that the boundary condition at \(x = 0\) and \(x = 1\) is verified literally and the numerical solution respect the maximum principle (see Figure 3).

Figure 1.  
Figure 2.  
Figure 3.
Acknowledgments. I would like to thank Boris Andreianov for his thorough reading and helpful remarks which helped me improve this paper.

REFERENCES


E-mail address: mgazibok@univ-fcomte.fr
EXPLICIT HIGHER ORDER SCHEMES FOR THE COUPLING OF DIMENSIONALLY HETEROGENEOUS FREE-SURFACE FLOW MODELS

CHRISTOPH GERBACHER
Department of Applied Mathematics, University of Freiburg
Hermann-Herder-Str. 10, D-79104 Freiburg, Germany

Abstract. The two-dimensional shallow water equations are widely used in the simulation of shallow free-surface flows. In some situations, even one-dimensional models may be accurate enough to capture the hydrodynamics of interest. In order to reduce overall numerical cost various methods for the numerical coupling of dimensionally heterogeneous models have been proposed in the literature. In this paper, we present a new Discontinuous Galerkin method of potentially arbitrary order for solving coupled 1d-2d shallow water equations based on an a-priori model decomposition. The accuracy of the method as well as its limitations are illustrated in a numerical experiment.

1. Introduction. In the simulation of free-surface flows depth-averaged models like the classical two-dimensional shallow water equations are widely used. They are computationally cheap compared to complex three-dimensional models, yet sufficiently accurate in many practical applications. One-dimensional shallow water models are often used to describe river flows that have a dominant downstream direction. Again, one-dimensional models require lesser computational effort than two-dimensional models. However, their use is somewhat limited. For example, one-dimensional models are incapable of accurately simulating the flooding of river banks.

Even if simplified models may not be accurate enough to capture all flow dynamics of interest, using the most general model available might be computationally inefficient. That is if complex flow behavior is restricted to small portions of the domain only. Multi-model approaches try to overcome these problems by using complex models where necessary and cheap models wherever possible. From a mathematical point of view these regions have to be identified and, if necessary, suitable coupling conditions have to be defined at the model interfaces.

Numerical methods for coupling one- and two-dimensional shallow water equations are presented in [4, 5, 8, 10]. These schemes are based on a Schwarz-like domain decomposition method: at a model interface adjacent solvers are coupled using boundary values or boundary fluxes obtained from the other side and are executed repeatedly in each time step until prescribed matching conditions across the interface are sufficiently satisfied.

In this paper we propose a different approach for the coupling of hierarchically nested one- and two-dimensional shallow water models. We derive a non-uniform...
Figure 1. Computational domain Ω seen from above (left) and cross section $S = S(x,t)$ (right).

Discontinuous Galerkin scheme of potentially arbitrary order to be carried out in the whole computational domain. By construction the method reduces to an essentially one-dimensional scheme in an a-priori fixed 1d flow region. A similar idea for adaptively coupling dimensionally heterogeneous shallow water models by means of second order Finite Volume schemes can be found in [9].

The paper is organized as follows. In Section 2 we present the mathematical models involved in the coupling. The coupled problem based on an a-priori model decomposition will be formulated in Section 3. In Section 4 we derive a non-uniform Discontinuous Galerkin method for the coupled problem. Finally, in Section 5, we study the accuracy and efficiency of our method in a numerical test case.

2. Mathematical models. In this section we briefly recall the time-dependent, non-linear shallow water equations in two and one spatial dimensions. For more detailed information on the derivation and mathematical properties of these systems we refer to [12].

2.1. The two-dimensional model. For the sake of simplicity we consider a straight channel $\Omega = \mathbb{R} \times (-\frac{\delta}{2}, \frac{\delta}{2})$, $\delta > 0$, see Figure 1. Let $T > 0$ be fixed and $\Omega_T = \Omega \times (0, T)$. The 2d shallow water equations read as follows:

$$
\begin{align*}
\partial_t h + \partial_x (hu) + \partial_y (hv) &= 0, \\
\partial_t (hu) + \partial_x (hu^2 + \frac{g}{2} h^2) + \partial_y (huv) &= 0, \\
\partial_t (hv) + \partial_x (huv) + \partial_y (hv^2 + \frac{g}{2} h^2) &= 0,
\end{align*}
$$

in $\Omega_T$. (1)

Here, $h$ denotes the position of the free surface, $(u, v)$ denote the depth-averaged horizontal velocity components along the $(x, y)$-axes, and $g$ is the gravitational constant. We complete the system (1) by initial conditions

$$(h, u, v) = (h_0, u_0, v_0), \quad \text{in } \Omega \times \{0\},$$

and solid wall boundary conditions

$$v\big|_{y=\pm \frac{\delta}{2}} = 0.$$ (3)

We assume that $h > 0$, then the system (1) is strictly hyperbolic.

In compact form the two-dimensional shallow water equations read as

$$\partial_t U + \partial_x F(U) + \partial_y G(U) = 0,$$

in $\Omega_T$, (4)
with \( \mathbf{U} = (h, hu, hv) \), and the analytical fluxes
\[
F(\mathbf{U}) = \begin{pmatrix}
    hu \\
    hu^2 + \frac{g}{2}h^2 \\
    hv
\end{pmatrix}, \quad G(\mathbf{U}) = \begin{pmatrix}
    hv \\
    hvu \\
    hv^2 + \frac{g}{2}h^2
\end{pmatrix}.
\] (5)

### 2.2. The one-dimensional model

For the derivation of the one-dimensional model we assume
i) that \( h, u \) do not depend on \( y \), and
ii) that \( v \equiv 0 \). Note that assumption ii) is consistent with the boundary condition (3).

Employing assumptions i), ii), system (1) simplifies to
\[
\begin{align*}
\partial_t h + \partial_x (hu) &= 0, \\
\partial_t (hu) + \partial_x (hu^2 + \frac{g}{2}h^2) &= 0, \\
\partial_x hv &= 0,
\end{align*}
\] in \( \Omega_T \).

Now integration over the channel width immediately yields the 1d shallow water equations:
\[
\begin{align*}
\partial_t \tilde{S} + \partial_x (\tilde{S}\tilde{u}) &= 0, \\
\partial_x (\tilde{S}\tilde{u}) + \partial_x (\tilde{S}\tilde{u}^2 + \frac{g}{2}\tilde{S}h) &= 0,
\end{align*}
\] in \( \mathbb{R} \times (0, T) \).

Here, \( \tilde{S} \) denotes the wetted cross-sectional area, and \( \tilde{h}, \tilde{u} \) are the averages of \( h \) and \( u \) along the cross-section.

The advantage of system (6) with regard to model coupling is that the original geometric problem framework is preserved: all quantities are defined on the space-time domain \( \Omega_T \). Apart from that, (6) is equivalent to the 1d shallow water equations. Therefore, we will consider (6) as an essentially one-dimensional model.

In the following we call the 2d shallow water equations (1) the fine model and (6) the coarse model. It is easy to see that a solution of the coarse model is a solution of (1) as well. Vice versa, a solution of the fine model solves (6) if and only if the coarse model assumptions are fulfilled. In that case we could pass to the simplified model and would end up with the same solution. Usually a flow will be of coarse type at most in some parts of the computational domain and the behavior of the flow may change with time.

### 3. Model decomposition and coupled problem

We consider an a-priori model decomposition \[10\]: the computational domain is split into several sub domains where either the coarse or the fine model shall be solved.

Again for the sake of simplicity we consider the case of two subregions
\[
\Omega_l = \{(x, y) \in \Omega \mid x < 0\}, \\
\Omega_r = \{(x, y) \in \Omega \mid x > 0\}.
\]

We solve the coarse model (6) in \( \Omega_{lt} \), and the full 2d model (1) in \( \Omega_{rt} \) only. We assume that for all times \( t \in [0, T] \) a flow under consideration meets the assumptions i), ii) made above in \( \Omega_l \).

Let \( \alpha \) be the characteristic function or fine model indicator function
\[
\alpha = \begin{cases}
0, & \text{in } \Omega_l, \\
1, & \text{in } \Omega_r.
\end{cases}
\]

Now set \( \mathbf{U}_\alpha = (h, hu, \alpha hv)^T \). Then the coupled problem simply reads
\[
\partial_t \mathbf{U}_\alpha + \partial_x F(\mathbf{U}_\alpha) + \alpha \partial_y G(\mathbf{U}_\alpha) = 0, \quad \text{in } \Omega_T, \] (7)
with the analytical fluxes $\mathbf{F}, \mathbf{G}$ as defined in (5). The system is completed by initial conditions (2) and boundary conditions (3). Note that admissible initial conditions are restricted by the model assumptions in the coarse model region $\Omega_l$.

4. Discretization. In this section we derive a higher order Discontinuous Galerkin (DG) method for the coupled problem (7). The key idea in our discretization for the coupled problem will be the choice of a spatially heterogeneous, piecewise continuous Finite Element space incorporating the flow structure assumed in the previous section.

4.1. Numerical mesh. In the following let $\mathcal{G} = \{E_i \mid i \in I\}$ be a not necessarily conforming quadrilateral grid covering the computational domain $\Omega$. For each grid element $E \in \mathcal{G}$ the set of intersections is defined as

$$I_E = \{e \mid e = E \cap E', E' \in \mathcal{G}\setminus\{E\}, \mathcal{H}^1(e) \neq 0\},$$

where $\mathcal{H}^1$ denotes the one-dimensional Hausdorff measure.

4.2. Discontinuous Galerkin method. We introduce the space of piecewise polynomial DG ansatz functions

$$V^k_G(\Omega) = \{v \in L^2(\Omega) \mid v|_E \in P^k(E) \text{ for all } E \in \mathcal{G}\},$$

where

$$P^k(E) = \left\{ p(x, y) = \sum_{\alpha = (\alpha_1, \alpha_2)} a_\alpha x^{\alpha_1} y^{\alpha_2} \mid (x, y) \in E, \ 0 \leq \alpha_1, \alpha_2 \leq k \right\}.$$

Note that the number of local degrees of freedom is $\dim(P^k(E)) = (k+1)^2$.

Multiplying (4) by a test function $\varphi \in V^k_G(\Omega) = [V^k_G(\Omega)]^3$ and integration by parts yields the following weak formulation: Find $U(\cdot, t) \in V^k_G(\Omega)$ such that

$$\int_{\Omega} \partial_t U \varphi \, dx = \sum_{E \in \mathcal{G}} \int_E (\mathbf{F}(U) \partial_x \varphi + \mathbf{G}(U) \partial_y \varphi) \, dx$$

$$- \sum_{E \in \mathcal{G}} \sum_{e \in I_E} \int_{e \cap \partial \Omega} \mathcal{F}(U|_E, U|_{E'}) \nu_E \varphi \, d\sigma$$

$$- \sum_{E \in \mathcal{G}} \sum_{e \in I_E} \nu_{E,y} G(U) \varphi \, d\sigma,$$

for all $\varphi \in V^k_G(\Omega)$. Here, $\nu_E = (\nu_{E,x}, \nu_{E,y})$ denotes the unit outer normal with respect to a grid element $E$. Note that on the boundary $\partial \Omega$ we have $\nu = (0, \nu_y)$. Like in Finite Volume schemes numerical fluxes are applied across inter-element discontinuities, denoted here by $\mathcal{F}(\cdot, \cdot)$. The same numerical fluxes as in standard Finite Volume methods can be used in the DG scheme. Admissible numerical fluxes for the shallow water equations can be found for example in [7, 12].

The spatial discretization (8) gives rise to a system of ordinary differential equations with initial conditions (2). We choose an explicit Strong Stability Preserving Runge-Kutta method [6] at least of order $k + 1$ to integrate this system in time.
4.3. DG scheme for the coupled problem. In the derivation of the coarse model (6) we made the following assumptions: i) the functions $h, u$ do not depend on $y$, and ii) $v \equiv 0$. A weak solution $U = (h, hu, hv)$ of (8) is thus a weak solution of the one-dimensional model (6), if $U$ is constant in the second spatial coordinate $y$ and $U_3 = 0$. The idea for the coupled scheme is to choose a subspace $W^k_G(\Omega) \subset V^k_G(\Omega)$ such that a function $w \in W^k_G(\Omega)$ is independent of $y$ in the coarse model region $\Omega_l$, and to consider the restriction of (8) to ansatz functions $\varphi \in W^k_G(\Omega) = [W^k_G(\Omega)]^3$.

We choose a numerical mesh as outlined in Figure 2: in the coarse model region the grid consists of columns that span the whole channel width in $y$-direction. We define

$$W^k_G(\Omega) = \left\{ w \in V^k_G(\Omega) \mid w|_E(x) = \sum_{0 \leq \alpha \leq k} a_\alpha x^\alpha \text{ for all } E \in G, E \subset \Omega_l \right\}.$$

The weak form of the coupled problem (7) now reads as follows: We are looking for $U^{1d/2d}(\cdot, t) \in W^k_G(\Omega)$ such that for all test functions $\varphi \in W^k_G(\Omega)$ it holds

$$\int_{\Omega_l} \partial_t U^{1d/2d} \varphi \, dx = \sum_{E \in G} \int_E \left( F(U^{1d/2d}) \partial_x \varphi + G(U^{1d/2d}) \partial_y \varphi \right) \, dx$$

$$- \sum_{E \in G} \sum_{e \in I_E} \left( \int_{e \cap \partial \Omega} \nu_{E,y} G(U^{1d/2d}) \varphi \, d\sigma \right)$$

$$- \sum_{E \in G} \sum_{e \in I_E} \left( \int_{e \cap \partial \Omega} \nu_{E,y} G(U^{1d/2d}) \varphi \, d\sigma \right).$$

Due to the choice of $W^k_G(\Omega)$ the above equation restricted to $\Omega_l$ simplifies to

$$\int_{\Omega_l} \partial_t U^{1d/2d} \varphi \, dx = \sum_{E \in G} \int_E F(U^{1d/2d}) \partial_x \varphi \, dx$$

$$- \sum_{E \in G} \sum_{e \in I_E} \left( \int_{e \cap \partial \Omega} \nu_{E,y} G(U^{1d/2d}) \varphi \, d\sigma \right).$$

for all $\varphi \in W^k_G(\Omega)$. Here we used the fact, that for each column element the sum of the boundary fluxes across upper and lower boundaries evaluates to zero. Note that we still have to assume that the third component of $U^{1d/2d}$ vanishes in the coarse model region $\Omega_l$. 

Figure 2. Schematic representation of an anisotropic grid. In the coarse model region $\Omega_l$ grid elements span the whole channel width.
5. **Numerical Results.** The following test case is devoted to [10, 11]. We simulate the motion of a solitary wave 200 meters along a straight channel of $(-100, 100) \times (-10, 10)$m. A cylindrical obstacle with radius $r = 4.5$m is centered at $(x, y) = (35, 0)$m. The initial data is
\[ h(x, y, 0) = \frac{1}{2} \text{sech}^2(\sqrt{\frac{1}{20}} x) + 5, \]
\[ (hu)(x, y, 0) = \frac{7\sqrt{5}}{5} (h(x, y, 0) - 5), \]
\[ (hv)(x, y, 0) = 0. \]

At $y \in \{\pm 10\}$ solid wall boundary conditions are imposed, and at $x \in \{\pm 100\}$ outflow boundary conditions are prescribed. We choose a model decomposition as sketched in Figure 2: for $x < 0$ the coarse model (6) shall be solved, the fine model will be computed in $x > 0$.

For the solution of (9) a blockwise structured grid with 31,324 elements was used with 400 column-type elements in the coarse model region. Locally, we used tensor product polynomials of second order in the fine model region and of second-zeroth order in the coarse model region to represent the discrete solution. Integration in time was accomplished using an explicit fourth-order Runge-Kutta scheme. A generic limiting approach as described in [2] was applied for stabilization. For comparison we computed the solution of (8) on a structured mesh, which corresponds to using the fine model everywhere. All computations were performed using the discretization modules DUNE and DUNE-FEM [1, 3].

The sequence in Figure 3 shows an outline of the position of the free surface $h$ at different time steps in a neighborhood around the obstacle. For $t < 3.5$s the wave profile is essentially one-dimensional everywhere. As the wave front hits the obstacle reflections move in opposite direction back into the coarse model region. As long as the flux across the model interface is constant the solutions of (8) and
Figure 4. Contour lines of the surface elevation $h$ shown for $x \in [-15, 50]$ at final time $t = 9s$ for the dimensionally heterogeneous 1d-2d simulation (top) and the 2d reference simulation (bottom).

Table 1. Comparison of third order 2d reference computation and coupled 1d-2d computation.

<table>
<thead>
<tr>
<th></th>
<th>Fully 2D</th>
<th>Coupled 1d-2d</th>
<th>Reduction [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid elements</td>
<td>62 924</td>
<td>31 324</td>
<td>49.78</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>1 698 948</td>
<td>838 548</td>
<td>49.36</td>
</tr>
</tbody>
</table>

(9) coincide. Compared to the fully 2d case, only about half the the number of grid elements and degrees of freedom are needed for the 1d-2d simulation, see Table 1.

From $t \approx 8.2s$ on the coarse model assumption is not valid any more. Figure 4 shows the numerical solution for both computations at final time $t = 9s$. Of course, in this test case, we cannot guarantee that in any part of the domain the flow is of coarse type throughout the whole computation. This illustrates the need for adaptive procedures to dynamically identify admissible coarse and fine model regions.

6. Conclusions. In this paper we presented an ansatz for coupling dimensionally heterogeneous shallow water models. We showed that the coupled problem can be reformulated as a DG scheme using spatially heterogeneous ansatz functions. The scheme proved to be reliable as long as the coarse model assumptions made in the derivation are fulfilled.

As demonstrated in Section 5 the strict assumptions made in the a-priori model decomposition cannot be justified in general. As the behavior of the flow is unknown and may change dynamically in time, this approach is limited to simple problem settings. Instead, adaptive methods that identify dynamically those regions, where simplified models can be used without loss of accuracy, are desirable. However, dynamic model adaptation represents an open problem. For a first attempt to
adaptively coupling 1d- and 2d-shallow water equations by means of Finite Volume schemes see [9]. The higher order ansatz presented in this paper will be extended to adaptive simulations in a forthcoming publication.

REFERENCES


E-mail address: christoph.gersbacher@mathematik.uni-freiburg.de
ON CAVITATION IN ELASTODYNAMICS

JAN GIESSELMANN

Institute of Applied Analysis and Numerical Simulation
University Stuttgart
Pfaffenwaldring 57, 70569 Stuttgart, Germany

ATHANASIOS E. TZAVARAS

Department of Applied Mathematics, University of Crete and
Institute for Applied and Computational Mathematics, FORTH
Heracleion, Crete, Greece

Abstract. Based on the seminal works by Ball (1982) and Pericak-Spector and Spector (1988) we investigate compressible, nonlinear elastodynamics as a model describing the onset of fracture or cavitation in a softening elastic material under tensile load. We explore a definition of singular solutions describing fracture via approximating sequences of smooth functions. Moreover, we use these approximating sequences to investigate the energy of such solutions, taking into account the energy needed to open a crack or hole. In particular, we find that the existence of singular solutions and the finiteness of their energy is strongly related to the behavior of the stress response function for infinite stretching. We will detail our findings in one space dimension.

1. Introduction. Starting with the work of Ball [1] compressible, nonlinear elasticity has been used as a model for fracture and cavitation in elastic materials such as rubber. For a study on the relation between the macroscopic behavior of a solid and its energy functional, see [5]. Ball constructed solutions to variational problems, which display discontinuities in the displacement field at the origin. These solutions, which are radially symmetric can be seen as describing cavitation. Ball computed the energy of these solutions comparing it to the energy of trivial solutions with uniform deformation. The upshot of his study is that for sufficiently large prescribed deformations on the boundary the solution with cavity is energetically favorable. Based on his findings Spector and coworkers, e.g. [7, 8, 6], studied the dynamic case getting similar results. One important feature of the singular solutions constructed in these works is that the normal Cauchy stress vanishes on the boundary of the cavity such that all the integrals needed in defining weak solutions and energies exist. However, these constructions do not account for any energy needed for crack and cavity creation. Therefore, we propose a new methodology which should take into account the layer structure of the solution at the onset of fracture.

2000 Mathematics Subject Classification. Primary: 74B20; Secondary: 74R99, 35L70, 35Q74.
Key words and phrases. Nonlinear Elastodynamics, Cavitation.

This research was supported by the EU FP7-REGPOT project “Archimedes Center for Modeling, Analysis and Computation”. AET is partially supported by the “Aristeia” program of the Greek Secretariat for Research.
We consider the equations of compressible, nonlinear elastodynamics, searching for a displacement field

\[ y_{tt} = \text{div} \tau(\nabla y), \text{ in } \mathbb{R}^d \times (0, T), \]  

for some finite time \( T > 0 \), where \( \tau : \mathbb{R}^{d \times d}_+ \to \mathbb{R}^{d \times d}_+ \) is the stress response function, which is related to an energy density function \( W : \mathbb{R}^{d \times d}_+ \to \mathbb{R} \) via \( \tau(F) = \frac{\partial W}{\partial F}(F) \).

We are interested in radially symmetric deformations

\[ y(x, t) = w(|x|, t) \frac{x}{|x|} \]  

with tensile deformations in the far field, i.e.,

\[ y(x, t) = \lambda x \text{ for } |x| > rt \]  

for \( r, \lambda > 0 \). In practice, we are interested in the case \( \lambda \gg 1 \). We focus on solutions of (1) having the form (2) and satisfying (3), which have a discontinuity at the origin. Let us note that using the velocity \( v := y_t \) and the strain \( u := \nabla y \) we can rewrite (1) as the following system of conservation laws:

\[ u_t = \nabla v, \quad v_t = \text{div}(\tau(u)). \]  

The outline of this contribution is as follows: In Section 2 we describe solutions in the one dimensional case. We determine their energy dissipation in Section 3. We would like to mention that basically the same results hold in several space dimensions, but the technical and notational complexity is considerably higher. Therefore, we refer to [4] for the details.

2. A special 1-dimensional solution including fracture.

2.1. A special ansatz function. We consider the one dimensional version of (1), i.e.,

\[ y_{tt} = \tau(y_x)x, \quad x \in \mathbb{R}, \ t > 0 \]  

describing the longitudinal or shearing motion of an one dimensional elastic bar. We impose a uniform deformation far away, i.e.,

\[ y(x, t) = \lambda x, \text{ for } |x| > rt, \]  

for some \( r \) sufficiently large. For large values of \( \lambda \) the bar under consideration will break and after breaking the continuum hypothesis which is crucial in deriving (5) fails. Therefore, (5) will not be an appropriate description of the physical situation. However, we expect that there is an intermediate range of values of \( \lambda \) from a range of loading where the model is valid to a range where it loses validity. In some situations it may arguably be possible to give a meaning to (5) in and past this intermediate regime. In particular, we think that (5) is a reasonable description at the onset of fracture which is the situation we want to explore.

We assume that the stress response function \( \tau \) and the energy density \( W \) satisfy

\[ \tau'(u) > 0, \ \tau''(u) < 0 \ \forall u > 0 \]  

\[ \lim_{u \searrow 0} \tau(u) = -\infty, \ \lim_{u \searrow 0} W(u) = \infty. \]  

(H1)  

(H2)

Given the first condition of (H1) the one dimensional version of (4) has the wave speeds \( \lambda_{1,2}(u) = \pm \sqrt{\tau'(u)} \) and is therefore hyperbolic. The second part of (H1) is related to softening response of the material and is a crucial requirement for our
Figure 1. Sketch of $W$ and $\tau$ satisfying (H1) and (H2).

analysis. The second hypothesis (H2) is placed to prevent a finite volume from being compressed down to zero. However, as we focus on tensile deformations this condition will not have any significant impact.

Smooth solutions of (5) satisfy the conservation law for energy

$$
\left( \frac{1}{2} |y_t|^2 + W(y_x) \right)_t - (\tau(y_x)y_t)_x = 0.
$$

This means that any change in mechanical (stored and kinetic) energy in a certain interval is due to the work performed on the boundary of said interval, as can be seen from the integral form

$$
\frac{d}{dt} \int_a^b \frac{1}{2} |y_t|^2 + W(y_x) \, dx = \tau(y_x)y_t|_a^b.
$$

In particular, (8) shows that if the deformation is uniform outside a certain interval $(-r, r)$, i.e., $y(x) = \lambda x$ for $|x| > r$, the energy inside $(-r, r)$ is conserved.

We focus on solutions in which stress and strain only depend on the self-similar variable $\xi = \frac{x}{t}$ which is achieved by the ansatz $y(x, t) = tY\left(\frac{x}{t}\right)$. One easily verifies that $y$ satisfies (5) if $Y$ satisfies

$$
\xi^2 Y''(\xi) = \tau(Y'(\xi))'.
$$

Introducing the velocity $V(\xi) := \xi Y'(\xi) - Y(\xi)$ and the strain $U(\xi) := Y'(\xi)$ in the self-similar variable we can rewrite (9) as

$$
-\xi U'(\xi) = V'(\xi)
$$

$$
-\xi V'(\xi) = \tau(U(\xi))'.
$$

We will test a class of self-similar solutions

$$
Y(\xi) = \begin{cases} 
\lambda \xi & \xi > -\sigma \\
-Y(0) + \alpha \xi & -\sigma < \xi < 0 \\
Y'(0) + \alpha \xi & 0 < \xi < \sigma \\
\lambda \xi & \sigma < \xi
\end{cases}
$$

where $\alpha$, $\lambda$, $\sigma$ and $Y(0)$ are positive parameters that satisfy $\lambda > \alpha$, and

$$
\lambda \sigma = Y(0) + \alpha \sigma,
$$

$$
\sigma = \sqrt{\frac{\tau(\lambda) - \tau(\alpha)}{\lambda - \alpha}}.
$$
For this deformation the velocity and strain distributions are given by

\[
U(\xi) = 2Y(0)\delta_{\xi=0} + \alpha\chi_{|\xi|<\sigma} + \lambda\chi_{|\xi|>\sigma},
\]

\[
V(\xi) = Y(0)\chi_{0<\xi<\sigma} - Y(0)\chi_{0<-\xi<\sigma}.
\] (13)

Due to the discontinuity of \(Y\) (corresponding to a delta distribution in \(U\)) it is not straightforward how such a function can be interpreted as a solution of (5) or (10). Before we give a definition of solution based on regularizations of the prospective solutions some remarks are in order.

**Remark 1.**
1. The solution ansatz (11) is similar to and in fact inspired by the dynamic cavitating solutions in three dimensions introduced in [7]. The main difference is that the solutions in [7] do not contain a delta measure in the strain.
2. For a given \(\lambda > 0\) there is a one parameter family of functions (11) satisfying (12).
3. The ansatz (11) has singularities at \(\xi = \pm\sigma\) and at \(\xi = 0\). The former are shocks. Due to the symmetry the Rankine Hugoniot conditions at both singularities amount to

\[
-\sigma(\lambda - \alpha) = -Y(0)
\]

\[
-\sigma(-Y(0)) = \tau(\lambda) - \tau(\alpha)
\]

and are equivalent to (12). The shock at \(\xi = \sigma\) belongs to the second characteristic family and the Lax shock admissibility criterion reads

\[
\lambda_2(U-) = \sqrt{\tau'(\alpha)} > \sigma = \sqrt{\frac{\tau(\lambda) - \tau(\alpha)}{\lambda - \alpha}} > \sqrt{\tau'(\lambda)} = \lambda_2(U+).
\]

Thus, it is satisfied provided \(\lambda > \alpha\). The same holds for the singularity at \(\xi = -\sigma\).
4. At positive times the solution \(Y\) has a discontinuity of size \(2tY(0)\) at the origin, which is associated to a crack whose boundary is moving according to

\[
y(0\pm, t) = \pm tY(0)
\]

such that \(Y(0)\) is the speed of the crack. In contrast (12) ensures the continuity of \(Y\) outside the origin.
5. For \(t \to 0\) we have \(y(x, t) \to \lambda x\) and \(v(x, t) \to 0\). Thus, the initial data is a configuration with homogeneous deformation which is at rest. The problem
with these initial data obviously permits the trivial solution \( y(x,t) = \lambda x \).

6. In view of (12) a straightforward calculation using \(-\xi \partial_\xi \delta_{\xi=0} = \delta_{\xi=0}\) in \( D' \) shows

\[-\xi u' = v' \]

in \( D' \). Hence, (11) satisfies (10)\(_1\). Thus, we will focus on giving a meaning to \( \tau(u) \) and (10)\(_2\) at the origin. This problem is strongly related to the concept of delta shocks, see section 7.5 of [3] and references therein. A main difference of our approach compared to those usually employed for delta shocks is that we exploit the structure from (5) to give a meaning to solutions. The solution for the p-system given in [3] can easily be handled in our framework, see [4].

2.2. Slic-solutions. Our next step is to give a meaning to the equation around \( x = 0 \) by introducing the notion of *singular limiting induced from continuum solution* (or short slic-solution). Roughly speaking a discontinuous function is called a slic solution provided it can be obtained as the limit of a sequence of smooth functions approximately solving the problem.

**Definition 2.1.** Let \( y \in L^1_{\text{loc}}(\mathbb{R} \times [0, \infty), \mathbb{R}) \) and let its extension to \( t < 0 \) via \( y(x,t) = \lambda x \) be denoted by \( y, \) as well. Let \( \phi \) be a mollifier: \( \phi \in C^\infty_c(\mathbb{R}), \phi \geq 0, \text{ supp } \phi = B_1 \) (the ball of radius 1), \( \int \phi = 1. \) Let \( \phi_n = n\phi(nx) \) and define

\[ y^n(x,t) = \phi_n * y = \int \phi_n(x-z)y(z,t) \, dz. \]

Then \( y \) is called a slic-solution of (5) provided

\[ \int_\mathbb{R} \int \int y^n \psi_{tt} + \tau(y^n_x) \psi_x \, dt \, dx \to 0 \]

for any \( \psi \in C^\infty(\mathbb{R} \times \mathbb{R}, \mathbb{R}) \).

**Remark 2.**
1. The definition of slic-solution is in fact independent of the particular choice of mollifier \( \phi. \)
2. Because of the special structure (11) of our solution \( y, \) mollification in \( x \) also makes it smooth in \( y. \) This is not true for general functions in the class (2). Thus, in a more general setting we would also need mollification in \( t \) in the definition of slic-solutions. We omit this for brevity.
3. A straightforward calculation shows that any standard \( H^1 \) weak solution of (5), which satisfies the correct initial and boundary data, is also a slic solution.

**Proposition 1.** If (12) is satisfied and

\[ L := \lim_{u \to \infty} \frac{\tau(u)}{u} = 0 \quad \text{(H3)} \]

then \( y \) defined in (11) is a slic-solution of (5).

**Proof.** We only give the main strategy of the proof, for details see [4]. Note that \( y \) is a weak solution of (5) away from the set \( \{(t,0) : t > 0\}. \) Moreover, the sequences \( (y^n)_{tt} \) and \( \tau(y^n_x)_x \) restricted to \( \{(x,t) : |x| > \frac{1}{n} \vee t < 0\} \) are uniformly bounded and converge pointwise. Thus, it is sufficient to show

\[ \lim_{n \to \infty} \int_0^\infty \int_{-\frac{1}{n}}^{\frac{1}{n}} \tau(y^n_x(x,t)) \psi_x(x,t) \, dx \, dt = 0 \]
for all $\psi \in C^1_c(\mathbb{R} \times \mathbb{R}, \mathbb{R})$. We can compute
\[
\lim_{n \to \infty} \int_{-\frac{1}{n}}^{\frac{1}{n}} \tau(y^n_x(x,t))\psi_x(x,t) \, dx = 2LtY(0)\psi_x(0,t),
\]
which proves the lemma.

3. The energy needed to open a crack. It seems reasonable to consider the energies of the approximate solutions $y^n$ as an approximation of the energy of the slic-solution $y$. As the approximate solutions are smooth they satisfy the energy identity
\[
\left(\frac{1}{2}(v_n)^2 + W(u_n)\right)_t - (v_n\tau(u_n))_x = ((v_n)_t - \tau(u_n)_x) v_n, \tag{14}
\]
where $v_n := y^n_t$ and $u_n = y^n_x$. Let us define the residual of $y^n$ in the wave equation (5) as $f_n$, i.e.,
\[
f_n := (v_n)_t - \tau(u_n)_x. \tag{15}
\]
This quantity could be seen as an exterior force which (if it were applied) would make $y^n$ the exact solution. Let $B = (-r,r)$ some interval containing a neighborhood of the whole wave fan at time $t$. Then (for sufficiently large $n$) $v_n|_{\partial B} = 0$. Thus, the energy of the solution inside $B$
\[
E_B[y^n] := \int_B \frac{1}{2}(v_n)^2 + W(u_n) \, dx \tag{16}
\]
evolves according to
\[
\frac{d}{dt}E_B[y^n] = \int_B ((v_n)_t - \tau(u_n)_x) v_n \, dx. \tag{17}
\]

**Proposition 2.** Let $v_n|_{\partial B} = 0$ and $n > \frac{2}{\sigma t}$ then the energy change rate is given by
\[
\frac{d}{dt}E_B[y^n] = Y(0)^2 \sigma + 2\sigma(W(\alpha) - W(\lambda)) + 2\int_0^\frac{1}{n} \tau(\alpha + 2\phi_n(x)tY(0))2\phi_n(x)Y(0) \, dx. \tag{18}
\]

**Proof.** We have
\[
\frac{d}{dt}E_B[y^n] = \int_{-r}^r v_n(v_n)_t + \tau(u_n)(u_n)_t \, dx = 2\int_0^r v_n(v_n)_t + \tau(u_n)(u_n)_t \, dx.
\]
We calculate (for $n$ large enough) and $x > 0$
\[
v_n(x,t) = -Y(0) \int_{-\sigma t}^0 \phi_n(x-z) \, dz + Y(0) \int_0^{\sigma t} \phi_n(x-z) \, dz
\]
\[
u_n(x,t) = \begin{cases} 
\alpha + 2Y(0)\phi_n(x) & : x < \frac{1}{n} \\
\alpha & : \frac{1}{n} < x < \sigma t - \frac{1}{n} \\
\alpha \int_0^{\sigma t} \phi_n(x-z) \, dz + \lambda \int_{\sigma t}^\infty \phi_n(x-z) \, dz & : \sigma t - \frac{1}{n} < x < \sigma t + \frac{1}{n} \\
\lambda & : x > \sigma t + \frac{1}{n}
\end{cases}
\]
\[
\begin{align*}
(v_n)_t(x,t) &= -Y(0)\sigma \phi_n(x + \sigma t) + Y(0)\sigma \phi_n(x - \sigma t) \\
(u_n)_t(x,t) &= 2Y(0)\phi_n(x) + (\alpha - \lambda)\sigma \phi_n(x - \sigma t)
\end{align*}
\]
and thus, as $n > \frac{2}{\sigma t}$,

$$\frac{d}{dt} E_B[y^n] = 2 \int_{\sigma t - \frac{1}{n}}^{\sigma t + \frac{1}{n}} Y(0)^2 \left( \int_0^{\sigma t} \phi_n(x - z) d z \right) \sigma \phi_n(x - \sigma t) d x$$

$$+ 2 \int_0^{\frac{1}{n}} \tau(\alpha + 2tY(0)\phi_n(x))2Y(0)\phi_n(x) d x$$

$$+ 2 \int_{\sigma t - \frac{1}{n}}^{\sigma t + \frac{1}{n}} \tau \left( \alpha + (\lambda - \alpha) \int_{\sigma t}^{\infty} \phi_n(x - z) d z \right) (-\lambda - \alpha) \sigma \phi_n(x - \sigma t) d x$$

$$=: I_1 + I_2 + I_3. \number{21}$$

We find

$$I_1 = -\int_{\sigma t - \frac{1}{n}}^{\sigma t + \frac{1}{n}} Y(0)^2 \sigma \frac{d}{dx} \left( \int_0^{\sigma t} \phi_n(x - z) d z \right)^2 d x = \frac{1}{2} Y(0)^2 \sigma. \number{22}$$

Moreover,

$$I_2 = -\sigma \int_{\sigma t - \frac{1}{n}}^{\sigma t + \frac{1}{n}} \frac{d}{dx} W(\alpha + (\lambda - \alpha) \int_{\sigma t}^{\infty} \phi_n(x - z) d z) d x = -\sigma (W(\lambda) - W(\alpha)). \number{23}$$

Inserting (22) and (23) into (21) we obtain the assertion of the Lemma.

The solution contains three waves, all of which contribute to the energy rate. There are the two shocks located at $\xi = \pm \sigma$ and the discontinuity of $y$ at 0. Both shocks satisfy the Lax criterion and therefore there is energy dissipation along them. Moreover, the energy dissipation at the shocks $\mu_\sigma, \mu_{-\sigma}$ can be calculated by classical Riemann problem theory, see [3, Sec. 8.5]. This gives

$$\mu_\sigma = -\sigma \left( -\frac{1}{2} Y(0)^2 + W(\lambda) - W(\alpha) \right) + Y(0)\tau(\alpha)$$

$$= Y(0) \left( -\frac{W(\lambda) - W(\alpha)}{\lambda - \alpha} + \frac{1}{2} (\tau(\lambda) + \tau(\alpha)) \right) < 0. \number{24}$$

because $\tau'' < 0$, and

$$\mu_{-\sigma} = \sigma \left( \frac{1}{2} Y(0)^2 + W(\alpha) - W(\lambda) \right) + Y(0)\tau(\alpha) < 0. \number{25}$$

As there is no energy dissipation/creation away from the three waves the energy balance can be expressed as

$$\frac{d}{dt} E_B[y^n] = \mu_{-\sigma} + p_c^n + \mu_\sigma \number{26}$$

where

$$p_c^n = 2 \int_0^{\frac{1}{n}} \tau(\alpha + 2\phi_n(x)Y(0))2\phi_n(x)Y(0) d x - Y(0)\tau(\alpha) \number{27}$$
can be interpreted as the work performed by the force \( f_n \) at the crack. It describes the work needed to open the crack. Its limiting contribution is given by
\[
\lim_{n \to \infty} p_n^c = \lim_{n \to \infty} 2 \left( \int_0^1 \tau(\alpha + 2n\phi(z)t\tau(Y(0)))2\phi(z)Y(0) \, dz - Y(0)\tau(\alpha) \right)
\]
\[
= \lim_{n \to \infty} 2 \left( \int_0^1 \tau(\alpha + 2n\phi(z)t\tau(Y(0)))2\phi(z)Y(0) \, dz - Y(0)\tau(\alpha) \right) = \left\{ \begin{array}{ll}
\infty & \text{if } \lim_{u \to \infty} \tau(u) = \infty \\
2(\tau_\infty - \tau(\alpha))Y(0) & \text{if } \lim_{u \to \infty} \tau(u) =: \tau_\infty < \infty.
\end{array} \right.
\]
(28)

Thus, the energy rate needed to open a crack is infinite in case \( \lim_{u \to \infty} \tau(u) = \infty \) while it is finite in case \( \lim_{u \to \infty} \tau(u) < \infty \). In that case the energy needed to create a crack can be computed as follows:

**Proposition 3.** Let \( \tau_\infty := \lim_{u \to \infty} \tau(u) < \infty \) then the total energy of the wave fan satisfies
\[
\frac{d}{dt} E_B[y^n] = \mu - \sigma + p_n^c + \mu_\sigma \to \infty \mu - \sigma + 2(\tau_\infty - \tau(\alpha))Y(0) + \mu_\sigma =: T > 0,
\]
where \( \mu_\sigma \) denotes the energy dissipation at the shocks and \( 2(\tau_\infty - \tau(\alpha))Y(0) \) is the cost for opening the crack. The total energy rate \( T \) is positive, i.e., energy is needed to open the crack.

**Proof.** Given Proposition 2 and (28) it only remains to show \( T > 0 \). Inserting (24) and (25) into the definition of \( T \) we get
\[
T = \sigma Y(0)^2 + 2Y(0) \left( \tau_\infty - \frac{W(\lambda) - W(\alpha)}{\lambda - \alpha} \right) > 2Y(0)(\tau_\infty - \tau(u^*)) > 0
\]
for some \( u^* \in (\alpha, \lambda) \).

**REFERENCES**


E-mail address: jan.giesselmann@mathematik.uni-stuttgart.de
E-mail address: tzavaras@tem.uoc.gr
OPERATOR-SPLITTING-BASED ASYMPTOTIC PRESERVING SCHEME FOR THE GAS DYNAMICS EQUATIONS WITH STIFF SOURCE TERMS

CHRISTOPHE CHALONS
Laboratoire de Mathématiques de Versailles, U.M.R. 8100
Université de Versailles Saint-Quentin-en-Yvelines, UFR des Sciences
bâtiment Fermat, 45 avenue des Etats-Unis
78035 Versailles Cedex, France

MATHIEU GIRARDIN AND SAMUEL KOKH
DEN/DANS/DM2S/STMF/LMEC – CEA Saclay, bât. 454 PC 47
91191 Gif sur Yvette Cedex, France

Abstract. We propose a numerical scheme for the gas dynamics equations with external forces and friction terms that is able to accurately approximate the flow in the large friction regime with a coarse spatial discretization. The key idea is to use a classic convection/source operator splitting and then to reduce the numerical diffusion involved with the approximation of the convective terms when the friction effects are dominant. The overall resulting scheme satisfies a discrete entropy inequality under a condition on the numerical diffusion reduction. Numerical tests are proposed in 1D and 2D that show a gain of accuracy.

1. Introduction. We are interested in the simulation of subsonic compressible flows where the driving phenomena are stiff source terms and material transport. Such a flow configuration may be encountered in several industrial processes like the flows within the core of a nuclear reactor. We consider the gas dynamics equations with external forces and friction and propose a formal asymptotic analysis for large friction. In this context, it is well known that classic operator splitting techniques fail to produce accurate approximate solutions when the small scales are not resolved, that is to say when coarse meshes are used. For a solution whose long-time behaviour is governed by the solution of a parabolic system, asymptotic preserving numerical schemes may be used to transpose this feature to the discrete setting without using very fine meshes (see e.g. [1, 3, 4, 7, 10, 11, 12] and the references therein). The present work is dedicated to the short time behaviour of the solution and relies on a careful asymptotic analysis of the classic operator splitting technique. In particular, we focus on the numerical diffusion created by the upwind part of the numerical flux function associated with the treatment of the convective part. This numerical diffusion is indeed what prevents the scheme from computing accurate solutions in the asymptotic limit of large friction coefficients. Based on this analysis, we propose a numerical diffusion reduction technique: the resulting scheme is an operator splitting algorithm with a good asymptotic behaviour. It

2000 Mathematics Subject Classification. Primary: 65M08, 76M12; Secondary: 65M12.
Key words and phrases. Euler system, asymptotic preserving, Lagrange-Projection, relaxation.
satisfies a discrete entropy inequality under a condition on the correction. The proposed diffusion reduction technique is similar to the one used for the low Mach limit in [8].

2. Governing equations and large friction asymptotic behaviour.

Governing equations. The gas dynamics equations with gravity and friction terms in Eulerian coordinates are given by

\[
\begin{align*}
\partial_t \rho + \partial_x (\rho u) &= 0, \\
\partial_t (\rho u) + \partial_x (\rho u^2 + p) &= \rho (g - \alpha u), \\
\partial_t (\rho E) + \partial_x ((\rho E + p) u) &= \rho u (g - \alpha u),
\end{align*}
\]

(1)

where \( \rho, u \) and \( E \) denote the density, the velocity, and the total energy of the fluid, \( g \) the gravitational acceleration and \( \alpha \) a friction parameter. The pressure law \( p = p(\rho, e) \) is assumed to be a given function of the density \( \rho \) and the internal energy \( e = E - \frac{u^2}{2} \) that satisfies the usual Weyl assumptions [15]. The sound speed \( c \) is given by \( c = \left[ \left( \frac{\partial p}{\partial \rho} \right)_e + \left( \frac{\partial p}{\partial e} \right)_\rho \right]^{1/2} \).

Large friction asymptotic behaviour. We are interested in the behaviour of (1) when the friction parameter \( \alpha \) goes to infinity. We model this flow regime by simply replacing \( \alpha \) with \( \frac{2}{\epsilon} \) in (1), where \( \epsilon > 0 \) denotes a small parameter. The asymptotic regime is obtained when \( \epsilon \to 0 \).

Let us then assume that the velocity \( u \) admits an asymptotic expansion in powers of \( \epsilon \) of the form \( u = u^0 + \epsilon u^1 + \mathcal{O}(\epsilon^2) \). Multiplying the second equation of (1) by \( \epsilon \) and letting \( \epsilon \) go to 0 gives \( u^0 = 0 \). Then system (1) reads

\[
\begin{align*}
\partial_t \rho + \epsilon \partial_x (\rho u^1) &= \mathcal{O}(\epsilon^2), \\
\partial_x p &= \rho (g - \alpha u^1) + \mathcal{O}(\epsilon), \\
\partial_t (\rho E) + \partial_x ((\rho E + p) u^1) &= \epsilon pu^1 (g - \alpha u^1) + \mathcal{O}(\epsilon^2).
\end{align*}
\]

(2)

Note that compared to (1), the fluxes now involve the first order corrector \( u^1 \) (or equivalently the small scale contribution \( \epsilon u^1 \)) instead of \( u \). From a numerical point of view, our objective is to propose numerical fluxes that are able to capture those small scales in the limit \( \epsilon \to 0 \).

Lagrange-Source-Projection operator splitting. We first propose an operator splitting between the terms accounting for the transport waves, the acoustic waves and the source terms. More precisely and using the chain rule for the space derivatives we split up the system (1) into the following three subsystems. The first subsystem describes the transport process and reads

\[
\begin{align*}
\partial_t \rho + \partial_x \rho u &= 0, \\
\partial_t (\rho u) + \partial_x (\rho u^2 + p) &= 0, \\
\partial_t (\rho E) + \partial_x ((\rho E + p) u) &= 0.
\end{align*}
\]

(3)

The second subsystem governs the acoustic phenomena, namely

\[
\begin{align*}
\partial_t \rho + \rho \partial_x u &= 0, \\
\partial_t (\rho u) + \rho u \partial_x u + \partial_x p &= 0, \\
\partial_t (\rho E) + \rho E \partial_x u + \partial_x (pu) &= 0.
\end{align*}
\]

Or equivalently with \( \tau = \frac{1}{\rho} \) the specific volume and the mass variable \( m \) such that \( \tau \partial_x = \partial_m \)

\[
\begin{align*}
\partial_t \tau - \partial_m u &= 0, \\
\partial_t u + \partial_m p &= 0, \\
\partial_t E + \partial_m (pu) &= 0.
\end{align*}
\]

(4)
3. Naive operator splitting numerical scheme. We begin with a natural discretization of (1) based on the previous Lagrange/Source/Projection decomposition. Let $\Delta x$ and $\Delta t$ represent the constant time and space steps. We set $x_{j+\frac{1}{2}} = x_{j-\frac{1}{2}} + \Delta x$ and $t^n = t^{n-1} + \Delta t$. In the sequel, $X^n_j$ denotes the approximate value of $X$ at time $t^n$ within the cell $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$. We propose a three-step numerical strategy based on the splitting of operator (4),(5),(3).

Regarding the Lagrange step (4) and using clear notations, we propose the following update formulas

$$
\begin{align*}
&\left\{ \begin{array}{l}
\gamma^n_j = \gamma^n_j + \frac{\Delta t}{\rho_j^n \Delta x} \left( u_j^{n+\frac{1}{2}} - u_j^{n-\frac{1}{2}} \right) ; \\
\rho_j^n L = \rho_j^n - \frac{\Delta t}{\rho_j^n \Delta x} \left( p_j^{n+\frac{1}{2}} - p_j^{n-\frac{1}{2}} \right) ; \\
E_j^n L = E_j^n - \frac{\Delta t}{\rho_j^n \Delta x} \left( (pu)_j^{n+\frac{1}{2}} - (pu)_j^{n-\frac{1}{2}} \right),
\end{array} \right.
\end{align*}
$$

where the numerical fluxes are defined by

$$
\begin{align*}
&u_j^{n+\frac{1}{2}} = \frac{1}{2} \left( u_{j+1}^n + u_j^n \right) - \frac{1}{2\Delta t} \left( \Pi_{j+1}^n - \Pi_j^n \right), \\
p_j^{n+\frac{1}{2}} = \frac{1}{2} \left( \Pi_{j+1}^n + \Pi_j^n \right) - \frac{\alpha}{2} \left( u_{j+1}^n - u_j^n \right), \\
\Pi_j^n = p(\rho_j^n, e_j^n),
\end{align*}
$$

and associated with a classic pressure relaxation process (see [2, 5, 6, 13, 14]). In order to ensure stability, the parameter $\alpha$ must be chosen sufficiently large according to the so-called subcharacteristic condition $\alpha > \rho c$. This scheme is shown to be nonlinearly stable under the CFL condition

$$
2\alpha \Delta t \leq \rho_j^n \Delta x.
$$

For the system (5), we propose a point-wise implicit evaluation

$$
\begin{align*}
&\tau_j = \tau_j^n L, \\
&\Pi_j = u_j^n L + g \Delta t - \alpha \Delta t \tau_j, \\
&E_j = E_j^n L + g \Delta t \Pi_j - \alpha \Delta t (\Pi_j)^2.
\end{align*}
$$

This implicit treatment is particularly important in the large friction regime since an explicit scheme would require $\Delta t = O(\frac{1}{\alpha})$, leading to $\Delta t = 0$ in the limit $\alpha \to +\infty$.

Regarding the transport step (4), we consider a standard upwind and time-explicit numerical scheme given by

$$
X_j^{n+1} = X_j - \frac{\Delta t}{\Delta x} \left( (u_{j+\frac{1}{2}}^+ - u_{j-\frac{1}{2}}^-) [X_j - X_{j-1}] + (u_{j+\frac{1}{2}}^+ - u_{j-\frac{1}{2}}^-) [X_{j+1} - X_j] \right),
$$

where $X \in \{\rho, \rho u, \rho E\}$ and $u^\pm = \frac{u \pm |u|}{2}$ (see for instance [9]). The update formula (10) is shown to be stable under the CFL condition

$$
\Delta t \left( (u_{j-\frac{1}{2}}^+)^+ - (u_{j+\frac{1}{2}}^-)^- \right) < \Delta x.
$$
For the sake of clarity, let us briefly recall the different steps of the overall method that shall be referred to as LSP-IMEX. Assume that \((\rho_j^n, (\rho u)_j^n, (\rho E)_j^n)\) is known, \((\rho_j^{n+1}, (\rho u)^{n+1}_j, (\rho E)^{n+1}_j)\) is computed by the following three steps:

(i) compute \((\tau_j^{\text{Log}}, u_j^{\text{Log}}, E_j^{\text{Log}})\) from \((\rho_j^n, (\rho u)_j^n, (\rho E)_j^n)\) with (6)-(7),

(ii) compute \((\tau_j, \varpi_j, \overline{E}_j)\) from \((\tau_j^{\text{Log}}, u_j^{\text{Log}}, E_j^{\text{Log}})\) with (9),

(iii) compute \((\rho_j^{n+1}, (\rho u)^{n+1}_j, (\rho E)^{n+1}_j)\) from \((\tau_j, \varpi_j, \overline{E}_j)\) with (10).

Asymptotic analysis. Let us study the asymptotic behaviour of this scheme in the large friction limit. For the sake of simplicity, we focus only on the first to obtain a scheme that provides good numerical results for large friction test cases with coarse meshes. The proposed correction can be easily understood as a numerical diffusion reduction technique.

\[ \partial_t \tau + \epsilon \partial_m u = O(\epsilon^2). \] (12)

We now perform a similar asymptotic analysis in the discrete setting by replacing \(\alpha\) with \(\frac{\epsilon}{2}\) and assuming expansions of the form \(u_j^n = u_j^{n,(0)} + \epsilon u_j^{n,(1)} + O(\epsilon^2)\). Injecting these expansions in the first equation of (7) gives

\[ u_{j+\frac{1}{2}} = \frac{1}{2} \left( u_{j+1}^{n,(0)} + u_j^{n,(0)} \right) - \frac{\Delta x}{2a} \left( \frac{\Pi_j^{n+1} - \Pi_j^n}{\Delta x} \right) + \frac{\epsilon}{2} \left( u_{j+1}^{n,(1)} + u_j^{n,(1)} \right) + O(\epsilon^2). \]

Then, if \(u_j^{n,(0)} = 0\) the first equation of (6) reads

\[ \tau_j^{\text{Log}} = \tau_j^n + \frac{\epsilon \Delta t}{\rho_j^n} \frac{\Delta x}{2} \left( \frac{u_{j+1}^{n,(1)} + u_j^{n,(1)}}{2} - \frac{u_j^{n,(1)} + u_{j-1}^{n,(1)}}{2} \right)
- \frac{\Delta t \Delta x}{\rho_j^n} \left( \frac{\Pi_j^{n+1} - 2\Pi_j^n + \Pi_j^{n-1}}{2a(\Delta x)^2} \right) + O(\epsilon^2). \]

The above equation is clearly not consistent with (12) because of the third term which is of order 1 with respect to \(\epsilon\). This term is dominant compared to the effects of order \(\epsilon\) we are interested in. Nonetheless, it is important to note that this problem no longer occurs if \(\Delta x = O(\epsilon)\). In other words, such a scheme requires a fine mesh to get accurate numerical results in the regime \(\epsilon \to 0\).

Moreover, the second equation of (9) leads to

\[ \varpi_j = \frac{u_j^{\text{Log}} + \gamma \Delta t}{1 + \alpha \epsilon^{-1} \Delta t}, \]

which implies that \(\varpi_j^{(0)} = 0\). Then (10) gives \(u_j^{n+1,(0)} = 0\).

It is important to emphasize that the bad asymptotic behaviour of the scheme comes from the upwind part of the numerical fluxes and not from the source terms discretization based on an operator splitting technique. The non-centred part of the flux \(u_{j+\frac{1}{2}}^*\) may be interpreted as numerical diffusion that becomes predominant in the asymptotic regime \(\epsilon \to 0\) and prevents the scheme from capturing the right interface velocity of order \(\epsilon\). As briefly discussed above, a possible way to circumvent this difficulty consists in using a grid size of order \(\epsilon\) to reduce the importance of the non-centred term in \(u_{j+\frac{1}{2}}^*\). Of course, this cannot be envisaged in practice. In the next section, we propose a correction of \(u_{j+\frac{1}{2}}^*\) involving the friction parameter \(\alpha\) to obtain a scheme that provides good numerical results for large friction test cases with coarse meshes. The proposed correction can be easily understood as a numerical diffusion reduction technique.
4. Suitable operator splitting numerical scheme and numerical diffusion reduction technique.

Numerical diffusion reduction technique. We propose a correction of the previous scheme to obtain the expected large friction asymptotic behaviour. We replace the definition of the relaxation fluxes (7) by

\[
\begin{align*}
 u_j^{t+\frac{\theta}{2}} &= \frac{1}{2} \left( u_j^n + u_j^0 + \frac{\theta_j + \frac{1}{2}}{2a} (\Pi_j^n - \Pi_j^0) \right), \\
 p_j^{t+\frac{\theta}{2}} &= \frac{1}{2} \left( \Pi_j^n + \Pi_j^0 - \frac{\theta_j}{2} (u_j^n - u_j^0) \right), \\
 \end{align*}
\]

and still use (6)-(9) and (10). This scheme will be referred to as LSP-IMEX COR. We recover the LSP-IMEX scheme by choosing \( \theta_j + \frac{1}{2} = 1 \). In order to obtain a good asymptotic behaviour, \( \theta_j + \frac{1}{2} = O(\epsilon) \) is expected. Since the parameter \( \theta_j + \frac{1}{2} \) allows to reduce the importance of the non-centred terms, it may be interpreted as a numerical diffusion reduction technique. As those non-centred terms are crucial for the sake of stability, a condition on this correction to obtain a discrete entropy inequality is expected and given in Theorem 4.1.

Asymptotic analysis. We replace \( \alpha \) by \( \frac{\theta}{2} \) and assume the same asymptotic expansions as before that we inject in the first equation of (13) to obtain

\[
 u_j^{t+\frac{\theta}{2}} = \frac{1}{2} \left( u_j^n + u_j^0 + \frac{\theta_j + \frac{1}{2}}{2a} (\Pi_j^n - \Pi_j^0) \right) + O(\epsilon^2)
\]

If \( u_j^n = 0 \) and \( \theta_j + \frac{1}{2} = e^2 \pi_j = O(\epsilon) \) then the first equation of (6) reads

\[
 \tau_j^{\text{Lag}} = \tau_j^n + \frac{\epsilon \Delta t}{\rho_j^n} \frac{1}{\Delta x} \left( \frac{1}{2} \left( u_j^n + u_j^0 + u_j^{n+1} \right) - \frac{1}{2} \left( u_j^n + u_j^{n+1} \right) \right)
 - \frac{\epsilon \Delta t}{\rho_j^n} \left( \theta_j + \frac{1}{2} \frac{(\Pi_j^n - \Pi_j^0)}{\Delta x} - \theta_j - \frac{1}{2} \frac{(\Pi_j - \Pi_j^{n-1})}{\Delta x} \right) + O(\epsilon^2),
\]

that is consistent with (12) as the third term is of order \( O(\epsilon) \) and consistent with 0. A straightforward analysis, similar to the one just performed for the density in Lagrangian coordinates, proves that the overall numerical scheme is consistent in the asymptotic limit with (2).

Main properties. We now give the main properties of the LSP-IMEX COR scheme.

**Theorem 4.1.** Under the CFL condition (8) and (11) the numerical scheme LSP-IMEX COR is well-defined and satisfies the following properties

(i) it is a conservative scheme for the density \( \rho \). It is also a conservative scheme for \( pu \) and \( pE \) when the source terms are omitted.

(ii) \( \rho_j^n > 0 \) for all \( j \) implies that \( \rho_j^n > 0 \) for all \( j \) and \( n > 0 \).

(iii) it preserves the large friction asymptotic behaviour in the sense that the scheme is consistent with (2) in the asymptotic limit \( \epsilon \to 0 \).

(iv) in addition, under the following condition on the correction \( \theta_j + \frac{1}{2} \)

\[
\frac{1}{2} (\alpha_j - u_j^{\text{Lag}})^2 + \frac{\Delta t}{\rho_j^n} \left( \frac{1}{2a^2} (p(\rho_j^{n+\frac{1}{2}} s_j^n) - p_j^{n+\frac{1}{2}})^2 - \frac{\gamma_j + \frac{1}{2}}{a} (p(\rho_j^{n+\frac{1}{2}} s_j^n) - p_j^{n+\frac{1}{2}}) \right) \\
+ \frac{\Delta t}{\rho_j^n} \left( \frac{1}{2a^2} (p(\rho_j^{n+\frac{1}{2}} s_j^n) - p_j^{n+\frac{1}{2}})^2 + \frac{\gamma_j + \frac{1}{2}}{a} (p(\rho_j^{n+\frac{1}{2}} s_j^n) - p_j^{n+\frac{1}{2}}) \right) \geq 0,
\]
where $\gamma_{j+\frac{1}{2}} = \frac{1-\theta_{j+\frac{1}{2}}}{2a} (\Pi_{j+1} - \Pi_j)$, 
$\rho_{j+\frac{1}{2}}^{R,s} = \frac{\rho_j^{s}}{1+\rho_j^{s} a^{-1} \left( u_j^{s} - (u_j^{s} + \gamma_j) \right)}$,
and $\rho_{j+\frac{1}{2}}^{L,s} = \frac{\rho_j^{s}}{1-\rho_j^{s} a^{-1} \left( u_j^{s} - (u_j^{s} + \gamma_j) \right)}$.

Verifying this discrete entropy inequality

$$(ps)_j^{n+1} \leq (ps)_j^{n} - \frac{\Delta t}{\Delta x} \left( G_{j+\frac{1}{2}} - G_{j-\frac{1}{2}} \right)$$

where $s_j^{n} = s(\rho_j^{s}, e_j^{s})$, $s$ denotes the mathematical entropy associated to the Euler system and $G_{j+\frac{1}{2}} = (u_j^{s} + \gamma_j) (ps)_j^{Lag} + (u_j^{s} - \gamma_j) (ps)_{j+1}^{Lag}$.

Note that, this condition in (iv) is of course true for $\theta_{j+\frac{1}{2}} = 1$ so that the LSP-IMEX scheme verifies this discrete entropy inequality.

5. Numerical results. We propose to test both LSP-IMEX and LSP-IMEX COR scheme against a test case that has been proposed in [4] and [7]. The friction coefficient is given by $\alpha = 10^6 \text{s}^{-1}$. Both LSP-IMEX and LSP-IMEX COR computations are performed with a time step defined by $2a\Delta t = \min(\rho_j^{s}) \Delta x$ which complies with (8) and (11).

We run our numerical test using a spatial discretization over 100 cells, 1000 cells and 10 000 cells for LSP-IMEX scheme. In figure 1, we can see that there is a large amount of numerical diffusion for coarse meshes and so the numerical solution is not accurate.

Then we set $\theta_{j+\frac{1}{2}} = \min \left( \frac{2a}{\alpha (\rho_{j+\frac{1}{2}} + \rho_j)}, 1 \right)$. This choice is nondimensional and verifies $\theta_{j+\frac{1}{2}} = \mathcal{O}(\epsilon)$ so that the correct asymptotic behaviour is expected when $\epsilon \to 0$. Moreover, we recover $\theta_{j+\frac{1}{2}} = 1$ when $\alpha \to 0$ or $\Delta x \to 0$. In other words, the correction is not activated when the friction parameter is not large enough or when we use a fine discretization. We run the same test for LSP-IMEX COR scheme. In figure 2, we observe that the approximate solutions are much more accurate than the one computed without the correction, especially for a coarse mesh. Let us note $J = \max (j \in \mathbb{N} | j \Delta x \leq 1)$, by plotting $n \mapsto \sum_j ((ps)_j^{n+1} - (ps)_j^{n}) + ((ps)_j^{n} - (ps)_j^{n})$, we observe in figure 3 that the average entropy inequality is verified.

We derive a multi-dimensional version of the LSP-IMEX COR scheme by using the 1D fluxes of the scheme and the rotational invariance of the equations. We consider the 2D domain $[0,1]^2$, discretized over a 5000-triangle grid. The friction parameter is set to $\alpha = 10^6 \text{s}^{-1}$ and the initial condition reads

$$\left\{ \begin{array}{ll}
(\rho, u, p) = (1.0, 0, 10000.0), & \text{if } (x - 0.5)^2 + (y - 0.5)^2 \geq 0.3, \\
(\rho, u, p) = (2.0, 0, 26390.2), & \text{if } (x - 0.5)^2 + (y - 0.5)^2 < 0.3.
\end{array} \right.$$

In figure 4, we observe that LSP-IMEX scheme leads to a large amount of numerical diffusion while LSP-IMEX COR scheme with $\theta = \frac{1}{\alpha \Delta x}$ gives much more accurate numerical results.

Acknowledgments. The authors are supported by the LRC MANON (CEA/DEN/DANS/DM2S and UPMC-CNRS/LJLL).
Figure 1. Profile at time $t = 0.01\text{s}$ of the velocity (left) and the pressure (right) obtained for a 100-cell, 1000-cell and 10 000-cell grid with the LSP-IMEX scheme and the reference solution (LSP-IMEX COR scheme with 10 000-cell mesh).

Figure 2. Profile at time $t = 0.01\text{s}$ of the velocity (left) and the pressure (right) obtained for a 100-cell and 1000-cell grid with the LSP-IMEX COR scheme and the reference solution (LSP-IMEX COR scheme with 10 000-cell mesh).

Figure 3. Variation of mathematical entropy averaged over the domain for each time step obtained for a 100-cell and 1000-cell grid with LSP-IMEX COR scheme.

REFERENCES


Figure 4. Profile at time $t = 0.01s$ of the velocity (left) and the pressure (right) obtained for a 5000-triangle mesh with the LSP-IMEX scheme (left) and the LSP-IMEX COR scheme (right).


E-mail address: christophe.chalons@uvsq.fr
E-mail address: mathieu.girardin@cea.fr
E-mail address: samuel.kokh@cea.fr
BLOW-UP OR NOT BLOW-UP AT THE HYPERBOLIC BOUNDARY FOR A SYSTEM FROM CHEMISTRY?

Christian Bourdarias and Marguerite Gisclon
Université de Savoie, LAMA, CNRS UMR 5127
73376 Le Bourget du lac, France

Stéphane Junca
Université de Nice, Labo. JAD, CNRS UMR 6621
Parc Valrose, 06108 Nice, France
and
Team Coffee, INRIA Sophia Antipolis Mediterranee
2004 route des Lucioles, BP 93
06902 Sophia Antipolis Cedex, France

Abstract. We consider a system of conservation laws modeling heatless adsorption of a gaseous mixture with two species and instantaneous exchange kinetics. In this model the velocity is not constant because the sorption effect is taken into account. Exchanging the roles of the $x, t$ variables we obtain a strictly hyperbolic system with a zero eigenvalue. We construct a solution with a velocity which blows up at the corresponding characteristic “hyperbolic boundary” $\{t = 0\}$. We introduce a “fractional BV” space which appears to be the critical space to avoid this blow-up.

1. Introduction. This paper deals the blow up for an one dimensional $2 \times 2$ hyperbolic system of conservation laws related to a particular isothermal gas-solid chromatography process, called “Pressure Swing Adsorption” (PSA), with two species. Each of the two species simultaneously exists under two phases: a gaseous and movable one with a common velocity $u(t, x)$ and concentration $c_i(t, x) \geq 0$ and a solid (adsorbed) with concentration $q_i$, $i = 1, 2$. One may consult [3] for a precise description of the process and a survey on various related models. In gas chromatography, velocity variations accompany changes in gas composition: it is known as the sorption effect. Here, this effect is taken into account through a constraint: $c_1 + c_2 = 1$. We assume that mass exchanges between the mobile and the stationary phases are infinitely fast thus the two phases are constantly at composition equilibrium and the PSA system reads:

$$\partial_t (c_1 + q_1^*(c_1, c_2)) + \partial_x (u c_1) = 0,$$  \hspace{1cm} (1)

$$\partial_t (c_2 + q_2^*(c_1, c_2)) + \partial_x (u c_2) = 0,$$  \hspace{1cm} (2)

$$c_1 + c_2 = 1.$$  \hspace{1cm} (3)

Notice that $c_1, c_2$ must satisfy $0 \leq c_1, c_2 \leq 1$.

2000 Mathematics Subject Classification. Primary: 35L65, 35L67, 35Q35.
Key words and phrases. Hyperbolic systems, conservation laws, weak solutions, Temple systems, blow up, chromatography, Riemann problem, Front Tracking Algorithm.
Adding (1) and (2) we get, thanks to (3):
\[ \partial_t (q_1^1 (c_1, c_2) + q_2^2 (c_1, c_2)) + \partial_x u = 0. \]
Thanks to (3), we denote \( c = c_1 \) then \( c_2 = 1 - c \) and the unknowns are the velocity \( u \) and the concentration \( c \). We write the PSA system under the form:

\[
\begin{align*}
\partial_x u + \partial_t h(c) &= 0, \tag{4} \\
\partial_x (u c) + \partial_t I(c) &= 0, \tag{5}
\end{align*}
\]
with

\[
\begin{align*}
h(c) &= q_1^1 (c, 1 - c) + q_2^2 (c, 1 - c) = q_1(c) + q_2(c), \\
I(c) &= c + q_1^1 (c, 1 - c) = c + q_1(c).
\end{align*}
\]

Any equilibrium isotherm related to a given species is always non decreasing with respect to the corresponding concentration and non increasing with respect to the others thus \( q_2^1 \leq 0 \leq q_1^1 \). We assume too that one gas is more active than the other: \( h' < 0 \). Then we have

\[
0 \leq q_1' \leq -q_2^1 \quad (H1).
\]

PSA system (4)-(5) is completed by initial and boundary data:

\[
\begin{align*}
\begin{cases} 
  c(0, x) &= c_0(x) \in [0, 1], \quad x > 0, \\
  c(t, 0) &= c_0(t) \in [0, 1], \quad t > 0, \\
  u(t, 0) &= u_0(t) \geq \alpha, \quad t > 0, \quad \text{for some constant } \alpha > 0.
\end{cases}
\end{align*}
\]  

(6)

2. General properties for the PSA system. In this section, we recall briefly some properties of PSA system obtained in [1, 2, 4]. It is possible to analyse PSA system in terms of hyperbolic system provided we exchange the time and space variables. PSA system is strictly hyperbolic with variable \( x \) as the evolution variable: the two eigenvalues are 0 (linearly degenerate) and \( \lambda = H(c)/u \) where

\[
H(c) = 1 + q_1^1(c) - c h'(c) = 1 + (1 - c) q_1^1(c) - c q_2^1(c) \geq 1. \quad (7)
\]

The system is strictly hyperbolic with \( u > 0 \), but at ( \( t = 0, x = X_\infty \) ) it becomes degenerate hyperbolic because \( \lambda \) tends to 0 when \( u \) tends to +\( \infty \). Then the boundary becomes twice characteristic. Using the function \( f(c) = q_1^1(c) - c h(c) \), the right eigenvector associated to \( \lambda \) is genuinely nonlinear in each domain where \( f'' \neq 0 \).

We also assume that

\[
f'' > 0 \quad \text{everywhere} \quad (H2).
\]

PSA system admits the two Riemann invariants

\[
c \text{ and } w = \ln(u) + g(c) \quad \text{where} \quad g'(c) = -\frac{h'(c)}{H(c)} > 0.
\]

Indeed, for smooth solutions we have:

\[
\begin{align*}
\partial_x c + \lambda \partial_t c &= 0, \\
\partial_x w &= 0.
\end{align*}
\]

The Riemann invariant \( W = u G(c) = e^w \) where \( G(c) = \exp(g(c)) \) plays a key role in the blow up mechanism. If \( W \) is decreasing with respect to \( x \) we have

\[
0 < u G(c) \leq u_6(t) G(c_6(t)) \quad \text{thus}
\]

\[
0 < u \leq \|u_6\|_\infty \sup_{c \in [0, 1]} G(c)/\inf_{c \in [0, 1]} G(c)
\]
and \( u \) is bounded. Then if \( W \) is non decreasing with respect to \( x \) we can expect the velocity \( u \) to increase and blow up. Then an important assumption throughout the paper is:

\[
G'' < 0 \ (H3).
\]

It means that \(-W\) is an admissible degenerate convex entropy for weak entropy solutions (with a zero entropy-flux). In contrary, notice that in some cases, for instance an inert gas and an active gas with a Langmuir isotherm as in [1], we have \( G'' > 0 \) then \( W \) is bounded and there is no blow up for the velocity in \( L^\infty \).

There are two families of smooth entropies for the PSA system \( u \psi(c) \) and \( \phi(uG(c)) \) where \( \phi \) and \( \psi \) are any smooth real functions and the corresponding entropy fluxes satisfy \( Q'(c) = h'(c) \psi(c) + H(c) \psi'(c) \) (cf [2]). The first family is degenerate convex (in variables \((u, m = uc)\)) provided \( \psi'' \geq 0 \). The second family is not always convex. Since \( G'' < 0 \) on \([0, 1]\), \(-uG(c)\) is a degenerate convex entropy, with entropy flux \( Q \equiv 0 \). So \((c, u)\) has to satisfy, in the distribution sense

\[
\frac{\partial}{\partial x} (uG(c)) \geq 0.
\]

3. The Riemann Problem. We solve the Riemann problem (1)-(2) with

\[
\begin{align*}
\{ c(t, 0) &= c^- \in [0, 1], \quad t < 0, \\
u(t, 0) &= u^- > 0, \quad t < 0, \\
\} \quad \{ c(t, 0) &= c^+ \in [0, 1], \quad t > 0, \\
u(t, 0) &= u^+ > 0, \quad t > 0. \quad (8)
\end{align*}
\]

We are classically looking for a selfsimilar solution: \( c(t, x) = C(z), \ u(t, x) = U(z), \ z = \frac{x}{t} \). Since we assume that \( f \) is convex, there is no \( \lambda \)-contact discontinuity and the boundary Riemann problem is always solved by a simple wave (see [2]). A 0-wave appears on the line \( \{ t = 0 \} \). As stated in [4], shock and rarefaction curves are monotonic.

**Proposition 1.** Two distinct states \( U^- \) and \( U^0 \) are connected by a 0-contact discontinuity if and only if \( c^- = c^0 \). The solution of the Riemann problem for \( x > 0 \) is

- \( (c, u) = (c^-, u^-) \) for \( t < 0 \),
- a 0-contact discontinuity for \( t = 0 \),
- a \( \lambda \) wave for \( t > 0 \),

![Figure 1. Solution of the Riemann problem when \( f'' > 0 \)](image)

As in a scalar conservation laws with piecewise constant data, no new shock can appear but shocks can disappear. It is a consequence of wave-interactions studied in [4]. We recall the following results concerning interactions:
• if a rarefaction interacts with a shock then we have a contact discontinuity and a rarefaction or a shock. If the shock and the rarefaction have the same strength (i.e. the same absolute variation of the concentration through the simple wave) we have only a contact discontinuity;
• if two shocks interact we obtain a contact discontinuity and a stronger shock;
• if a shock interacts with a contact discontinuity, we obtain a contact discontinuity and a shock with the same strength;
• if a rarefaction interacts with a contact discontinuity, we obtain a contact discontinuity and a rarefaction with the same strength.

More precisely, if a shock disappears after an interaction with a stronger rarefaction, by convention we follow the characteristic speed associated with the eigenvalue $\lambda = H(c)/u$, and the strength of the shock is 0.

We obtained in [1, 2] an existence result for weak global entropy solutions with large BV data for the concentrations and only $L^\infty$ data for the velocity: as in [8], the zero eigenvalue makes the existence possible of stratified solutions in the sense that $u(t, x) = u_b(t) \psi(t, x)$ where $\psi$ is as regular as the concentration $c$ and more than the “initial” data $u_b$ (see [4]). As we will see below, we cannot expect in general to have weak global entropy solutions with $L^\infty$ data for the concentrations because it is possible, in that case, to build a blow up solution for some particular isotherms related for instance to ammonia or water vapor.

4. Temple class and PSA system. It is well known ([7]) that blow up cannot occur for Temple systems with $L^\infty$ data. For PSA system we have:

Lemma 4.1. PSA system (4)-(5) is a Temple system if and only if $\partial_x W = 0$ for the entropy solution of any Riemann problem.

Proposition 2. If $G'' = 0$ (for example if the two isotherms are linear or if $q_1 = 0$ with $q''_2 = 0$ ) then PSA system (4)-(5) is a Temple system.

Remark 1. For an inert gas (for instance the first one: $q_1 = 0$) and an active gas with strictly convex isotherm ($q''_2 > 0$), PSA system is not in the Temple class. For example, it is the case if the active gas is the ammonia or the water vapor. For other examples, see [2, 4].

In the following, we use this assumption to avoid the Temple class:

$PSA$ system (4)-(5) is not a Temple system $(H_4)$.

5. Blow up for velocity. It is already known that systems of two hyperbolic conservation laws may blow up in the $L^\infty$ norm. In [10], R. Young built a example involving two Burgers equations, linearly coupled at the two boundaries. In our example, we also loose the strict hyperbolicity of the system at the blow up point, but the blow up takes place only at the characteristic boundary $\{t = 0\}$ which becomes twice characteristic and only the velocity blows up (cf [5]). The context of our example is physically motivated and the various assumptions are physically relevant but a blow up along the $x$-axis at $t = 0$ is of course artificial. Nevertheless it illustrates what may occur when $BV$ regularity is not ensured for the velocity at the physical boundary.

Theorem 5.1 ([5]). Assume $(H1)-(H2)-(H3)-(H4)$. For any $T > 0$, $X_\infty > 0$, there exists a set of initial and boundary data (6) and a corresponding weak entropy
solution on \([0, T] \times [0, X_\infty]\) of PSA system (4)-(5) such that
\[
\| u \|_{L^\infty([0,T] \times (0,X_\infty)])} = +\infty.
\]
Actually the solution built to prove this theorem has special features.

- The velocity only blows up at the boundary \(\{t = 0\}\) when \(x \to X_\infty\), with arbitrary small data.
- \(\forall X \in [0,X_\infty[, u, c \in L^\infty(0,T; BV(0,X)) \cap L^\infty(0,X; BV(0,T))\). The concentration \(c\) remains bounded while \(u\) blows up.
- Let \(\Omega\) be a neighborhood of the critical point \((t = 0, x = X_\infty)\) such that \(\Omega \subset [0,T] \times [0,X_\infty]\). Outside \(\Omega\), we can prove that \((u,c)\) has a piecewise smooth structure, so the blow up occurs only at the boundary. Indeed, there is an accumulation of wave-interactions near \((t = 0, x = X_\infty)\).
- To build such a solution, we have necessarily to choose the boundary concentration \(c_0(x) = c(0,x)\) not in \(BV(0,X_\infty)\). Else, there is no blow up, see [1, 2].
- The blow-up can be avoid with less constraints on the initial data. Below, we introduce a critical “fractional BV” space to control the velocity at the boundary.

The main ingredient for the construction of the blow-up solution consists in the resolution of two consecutive boundary Riemann problems which leads to increase the velocity without increasing the concentrations.

![Figure 2. The two boundary Riemann problems](image)

For the first boundary Riemann problem the data are \((c_-, c_+, u_0)\) chosen in such a way that the solution is a shock wave and we set \(u_1 = \mathcal{R}(c_-, c_+, u_0)\). For the second problem the data are \((c_+, c_-, u_1)\), the solution is necessarily a rarefaction wave and we set \(u_2 = \mathcal{R}(c_+, c_-, u_1)\). With the assumptions, notice that for the shock curves we have \(c_- < c_+\) (see [1]) and \(u\) is not monotonic along the process because \(u_0 \leq u_2 \leq u_1\). We introduce now the amplification coefficient \(\mathcal{R}(c_-, c_+)\) defined by \(u_2 = \mathcal{R}u_0\).

**Lemma 5.2.** We have the following properties:

1. \(\mathcal{R}(c_-, c_+) \geq 1\),
2. \(\mathcal{R}(c_-, c_+) \equiv 1\) if and only if the system is in the Temple class,
3. \(\mathcal{R}\) is analytic then \(\mathcal{R}(c_-, c_+) > 1\) for all \(c_- < c_+\).
Notice that the commonly used isotherms are analytic, then \( R \) also.

Let \( N > 1 \) be a fixed integer, \( u_0 > 0 \) and \( 0 = x_0 < x_1 < \cdots < x_{2N-1} < X = x_{2N} \). We impose constant initial data and piecewise constant concentration at the boundary: for \( 0 < t < T \), \( 0 < x < X \), \( k = 0, 1, \ldots, N-1 \):

\[
\begin{aligned}
  c(t,0) &= \bar{c}, \\
  u(t,0) &= u_0, \\
  c(0,x) &= \begin{cases} 
  c_k & \text{if } x_{2k} < x < x_{2k+1}, \\
  \bar{c}_k & \text{if } x_{2k+1} < x < x_{2k+2}.
\end{cases}
\end{aligned}
\]  

(9)

Riemann problems at the boundary are alternatively solved by a shock or a rarefaction and we construct the solution of PSA system with the Front Tracking Algorithm as in [4]:

**Theorem 5.3** ([5]). With data (9) where \( \underline{c}_k \) and \( \bar{c}_k \) are constants such that

\[ 0 < \underline{c} = \underline{c}_k < \bar{c} = \bar{c}_k < 1 \]  

there exists an unique weak entropy solution in the class of piecewise Lipschitz functions. Furthermore, this solution has exactly \( N \) shock curves on \([0,T] \times [0,X] \).

There exists a domain \( Z \) (see Fig. 3), a neighborhood of the boundary \( \{t = 0\} \) where boundary Riemann problems do not interact. In this domain \( Z \) we have an explicit solution. Let us denote by \( u_k \) the value of \( u(x,0^+) \) when \( x_{k-1} < x < x_k \) for a given \( k > 0 \). Since \( f'' > 0 \) and \( h' < 0 \) we have \( N \) shocks emerging from \( ((x_{2k}, t = 0))_{k=0}^{N-1} \) and \( N \) rarefactions from \( ((x_{2k+1}, t = 0))_{k=0}^{N-1} \). Furthermore \( u_{2k} < u_{2k+1}, u_{2k} < u_{2k+2} = R u_{2k} \) and \( u_{2k} = R^k u_0 \). Take \( \delta > 0 \) to discretize the rarefactions as in [4] and let shocks, rarefactions and contact discontinuities interact to obtain an approximate entropy solution on \([0,T] \times [0,X] \).

Notice that \( u \) is less regular than \( c \) because there are contact discontinuities emerging at each interaction between two waves.

![Figure 3. First interactions and free domain Z](image)

We use data (9) with \( N = +\infty \) i.e. \( (x_k)_{k \in \mathbb{N}} \) is an increasing sequence such that \( \lim_{k \to +\infty} x_k = X_\infty \). Then, the concentration remains bounded at the boundary but the \( BV \) norm of \( c \) and the \( L^\infty \) norm of \( u \) blow up at the boundary \( t = 0 \). Near \((t = 0, x = X_\infty)\), there is an accumulation of oscillations for the concentration and an accumulation of interactions between shocks and rarefactions.
The subset \( Z \) of \([0, T] \times [0, X_\infty]\) is a neighborhood of the vertical segment \( \{0\} \times [0, X_\infty]\). Indeed, the first interaction takes place at \((t^1, x^1)\): since a contact discontinuity propagates vertically, we have first to cut the set \( \{t > t^1, x > x^1\} \).

We do the same work, for all first interactions of the \(2N\) Riemann problems issuing from the boundary. Then \( Z = \{0 < x < X_\infty, \ 0 < t < z(x)\} \) where the function \( z : [0, X_\infty] \to [0, T] \) is piecewise constant on \([0, X]\) for any \( X < X_\infty\), non increasing, positive before \( X_\infty\), \( z(0) = T \), \( z(X_\infty) = 0 \). On \( Z \), we exactly know the solution and

\[
\lim_{(t,x) \to Z, x \to X_\infty} u(t, x) = +\infty.
\]

Let \( 0 < X < X_\infty \). By previous construction, we get an unique piecewise smooth entropy solution \( U^X \) on \([0, T] \times [0, X]\). If \( 0 < X < Y < X_\infty \), we get \( U^Y \) and by uniqueness, \( U^X = U^Y \) on \([0, T] \times [0, X]\). So, as for an ordinary differential equation, we can consider the unique maximal solution on \([0, T] \times [0, X_\infty]\). Before interaction between the solutions of the Riemann problems issuing from the boundary, we know explicitly \( U \) in \( Z \). From geometric increasing of \( u(0, x) \) when \( x \to X_\infty \), we get a blow up for \( u \) at \( t = 0 \), \( x = X_\infty \).

Furthermore, the characteristic slope \( 1/\lambda \) goes to infinity near \( \{0\} \times \{X_\infty\} \) but outside a suitable “triangular neighborhood” of this point we get a determination domain where the solution has \( BV \) regularity and there is no blow up. This determination domain has the form

\[
T = \{(t, x) : 0 < t < T, \ 0 < x < \min(X + \frac{t}{\lambda}, X_\infty)\}
\]

for any \( X \) in \([0, X_\infty]\).

6. \( BV^{1/3} \) space. We recall the classical result: shocks curves and rarefaction curves have a second order contact (see Theorem 8.2.2, page 209 in [9]). Then the amplification coefficient writes

\[
R(c_- , c_+) = 1 + O(|c_+ - c_-|^3).
\]

According to the above cubic order, we introduce a “fractional \( BV \)” space as follows:

**Definition 6.1 (\( BV^{1/3} \)).** Let \( I \) be a non empty interval of \( \mathbb{R} \). We note \( S(I) \) the set of the subdivisions of \( I \), that is the set of finite subsets \( \sigma = \{x_0, x_1, \cdots, x_n\} \subset I \) with \( x_0 < x_1 < \cdots < x_n \). Let \( u \) be a real function on \( I \). The new total variation of \( u \) on \( I \) is defined by

\[
TV^{1/3} u(I) = \sup_{\sigma \in S(I)} \sum_{i=1}^n |u(x_i) - u(x_{i-1})|^3.
\]

The set \( BV^{1/3}(I) \) is the set of functions \( u : I \to \mathbb{R} \) such that \( TV^{1/3} u(I) < +\infty \).

We define the \( BV^{1/3} \) semi-norm by \( |u|_{BV^{1/3}(I)} = (TV^{1/3} u(I))^{1/3} \).

The following inclusions hold \( BV(I) \subset BV^{1/3}(I) \subset L^\infty(I) \). More details and applications to conservation laws about spaces \( BV^s \) can be found in [6].

**Theorem 6.2.** Assume (H1)-(H2)-(H3)-(H4). If \( c_0, c_b \) belong to \( BV^{1/3} \) then there is no blow up of the velocity at the characteristic boundary.

**Sketch of proof:** we explain the proof for special set of initial and boundary value data. We use the same notations as in (9). The initial concentration \( c_0 \) in \( BV^{1/3} \) yields \( \sum k |c_{0k} - \bar{c}_{0k}|^3 < +\infty \). To compute the speed at the hyperbolic
characteristic boundary, we have the iterative relation:

\[ u_{2(k+1)} = R(c_k, \tau_k)u_{2k}. \]

Thus, \( u \) is bounded at the boundary if and only if \( \prod_k R(c_k, \tau_k) \) converges. This is achieved through (10).

**Remark 2.** If the initial concentration \( c_0 \) does not belong to \( BV^{1/3} \), the blow-up can occur.

For instance, if the sequences \( (c_k) \) and \( (\tau_k) \) converge towards the same limit but \( \sum_k |c_k - \tau_k|^3 = +\infty \) then \( \prod_k R(c_k, \tau_k) = +\infty \) and the velocity blows up at the boundary.

Because of (10), the space \( BV^{1/3} \) appears to be the critical space to control the blow-up at the boundary. So, we expect to have the following result:

**Conjecture:** if \( u_b \in L^\infty \) and \( c_0, c_b \in BV^{1/3} \) then there is existence of an entropy solution for PSA System (4)-(5) in \( BV^{1/3} \).

**REFERENCES**


E-mail address: bourdarias@univ-savoie.fr
E-mail address: gisclon@univ-savoie.fr
E-mail address: junca@unice.fr
FULLY DEVELOPED TURBULENT MIXING IN AN ANNULAR SECTOR

HYUNKYUNG LIM AND YIJIE ZHOU
Department of Applied Mathematics and Statistics, Stony Brook University
Stony Brook, NY 11794, USA

VALMOR F. DE ALMEIDA
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6181, USA

JAMES GLIMM
Department of Applied Mathematics and Statistics, Stony Brook University
Stony Brook, NY 11794, USA
and
Computational Science Center, Brookhaven National Laboratory
Upton, NY 11793-6000, USA

Abstract. We review recent progress on the characterization of turbulent fluid flow mixing and relate these ideas to high-speed, two-phase Taylor-Couette flow with application to mixing in a centrifugal contactor. Here we began to investigate the mixing and dispersion of organic/aqueous phases with the goal of predicting the underlying interfacial area. The general ideas are more broadly applicable and have been applied to the study of Rayleigh-Taylor and Richtmyer-Meshkov fluid mixing, combustion in the engine of a scram jet and the analysis of inertial confinement pellet simulations.

1. Introduction. The analysis in this article pertains to the high-speed, two-phase, Taylor-Couette (TC) flow, as applied to annular mixing in chemically reactive flows. We report on two aspects of what is an ongoing study, namely diffusion and convection. Interfacial chemical reactions in such flows are of interest in this research. Therefore the molecular diffusion coupled with turbulent convection (or turbulent dispersion), and the interfacial area are quantities that influence the bulk macroscopically observed chemical transport.

In Sec. 2, we discuss some new developments in turbulent mixing and turbulent transport. In Sec. 3, we relate the subgrid scale terms, which drive the turbulent mixing, to a renormalization group expansion, and in Sec. 4 we present simulations which study the evolving interfacial area. Conclusions are in Sec. 5.

Key words and phrases. Two-phase mixture, turbulent fluid flow mixing, Taylor-Couette flow, chemical reactions, centrifugal contactor.

This research is being performed using funding received from the DOE Office of Nuclear Energy’s Nuclear Energy University Programs. Supported in part by the Battelle Energy Alliance LLC 00088495 (sub-award with DOE as prime sponsor), Leland Stanford Junior University 2175022040367A (sub-award with DOE as prime sponsor), Army Research Office W911NF0910306, Department of Energy DEFC0206ER25770. Computational resources were provided by the Stony Brook Galaxy cluster, the Stony Brook/BNL New York Blue Gene/L IBM machine, and the Idaho National Laboratory, Center for Advanced Modeling and Simulation.
2. Theory for Turbulent Mixing. Chemical species in liquids diffuse relatively slowly, so that diffusion-limited chemical reactions are common. Vigorous stirring leads to high-Reynolds-number flow and turbulence. At high Reynolds numbers, turbulent transport is usually more important than molecular transport. In Table 1, we list molecular and turbulent properties for the TC flow to be studied in Sec. 4. The Schmidt number, the ratio of viscosity to species diffusion, is a dimensionless measure of species diffusion; it influences the rate of chemical reactions, especially diffusion-limited ones.

Since in this TC flow study mass transfer is not allowed for the time being, there is no molecular diffusion for current simulations, and we compare this flow with a Rayleigh-Taylor (RT) flow [9, 6] for the mixing of fresh and salt water at a comparable Reynolds number. See Table 1. Turbulent viscosity and turbulent species diffusion rates are mesh dependent, but in a recent study we observed quite stable values for their ratio, i.e. the turbulent Schmidt number, at high Reynolds numbers [8]. From Table 1, we conclude that turbulent viscosity plays a small role in our large eddy simulations (LES) of TC flow, but comparison to our RT flow suggests a strong role for turbulent diffusion and the turbulent Schmidt number, once chemistry is considered, in addition to the hydrodynamics.

<table>
<thead>
<tr>
<th></th>
<th>RT</th>
<th>TC (aqueous)</th>
<th>TC (organic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>molecular kinematic viscosity (cm$^2$/s)</td>
<td>$1.0 \times 10^{-2}$</td>
<td>$9.9 \times 10^{-3}$</td>
<td>$1.9 \times 10^{-2}$</td>
</tr>
<tr>
<td>turbulent kinematic viscosity (cm$^2$/s)</td>
<td>$8.0 \times 10^{-4}$</td>
<td>$3.6 \times 10^{-4}$</td>
<td>$2.3 \times 10^{-4}$</td>
</tr>
<tr>
<td>molecular diffusion (cm$^2$/s)</td>
<td>$1.8 \times 10^{-3}$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>turbulent diffusion (cm$^2$/s)</td>
<td>$2.4 \times 10^{-3}$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>molecular Schmidt number</td>
<td>$5.6 \times 10^2$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>turbulent Schmidt number</td>
<td>$3.3 \times 10^{-1}$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Reynolds number</td>
<td>$\approx 25 \times 10^4$</td>
<td>$25 \times 10^4$</td>
<td>$12 \times 10^9$</td>
</tr>
</tbody>
</table>

Table 1. Viscosity and mass diffusion for Rayleigh-Taylor and Taylor-Couette flow.

For chemically reacting flows, mixing at the atomic or molecular level is relevant, and thus the convergence of small scale fluctuations and not just mean values is important. We have begun a systematic study of convergence of solution fluctuations, based on the notion of w* convergence of a solution to a Young measure [3]. In this formulation, solution values are grouped into supercells, in a coarse graining that reduces the spatial resolution, but within each supercell, the ensemble of values is treated as a finite approximation to the fluctuating solution probability density function (PDF) of possible state values [4], thereby enhancing the resolution of the statistical fluctuations. A tool to allow ready support for w* convergence studies is available.

http://www.ams.sunysb.edu/~rkaufman/api/

We obtain numerical evidence (verification) and some experimental confirmation (validation) for the convergence of fluctuation second moments and $L_1$ norm convergence of the cumulative distribution functions (CDFs) [8, 7]. With the convergence of the full CDFs, we also obtain convergence of nonlinear functions of the flow field, such as reactive chemistry. We thus enable finite rate chemistry, in which the only chemical law added to the fluid flow is that of Arrhenius chemistry. Omitted is a range of closure models otherwise needed to capture the internal structure
of the chemical reaction; often such closure models are a prime source of solution uncertainty and error.

3. **An RNG approach to the definition of SGS terms.** We re-examine the construction of the subgrid scale (SGS) of a LES from the point of view of renormalization group (RNG) theory [2].

   The SGS terms, defined in terms of cell averages of the primitive variables, have the form of flux terms averaged over cell faces, of the form \( F(U) - F(U) \), where \( F \) is the flux term from the conservation law \( U_t + \nabla \cdot F = 0 \). Generally \( F \) is nonlinear, and the subgrid terms are nonzero. Using Favre averages, as is conventional, all terms are either quadratic or cubic in their nonlinearities for a closure of the Navier-Stokes equation [2]. If we define the SGS terms relative to a grid \( M_n \), we think of an \( L_2 \) space defined on each cell face of \( M_n \), with the \( L_2 \) norms and inner product defined relative to the density weight function \( \rho \) restricted to that face. We introduce projections \( E_n \) and \( F_n = I - E_n \) onto the space of functions constant on the face, and its orthogonal complement, respectively. In this notation, one can identify the quadratic SGS terms (unclosed) as a covariance, namely \( \langle aF_n b \rangle \) where \( a \) and \( b \) are primitive variables from the multispecies Navier-Stokes equation. Similarly, the cubic terms are closely related and defined as a cumulant. The expansion moves to a once refined grid, \( M_{n+1} \), and related projections \( E_{n+1} \) and \( F_{n+1} \), writing the unclosed term as \( \langle a(F_n E_{n+1} + F_n F_{n+1}) b \rangle \). The first term is defined on \( M_{n+1} \), and the second, which is not, can be further expanded. Continuing to all orders yields the RNG expansion for the unclosed SGS terms. A related expansion applies to the third order nonlinearity, the cumulants.

   The expansion given here is derived in [2], wherein relations to the dynamic setting of model coefficients, and resolution of solution nonuniqueness is explored.

   Assuming a K41 [5] bound on a constant density incompressible velocity field, it is easy to see that the expansion is convergent. In fact the K41 bound guarantees equicontinuity of the velocity field, so that the trace or restriction of the velocities to cell faces is well defined. The definition of the projection operator \( F_j \) forces the vanishing of Fourier coefficients taken in the surface transverse directions, for wave numbers in a bounded region, expanding as truncation error terms for higher order terms in the expansion are considered.

4. **Fully developed high-speed two-phase Taylor-Couette flow.** We are concerned with the turbulent mixing TC flow of a two-phase (organic and aqueous) mixture in a centrifugal contactor. This device is used in solvent extraction processing and other separation operations. It is critical in solvent extraction to create a large interfacial area when mixing the phases. Therefore the study of the evolution of the interface between phases from a statistical perspective is important as it sheds light on the observed, or macroscopic chemical reaction rates, and indirectly on the microscopic rates for interfacial chemistry.

   We present a preliminary characterization of the flow regime for a high-speed, two-phase Couette flow contactor. We refer to [11] for a detailed description of the contactor device and the simulation code used here to model it. The TC flow occurs in the mixing zone between two concentric rotating cylinders. In the simulation considered in this paper, the inner cylinder rotates at a high speed while the outer cylinder is fixed. The mixing of this TC flow is promoted by the shear stress and centrifugal force.
We postulate three possible flow regimes and in our simulations, at least, exclude one (global TC mixing). The remaining two regimes which have stable two continuous phases could be called centrifuge regimes, with the heavy fluid (aqueous) on the outside and the light fluid (organic based) on the inside. Our simulations are not of a scope to distinguish between large or small amounts of the dispersed phase material entrained in each of the two continuous phase regimes. Our main conclusion is that the interfacial area is significantly higher (a factor of 10) above that of a cylindrical interface. In addition, our simulations suggest an important role for turbulent diffusion in promoting rapid chemistry.

We solve the two phase incompressible Navier-Stokes equations in cylindrical coordinates using an interface tracking method to represent the sharp discontinuity and surface tension forces between the organic and aqueous phases. The detailed numerical algorithm is discussed in [1, 10, 11] with verification and validation for Couette flow [11]. We consider the two phase turbulent TC flow in a subdomain of the full contactor to reduce the computational cost. The physical fluid and geometric parameters are listed in [11]. We initialize the flow in a maximally unstable configuration, with the aqueous phase (heavy fluid) in the inner region and the organic phase (light fluid) in the outer region between the cylinders. The initial perturbed interface between the two phases is positioned to achieve nearly a 0.5 volume fraction for the organic and aqueous phases. We use periodic boundary conditions in the axial and azimuthal directions and no-slip boundary conditions at the inner and outer cylinder walls.

![Figure 1](image)

**Figure 1.** Growth of interfacial area with time in an annular sector with volume of 0.0221 cm$^3$.

Fig. 1 shows the growth of the interfacial area from a time after the initial fluctuation to a later time with a presumed statistically steady state. We observe that a transient chaotic flow regime with a high surface area is followed by a statistically
stable flow regime at a later time. This state has a surface area much higher than that of a simple cylindrical surface to divide the two fluids, but it is also significantly lower than the peak of the interfacial area during the transient regime. The transient regime has an interfacial area some seven times larger than that observed for the late time flow. The late time interface surface area is itself about 10 times that of the simple initial cylindrical interface.

The transient chaotic flow regime has significant breakup of the interface between the two phases into small droplets and thus an extensive interfacial area (Fig. 2). We found that the initial unstable configuration of the two continuous phases approaches a stable two phase configuration at a late time as a statistically steady state through the transient chaotic flow regime. The droplets are formed at the unstable interface and they migrate in their respective stable directions and then finally are segregated into the stable continuous phase. The statistically stable flow regime has two continuous phases with the organic phase on the inner region and the aqueous phase on the outer region, and with a few droplets of the dispersed phase embedded in it (Fig. 3). We call this statistically stable configuration the centrifuge mode.

In Fig. 4 and Fig. 5, we show the statistical distribution at two different time regimes. In summary, we find many small droplets, but on a volume weighted basis, they contribute little. The histogram in Fig. 4 shows the total volume of droplets vs. the droplet diameter size. Fig. 5 shows the frequency of droplets vs. the droplet diameter size. We found small droplets primarily on a transient basis only, and in a
late time statistical steady state we observe a flow segregated into distinct domains, each with a dominant fluid as a continuous phase.

The main result is the existence of two separate continuous phase regions in steady state at late time. Our conclusions are preliminary, in that they may be influenced by

1. The small simulation domain size, which may inhibit creating and breaking of interface waves, and subsequent growth of small scale mixing;
2. Influence of air as a third phase on practical application of solvent extraction;
3. Some modification of the device rotation rate, as a compensation for item 2;
4. Possible further evolution of the flow at still later times.

**Figure 4.** Total volume of droplets vs. diameter at a transient chaotic regime (above) and a statistically stable flow regime (below).
5. **Conclusions.** We have examined statistical flow regimes for a two-phase, high-Reynolds number TC mixer. Our preliminary results indicate that an almost bicontinuous regime is obtained for the current operating parameters. A residual dispersion of both phases prevails in the vicinity of the turbulent interface. This state is obtained from evolving the initial flow state and phase configuration which had the heavy phase near the inner cylinder and the organic phase near the outer cylinder. Therefore intense mixing was observed in the simulation before the bicontinuous regime was established. Further studies, especially in larger flow domains, are planned. The flow regime depends on the specific parameters of operation of the mixer.
We have sketched recent contributions to the theory and computational practice of turbulent mixing which may prove useful for further studies of this type.

REFERENCES


E-mail address: hyulim@ams.sunysb.edu
E-mail address: yzhou@ams.sunysb.edu
E-mail address: dealmeidav@ornl.gov
E-mail address: glimm@ams.sunysb.edu
FLUID SYSTEMS: APPROXIMATION BY RELAXATION AND COUPLING

Frédéric Coquel
CNRS, UMR 7641
CMAP École Polytechnique
Palaiseau, F-91128, France

Edwige Godlewski and Nicolas Seguin
UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions
CNRS, UMR 7598, Laboratoire Jacques-Louis Lions
Paris, F-75005, France

Abstract. Following a previous work [1], we are interested in the interface coupling of two (general) fluid models, differing only by their equation of state. We proposed in [5] a ‘selective relaxation’ framework for these fluid models, among which we find Euler system and the ideal MHD. Then, the (non linear) relaxation system may involve only linearly degenerate fields, which renders the Godunov scheme efficient since the solution of the Riemann problem is explicit and the resulting scheme for the fluid system is particularly simple. Moreover, the equivalence between the Eulerian and Lagrangian frames permits to derive immediately the relaxation system for the fluid model expressed in Eulerian coordinates, benefiting from the same nice properties. A particular choice of transmitted variable in the interface coupling of the relaxation systems enables us to derive a simple conservative scheme which ensures the numerical interface coupling of the two fluid models. This extends the results of [2] dedicated to the coupling of two Euler systems.

1. Introduction. Following a previous work [1] on general fluid models as described by Després in [6], we proposed recently in [5] a relaxation framework for these fluid models, among which we find Euler system and the ideal MHD. Our approach is an extension of the Suliciu approach [9] which concerns the Euler system, in particular, the relaxation is selective. In the present work, we use the equivalence in between Lagrangian and Eulerian frameworks to derive relaxation approximation from the simpler Lagrangian setting. Indeed, in the latter, because of the simple structure of the system of PDE’s representing fluid models, which involves a subset of linear equations, the relaxation approximation procedure may be performed only on the nonlinear flux terms. Then, the (non linear) relaxation system may involve only linearly degenerate fields, a situation we shall refer to as “totally linearly degenerate” (or tLD for short).

Following Yong’s results [11][12], several stability properties are proved in [5] in order to justify the relaxation procedure and its efficiency in the numerical approximation of the entropy weak solutions of the equilibrium system, i.e., of the original

2000 Mathematics Subject Classification. Primary: 35L50, 76M12; Secondary: 76N15.

Key words and phrases. Relaxation, Fluid, Finite Volume.

The second and third authors are supported by LRC Manon.
nonlinear system of PDE’s, extending the results in [4] devoted to the Euler system (see also [3]).

The Godunov scheme is efficient in the case of our tLD system for which the solution of the Riemann problem is explicit; moreover the resulting scheme for the original fluid system is particularly simple. Indeed, the Godunov solver for the homogeneous relaxation system results in an HLLC-type solver for the equilibrium system which satisfies discrete entropy inequalities under a natural Gibbs principle, valid under natural sub-characteristic conditions.

Moreover, the equivalence between the Eulerian and Lagrangian frames (see [8]) allows us to derive a relaxation system together with a numerical method for the system in Eulerian coordinates with the same nice properties.

We are now interested in extending to these general fluid models the results in [2] concerning the coupling of two Euler systems. We thus consider the situation of a one dimensional flow with a fixed interface and assume that, on each side of the interface, the flow is modeled by a fluid system, each with its own equation of state. Then some exchange of information is needed at the interface, and this can be modeled either by imposing a conservative coupling (the flux is continuous) which we call ‘flux coupling’, or by ensuring the transmission of some selected variables, which we call ‘state coupling’. We refer to [2] and references therein for more details.

The coupling of two fluid systems, written in Lagrangian variables, and only differing by their pressure law was already considered in [1], in a quite general case. We now focus on the conservative coupling of two systems in the Eulerian frame.

In Lagrangian coordinates, the interface \( x = 0 \) is characteristic, and may be thought of as a material interface between two fluids with different equations of state, the conservative coupling is physical. It can also be understood as some kind of state coupling where the choice of transmitted variables is that of the Riemann invariants associated to the null eigenvalue (think of \( u, p \) for Euler system). In [1] we detailed how to write the system of PDE in terms of these Riemann invariants for a general fluid system, then these invariants beeing continuous at the interface, the conservative coupling follows easily.

In Eulerian coordinates, a fixed interface is artificial, it has no physical meaning and we may use a state coupling approach with any choice of transmitted variables. We may also want to perform a conservative coupling. Since now 0 is not an eigenvalue, there is no associated Riemann invariant offering, as in the Lagrangian setting, an obvious choice of variables which, once transmitted, ensures a continuous flux. However, this can be done at the numerical level, by introducing the above mentioned tLD relaxation system and the resulting relaxation scheme.

The conservative coupling is performed on the relaxation systems associated to each fluid model, and the main ingredient is Godunov’s scheme. As already said, one can compute explicitly the solution of any Riemann problem on each side and at the interface, we also need to solve a Riemann problem, which is now a coupled Riemann problem (or CRP, see section 4.2 below). We solve the CRP with a well chosen set of transmitted variables. The resulting ‘full Godunov scheme’ is followed by an instantaneous relaxation step and results in a scheme for the coupling of the two original fluid systems for which the flux is continuous.

In order to construct our numerical scheme, we need to define the relaxation system for a general fluid model. We recall in section 2 this framework, first introducing the equilibrium model, then the relaxation system, skipping most details for which we refer to [5]. Section 3 is devoted to the presentation of the relaxation
scheme. The coupling of two fluid models is considered in section 4, relying on [1] [2] for a more precise description of our interface coupling approach.

2. Relaxation of fluid systems.

2.1. Quasilinear form of the equilibrium model in Lagrangian frame. Assuming some properties, such as Galilean invariance and null entropy flux, the systems we study, which we call fluid models, may be written in Lagrangian coordinates in a simple quasilinear form, equivalent for smooth solutions,

\[
\begin{aligned}
\partial_t \mathbf{v} - N \partial_u \mathbf{u} &= 0 \\
\partial_t \mathbf{u} - N^T \partial_p \phi(\mathbf{v}, s) &= 0 \\
\partial_t s &= 0,
\end{aligned}
\]

(1)

where \( \mathbf{u} \in \mathbb{R}^d \), is the velocity vector, \( \mathbf{v} \in \mathbb{R}^r \), represents the state variable vector, \( s \) the entropy, and introducing the total energy \( e = \frac{1}{2} |\mathbf{u}|^2 + \epsilon(\mathbf{v}, s) \), we have \( \phi(\mathbf{v}, s) = \epsilon_\mathbf{v} \), and \( N \) is a constant \( r \times d \) matrix. For discontinuous solutions, the entropy equation must be replaced by the energy conservation one.

The spectrum is real valued and symmetric, and if we moreover assume that \( d = r \) and that \( N \) is a square \( r \times r \) invertible matrix, there is only one null eigenvalue, associated to the conservation of entropy.

As example of (1), one finds the Euler system, with \( d = r = 1 \), \( \mathbf{v} = \tau = 1/\rho \), the specific volume, \( u \) the velocity, \( \phi = -\rho \), where \( \rho \) is the pressure law expressed in terms of \((\tau, s)\), i.e. \( p = p(\rho, \epsilon) = \rho(\tau, s) \) \( (\rho, \epsilon, s) \) denote the density, internal energy and entropy); the function \( \rho \) is usually assumed to satisfy \( \partial_t \rho > 0 \). The ideal MHD system also enters this framework (see [5]).

2.2. The Eulerian frame. Thanks to the equivalence between the Lagrangian and Eulerian settings (for which we refer to [8] and [10]), we may write system (1) in an Eulerian frame. The following result is a direct application of [8] which concerns general conservative systems and weak solutions.

**Proposition 1.** Let \((\mathbf{v}, \mathbf{u}, s)\) where \( \mathbf{v} = (\tau, v_2, \ldots, v_r)^T, \mathbf{u} = (u, u_2, \ldots, u_d)^T \), be a solution of (1) where \( N \) is a rectangular \( r \times d \) constant matrix with first line equal to \((1,0,\ldots,0)\) and \( \phi : \mathbb{R}^{r+1} \to \mathbb{R}^d \) is a smooth mapping. Assume that \( \tau > 0 \). Defining \((\overline{U}, \overline{U}) \in \mathbb{R}^r \times \mathbb{R}^d\) by

\[
\overline{U}_i = \tau = \frac{1}{\tau}, \quad \overline{U}_i = \rho u_i, \quad i = 2, \ldots, r, \quad \overline{U}_1 = \rho u_i, \quad \overline{U}_i = \rho u_i, \quad i = 2, \ldots, d
\]

(2)

setting \( U = (\overline{U}, \overline{U}, \rho s) \in \mathbb{R}^n \); defining \((\overline{F}(U), \overline{f}(U)) \in \mathbb{R}^r \times \mathbb{R}^d\) by \( \overline{F}_i(U) = \rho u_i \);

\[
\overline{f}_i(U) = -(N \overline{U}_i) + u \overline{U}_i, \quad i = 2, \ldots, r, \quad \overline{f}_i(U) = -(\rho u_i^2 - (N^T \phi_1) + u \overline{U}_i), \quad i = 2, \ldots, d
\]

(3)

where \( \phi(\overline{U}, s) = \phi(\mathbf{v}, s) \) and setting \( F(U) = (\overline{F}(U), \overline{f}(U), \rho s u) \), then \( U \) is solution of the system

\[
\partial_t U + \partial_x f(U) = 0,
\]

which writes equivalently

\[
\begin{aligned}
\partial_t \rho + \partial_x \rho u &= 0 \\
\partial_t \overline{U}_i + \partial_x (\rho u_i^2 - (N^T \phi_1) + u \overline{U}_i) &= 0, \\
\partial_t \rho u_i + \partial_x (\rho u_i^2 - (N^T \phi_1)) &= 0 \\
\partial_t \overline{U}_i + \partial_x (\rho u_i^2 - (N^T \phi_1) + u \overline{U}_i) &= 0, \\
\partial_t \rho s + \partial_x \rho s u &= 0.
\end{aligned}
\]

(4)
2.3. Relaxation system in Lagrangian frame. In order to approximate the nonlinear fluid system we start from its simple form (1), equivalent for smooth solutions (a now straightforward correction may be added later on in order to extend the approach to discontinuous solutions for which the energy conservation equation must be taken in place of the trivial entropy conservation equation, details can be found in [5]). We introduce a larger relaxation system, obtained by introducing new dependent variables $V$ found in [5]). We introduce a larger relaxation system, obtained by introducing new dependent variables $V$ as some relaxation parameter goes to 0 (in fact we write the parameter in the form $\frac{1}{\epsilon}$ and thus $\lambda \rightarrow \infty$), and such that the limit of the relaxation system (as $\lambda \rightarrow \infty$) reduces to (1). Such a relaxation procedure is now classical for Euler system (see references in [4]).

Because of the trivial entropy equation, we first consider the nonlinear system (1) in the isentropic case. In order to simplify the notations, we set $\epsilon = \epsilon(v)$, $e_v = e'$, thus (1) reads

\[
\begin{align*}
\partial_t v - N \partial_x u &= 0 \\
\partial_t u - N^T \partial_y e'(v) &= 0.
\end{align*}
\]

The total energy now takes the place of a mathematical entropy since it is a convex function of $(v, u)$, provided the matrix $e_{v,v}(v) = e''(v)$ is (symmetric) positive-definite, which we will assume thereafter.

For the approximation of system (5), we consider the larger relaxation system of $2r + d$ equations

\[
\begin{align*}
\partial_t v - N \partial_x u &= 0 \\
\partial_t u - N^T \partial_y V &= 0 \\
\partial_t V &= \lambda(v - V)
\end{align*}
\]

with

\[
W = W(v, V) = e'(V) + \theta'(v) - \theta'(V),
\]

where $V \in \Omega_v \subset \mathbb{R}^r$, and $\theta : \mathbb{R}^r \rightarrow \mathbb{R}$ is some $C^2$ mapping on which we will make more assumptions hereafter. We now note $U = (v, u, V)^T$, and $\Omega \subset \Omega_v \times \mathbb{R}^d \times \mathbb{R}^r$ the set of states $U$. In the sequel, we will partition $(2r + d) \times (2r + d)$ matrices in blocks relative to the above decomposition of system (6): $\mathbb{R}^{2r+d} = \mathbb{R}^r \times \mathbb{R}^d \times \mathbb{R}^r$. Also, we note either $\nabla_v \theta, \theta_v$ or $\theta'$ the derivative of $\theta$ (similarly for $e$). For the system with entropy, all the computations could be done in exactly the same way, for instance $\theta = \theta(v, s)$ would depend on $s$, only replacing the derivatives by partial derivatives.

Formally, as the relaxation parameter $\lambda \rightarrow \infty$, $v - V \rightarrow 0$ and at equilibrium, $W(v, V) = e'(V)$, we recover system (5) (recall we have assumed we are in the isentropic case). We will denote by $U_{eq}$ the manifold of equilibrium states $U_{eq} = (v, u, V)^T$ i.e. states satisfying $V = v$. We need to make assumptions in order that this ‘relaxation procedure’ is indeed stable. Note that for a state at equilibrium, $V \in \Omega_v$, otherwise $V$ is only supposed to lay in a neighborhood of $\Omega_v$ in $\mathbb{R}^r$.

The projection operator is denoted when necessary by $P : U = (v, u, V)^T \mapsto (v, u)$ and the equilibrium mapping by $M : (v, u, V) \mapsto (v, u, v)$. The source term is written $\lambda G(U)$ with $G(U) = (0, 0, v - V)^T$, thus satisfying $PG(U) = (0, 0)^T$.

2.4. Stability results. From now on, we assume that $\theta$, as $e$, is convex i.e., we also assume $\theta''(V)$ is symmetric positive definite, we will make below a more precise assumption. We often identify these forms with matrices and note equivalently $\theta''(V)v, v$ or $v^T \theta''(V)v$.

If we take a quadratic $\theta, \theta(v) = \frac{1}{2}(v, \Lambda v)$ with $\Lambda$ a positive definite $r \times r$ constant matrix, then $\theta''(V) = \Lambda$ is constant and the system is linearly degenerate.
This property, which is not needed for the stability results we recall now, is however a keypoint for the efficiency of our numerical scheme. Then (7) is indeed a linearization of the nonlinear term $\epsilon'(v)$.

We introduce the mathematical entropy (which is in fact an energy), noted $\Sigma(v,u,V)$, which is dissipated, and coincides with the entropy of the system at equilibrium, i.e. the total energy $e$ of the system at equilibrium which writes

$$\Sigma(v,u,V) = \frac{1}{2}|u|^2 + \epsilon(V).$$

Define

$$\Sigma(v,u,V) = \frac{1}{2}|u|^2 + \epsilon(V) + \theta(V) - \theta(V) + ((\epsilon' - \theta')(V), v - V).$$

(8)

Lemma 2.1. Let the function $\Sigma$ be defined by (8); smooth solutions of (6) satisfy

$$\partial_t \Sigma - \partial_x (u, N^T W - \Theta u) = \lambda((\epsilon'' - \theta'')(V)(v - V), v - V).$$

(9)

In order to check the entropy dissipation, since the right hand side in (9) is negative if the matrix $\theta'' - \epsilon''$, which is obviously symmetric, is positive definite, we shall thus make the (rough) assumption

(H) $\theta''(V) - \epsilon''(V)$ is positive definite for all $V$ under consideration.

Under some more precise hypothesis (on the set of $V$ where (H) holds), one can prove that $\Sigma$ is strictly convex on the set of equilibrium states; also a minimum principle holds, and one gets entropy dissipation together with Chapman-Enskog dissipativity. Finally, one can prove an approximation result, which says, roughly speaking, that the solution of (5) may be approximated (for large $\lambda$) by the projection of the solution of (6), again, we refer to [5] for details.

2.5. Riemann problem. We have in mind the use of Godunov’s scheme and begin by an easy result which will allow us to compute explicitly the solution of the Riemann problem.

Lemma 2.2. Assume that $\theta''$ is constant. Then, all the characteristic fields of the homogeneous part of system (6) are linearly degenerate.

Proof. We assume $\theta''$ is constant and thus $\Theta = N^T \theta''(v) N$ is constant too. We check easily that all solutions of the system (6) also satisfy

$$\partial_t N^T W - \Theta \partial_x u = 0.$$ 

(10)

If we consider this equation in place of the first equation in (6), since all the matrices are constant, the resulting homogeneous system is linear. The mapping $v \mapsto W(v,V) = \epsilon'(V) + \theta'(v) - \theta'(V)$ is linear ($\theta''$ is constant) and invertible from the set of state variables ($v$ is a state variable) on its range since $W_v = \theta''$ is positive definite. Hence $(N^T W, u, V)$ defines an admissible change of variables, and thus all the fields of system (6) are also linearly degenerate.

Then, the wave curves (involved in the solution of the Riemann problem) are integral curves of the fields of eigenvectors, they may be found by writing the Rankine–Hugoniot relations across a discontinuity propagating with speed $\sigma$. From the equations

$$\begin{cases}
-\sigma[u] - N^T[W] = 0, \\
-\sigma N^T[W] - \Theta[u] = 0.
\end{cases}$$
we get
\[ \sigma^2[u] = -\sigma N^T[W] = \Theta[u], \]
hence \([u]\) appears as an eigenvector of \(\Theta\) associated to \(\sigma^2\), then the first Rankine–Hugoniot equation will give \([v]\) across \(x = \sigma t\). The matrix \(\Theta\) is positive definite and has \(d\) positive eigenvalues \(\sigma_i^2 > 0\), an orthonormal basis of eigenvectors, say \(s_i, 1 \leq i \leq d\). The linear system
\[
\begin{cases}
\partial_t N^T W - \Theta \partial_x u = 0 \\
\partial_t u - N^T \partial_x W = 0
\end{cases}
\] (11)
has a matrix which is diagonalizable in a basis of the form \((\pm \sigma_i s_i, -s_i), 1 \leq i \leq d\). For the full system
\[
\begin{cases}
\partial_t N^T W - \Theta \partial_x u = 0 \\
\partial_t u - N^T \partial_x W = 0 \\
\partial_t V = 0
\end{cases}
\] (12)
setting \(V = (N^T W, u, V)\), we have, with this choice of dependent variables, a Jacobian matrix with eigenvectors \((\pm \sigma_i s_i, -s_i, 0), 1 \leq i \leq d\), each associated to an eigenvalue \(\pm \sigma_i\). And, associated to 0, are the vectors \(r_i = (0, 0, e_i), 1 \leq i \leq r\), with \(e_i = (0, \ldots, 1, 0, \ldots, 0)^T\) and 1 in ith position. We call all the eigenvectors \(R_i\) (they form a basis) and let \(I\) be a dual basis. For a given initial condition \(V_0 = (N^T W_0, u_0, V_0)\), the explicit solution of the Cauchy problem is given by
\[ V(x, t) = \sum_i (I, V_0)(x - \sigma_i t)R_i. \]
We can define \(v\) from \(W\), and thus \(U(x, t)\) from \(V(x, t)\).
Practically, the Riemann problem will be solved for given left and right states at equilibrium, which defines immediately \(V\) everywhere since it jumps only across the field \(\mu = 0\); and for the determination of the \(d + r\) remaining components of the \(2r\) intermediate states, we use that both \(Nu + \sigma v\) and \(\sigma u+N^T W\) are continuous across the wave of speed \(\sigma\) (it gives \((r+d) \times 2r\) equations minus \(2d\) equations for the continuity of the \(\sigma_i\)’s), and naturally the fact that \(u\) and \(N^T W\) are continuous across the field \(\mu = 0\) (\(2d\) equations).

2.6. The Eulerian relaxation model. We now use the results of section 2.3 to write our relaxation model in Eulerian coordinates. We state without proof the analogue of proposition 1.

**Proposition 2.** Let \((v, u, s, V)\) where \(v = (\tau, v_2, \ldots, v_r)^T, u = (u_2, \ldots, u_d)^T\) be a solution of
\[
\begin{cases}
\partial_t v - N \partial_x u = 0 \\
\partial_t u - N^T \partial_x W = 0 \\
\partial_t s = 0 \\
\partial_t V = \lambda(v - V)
\end{cases}
\] (13)
with
\[ W = W(v, V, s) = \phi(V, s) + \theta'(v) - \theta'(V), \]
where \(N\) is a rectangular \(r \times d\) constant matrix with first line equal to \((1, 0, \ldots, 0)\), \(\phi : \mathbb{R}^{r+1} \to \mathbb{R}^d\) and \(\theta : \mathbb{R}^r \to \mathbb{R}\) are smooth mappings. Assume that \(\tau > 0\). Defining \((U, U) \in \mathbb{R}^r \times \mathbb{R}^d\) by
\[ U_1 = \phi \equiv \frac{1}{\tau}, U_i = qv_i, i = 2, \ldots, r; U_1 = qu, U_i = qu_i, i = 2, \ldots, d \] (15)
setting \( U = (\overline{U}, U_s, q, qV) \in \mathbb{R}^{n+r} \); defining \((\overline{f}(U), f(U)) \in \mathbb{R}^r \times \mathbb{R}^d \) by

\[
\overline{f}_1(U) = qu, \quad \overline{f}_i(U) = -(N^T U)_{\overline{i}}, i = 2, \ldots, r,
\]

\[
f_i(U) = -(N^T \tilde{W})_{\overline{i}} + u_{\overline{i}}, i = 1, \ldots, d
\]

(16)

where

\[
\tilde{W} = \tilde{W}(U, V, s) = \mathcal{W}(v, V, s) + \overline{\theta}'(V) - \overline{\theta}'(V),
\]

and setting \( f(U) = (\overline{f}(U), f(U), qsu, quV) \), then \( U \) is solution of the system

\[
\partial_t U + \partial_x f(U) = r,
\]

(17)

where \( r = (0, 0, 0, \lambda \phi(\mathcal{V} - V))^T \), which writes equivalently

\[
\begin{align*}
\partial_t q + \partial_x q u &= 0 \\
\partial_t U_i + \partial_x (\frac{1}{2} (N^T U)_{i} + u U_i) &= 0, \quad i = 2, \ldots, r, \\
\partial_t q u + \partial_x (g u^2 - (N^T \overline{W})) &= 0, \\
\partial_t U_i + \partial_x (\frac{1}{2} (N^T \overline{W})_{i} + u U_i) &= 0, \quad i = 2, \ldots, d \\
\partial_t q s + \partial_x q s u &= 0 \\
\partial_t q V + \partial_x q u V &= \lambda \phi(V - V).
\end{align*}
\]

Note that we have stated the result with a separate treatment of the variable \( V \) for clarity; it might have been incorporated in some augmented \( \overline{U} \).

Let us precise the expression of the energy. Recall (8) and define \( \overline{\Sigma}(U) = \Sigma(v, U_s, V) \) where \( U \) is given in terms of \( (v, U_s, V) \) in proposition 2, which writes

\[
\overline{\Sigma}(U) = \frac{1}{2} |u|^2 + \epsilon(V, s) + \theta(V) - \theta(V) + ((\epsilon(V, s) - \theta(V))(V - V)).
\]

(19)

In Eulerian coordinates, smooth solutions of (18) satisfy (dropping the notation \( \tilde{\overline{\Sigma}} \))

\[
\partial_t q \Sigma + \partial_x \left( - (u, N^T W) + q u \Sigma \right) = \lambda \left( (\epsilon V, V, s) - \theta''(V)\right)(v - V), v - V,
\]

and thus weak solutions of (18) for \( \lambda = 0 \) also satisfy the energy conservation equation

\[
\partial_t q \Sigma + \partial_x \left( - (u, N^T \overline{W}) + q u \Sigma \right) = 0.
\]

(20)

Now, let us recall that linearly degenerate fields in the Lagrangian frame give linearly degenerate fields in Eulerian coordinates (see also [7]). Thus one obtains directly

**Lemma 2.3.** Assume that \( \theta'' \) is constant. All the characteristic fields of the relaxation system (18) are linearly degenerate.

We state the following result which implies that we can use Godunov’s scheme for the homogeneous part of the relaxation system.

**Proposition 3.** Given two constant states \( U_l, U_r \), the solution \( W_R(x/t; U_l, U_r) \) of the Riemann problem for (18) consists of (at most) \( 2r + 1 \) contact waves propagating at speed \( u \pm \tau \sigma_i, i = 1, \ldots r \) and \( u \) (with multiplicity \( d + 1 \)), where the \( \sigma_i^2 \) are the eigenvalues of the matrix \( \Theta = N^T \theta'' N \); these waves separate constant states which coincide with those of the Riemann problem for (13) when expressed in terms of the same variable \( U \).

**Example:** Euler system.

For the full Euler system, we deduce from (18) that the relaxation system writes
\[
\begin{align*}
\partial_t \rho + \partial_x (\rho u) &= 0 \\
\partial_t (\rho u) + \partial_x (\rho u^2 + \Pi) &= 0 \\
\partial_t \rho s + \partial_x (\rho su) &= 0 \\
\partial_t (\rho T) + \partial_x (\rho Tu) &= \lambda \rho (\tau - T)
\end{align*}
\]

(21)

with (using the notation $-\Pi$ instead of $\mathcal{W}$)

\[
\Pi = \tilde{\Pi}(\tau, s, T) \equiv \tilde{p}(T, s) + a^2(T - \tau).
\]

(22)

In this extended Euler system, the pressure law is linear in $\tau$, at the expend of adding a new variable $T$, which itself is simply convected with the flow, thus leading to a linearly degenerate system (for what concerns the homogeneous part), for which $u$ is a double eigenvalue, and the other eigenvalues are $u \pm a\tau$ since the Eulerian ‘sound speed’ $c$ is given by $c^2 = -\tau^2 \partial_\tau \Pi$.

3. The relaxation solver. We come to the numerical approximation procedure for the fluid system. For simplicity, we assume $d = r$, $N$ is a square invertible matrix. We use classical notations for a finite volume scheme, and, starting from $(u_j^0, v_j^0), j \in \mathbb{Z}$ given by the means of the initial data $u_0(x), v_0(x)$ on a uniform mesh $(x_{j-1/2}, x_{j-1/2})$, we describe the updating $(u_j^n, v_j^n) \rightarrow (u_j^{n+1}, v_j^{n+1})$.

3.1. Description. The numerical approximation may be described by three steps: (i) a reconstruction step involving the equilibrium mapping $(u_j^n, v_j^n) \mapsto \mathcal{M}(u_j^n, v_j^n)$ and a piecewise constant reconstruction; (ii) an evolution step involving the exact solution of a juxtaposition of Riemann problems for the relaxation system (6) (with $\lambda = 0$), and (ii) a projection step with instantaneous relaxation ($\lambda \rightarrow \infty$) to get an equilibrium state, followed by a projection on piecewise constants and retaining only its projection part $\mathcal{P}$.

This last step needs a little tricky complement for the full system with energy, however, the resulting numerical scheme is indeed very simple. The way it is built enables us to obtain good stability properties. The approach is thoroughly described in [5] starting from the much simpler Lagrangian framework.

3.2. Relaxation solver in Eulerian coordinates. We may deduce the properties from the equivalence between the two Lagrangian/Eulerian formulations at the discrete level following [8]. Indeed, since the solver is an exact Godunov solver for the relaxation system, they can be obtained from the equivalence between the two Lagrangian/Eulerian formulations at the continuous level. We do not give the details which can be found in [5], and only state the final result.

Let us note here $U = (\mathcal{U}, \underline{U}, ge)$ the state variable of the fluid system, $V = (\mathcal{U}, \underline{U}, gs, g\mathcal{V})$ the state variable of the augmented relaxation system. We note again $\mathcal{M}$ the equilibrium mapping $\mathcal{M}(U) = V = (\mathcal{U}, \underline{U}, gs, g\mathcal{V})$ which allows to reconstruct an equilibrium state, where $s = s(U)$.

**Proposition 4.** Under CFL 1/2, the resulting global relaxation scheme can be written

\[
U_{j+1}^n = U_j^n - \mu \left( G_{j+1/2}^n - G_{j-1/2}^n \right), \quad j \in \mathbb{Z}, n \geq 0,
\]

(23)

with the components of the flux $G_{j+1/2}$ given by the corresponding flux components of (18), (20) evaluated on state $W_{j+1/2}^n$,

\[
W_{j+1/2}^n = W_{R}(0+; \mathcal{M}(U_j^n), \mathcal{M}(U_{j+1}^n)),
\]

(24)
and $W_R$ denotes the solution of the Riemann problem for (18) given by proposition 3. Moreover, the global relaxation solver satisfies a discrete entropy inequality

$$
(\rho s)^{n+1}_j \leq (\rho s)^n_j - \mu (G^n_{\rho s,j+1/2} - G^n_{\rho s,j-1/2}), \quad j \in \mathbb{Z}.
$$

(25)

4. Coupling. We now come to the coupling of two fluid systems at a fixed interface.

4.1. Framework. We assume that there is some fixed interface across which the ‘pressure law’ $\phi(v, s) = \epsilon_v$ of the fluid model (in Eulerian coordinates) varies, then we have two systems (4) with slightly different ‘pressure laws’

$$
\phi = \begin{cases} 
\phi_L, & x > 0, \\
\phi_R, & x > 0,
\end{cases}
$$

(26)

thus, with the notations of Proposition 2 two different fluxes $f_L$ and $f_R$. We introduce the two respective relaxation systems (18) with corresponding expressions for $W$ in (14).

Then, we consider the resulting relaxation scheme on each side of the interface. This numerical scheme involves the usual Godunov scheme for the homogeneous parts of (18) on each side. For simplicity, we assume here the mesh such that the interface is at $x_{1/2} = 0$. For the flux at the interface $x = 0$, we also consider the Godunov scheme, which means that we solve the coupled Riemann problem (CRP for short) which we introduce now.

4.2. Coupled Riemann problem. We consider the Riemann data

$$
U = \begin{cases} 
U_L, & x < 0 \\
U_R, & x > 0
\end{cases}
$$

(27)

where $U_{L,R}$ are two given constant states. Solving the coupled Riemann problem means solving the problem (18)/(14) together with (26) and with data (27). As presented in [2], this problem is considered as two initial boundary value problems, one on each side of the interface $x = 0$, and the boundary (Dirichlet) conditions are imposed in a weak form, as detailed in [1]. Moreover, it is explained in [1] that we can choose to characterize the (Dirichlet) boundary data in any admissible set of variables, using for instance ‘primitive’ variables, a fact which we refer to as transmission of primitive variables.

The conservative coupling will be ensured if we choose carefully the variables which we transmit when solving the CRP which means in such a way that the left and right fluxes taken on the solution of the CRP at $x = 0$ coincide. Because of the linear degeneracy of the system, the computations are again explicit.

**Example:** Euler system (revisited). In [2], we ensure the numerical conservative coupling of two Euler systems differing by their pressure law $p = p_L(\rho, \varepsilon) = \tilde{p}_L(\tau, s)$ in $x < 0, p = p_R(\rho, \varepsilon) = \tilde{p}_R(\tau, s)$ in $x > 0$. We introduce the two corresponding relaxation systems (21)/(22), and solve the CRP with transmission of the primitive variables $V = (\tau, u, \Pi, \Sigma)^T$. Then the three components $(\rho u, \rho u^2 + \Pi, \rho \Sigma + \Pi)$ of the flux are continuous, resulting in a conservative scheme for the numerical coupling of the two Euler systems.

4.3. Choice of transmitted variables for the general fluid system. In order to couple numerically the two fluid models, there only remains to precise the choice of the transmitted variables involved when solving the coupled Riemann problem. What we did for the Euler system extends naturally; we choose, using the notations of Proposition 2, the transmission of the $3d$ variables $\rho, v_i, i = 2, \ldots, d, u, u_i, i =$
2, ... d, W. Note that this defines indeed an admissible change of variables, the mapping \( V \mapsto W \) is invertible under assumption (H).

The analogue of Theorems 2 and 3 in [2] holds. The resulting scheme can be written in conservative form, which means we get the analogue of Proposition 4 where in (24) we solve the Riemann problem for \( f_L \) in \( x < 0 \), i.e. for \( j < 0 \) (resp. \( f_R \) in \( x > 0 \), i.e., for \( j > 0 \)) and for the flux at the interface \( x_{1/2} = 0 \), i.e., for \( j = 0 \), we consider the solution of a CRP, with the above choice of transmitted variables which leads to a unique flux.

Moreover, a discrete entropy inequality holds on each side of the interface, and a formal Lax-Wendroff type theorem says that if the scheme converges (in a usual sense, see [2], which in practice one seldom gets) the solution is a weak solution of a conservative system (4) with \( \phi = (1 - H(x)\phi_L + H(x)\phi_R \) (\( H \) the Heaviside function).

REFERENCES


E-mail address: frederic.coquel@cmap.polytechnique.fr
E-mail address: edwige.godlewski@upmc.fr
E-mail address: nicolas.seguin@upmc.fr
SLOW EROSION OF GRANULAR FLOW: 
CONTINUOUS AND DISCONTINUOUS PROFILES

DEBORA AMADORI
DISIM, Università degli Studi dell’Aquila
via Vetoio, 1 67010 Coppito, L’Aquila, Italy

RINALDO M. COLOMBO
Unità INdAM c/o DI, Università degli studi di Brescia
Via Branze 38 25123 Brescia, Italy

GRAZIANO GUERRA
Dip. Matematica e Applicazioni, Università di Milano-Bicocca
via R. Cozzi, 53 20125 Milano, Italy

WEN SHEN
Mathematics Department, The Pennsylvania State University
University Park, PA 16802, USA

Abstract. We consider an integro-differential conservation law that models
the slow erosion of the granular flow. In the model, the flux contains an integral
term in the space variable. Depending on the assumptions on the erosion rate,
the solutions exhibit various types of singularities, including blowing up of
the gradient. We establish existence of BV solutions and their continuous
dependence on the data, obtaining a Lipschitz semigroup. For the case with
continuous profile, the solutions are unique.

1. Introduction. We study an integro-differential equation that describes the slow
erosion of the granular flow
\[
U_t(t, x) - \left( \exp \int_x^{+\infty} f(U_x(t, y)) \, dy \right)_x = 0.
\] (1)

Here \(x\) is the space variable, \(U\) is the height of the profile, and \(t\) represents the total
mass of passing moving layer. We assume that the slope of the profile \(U_x\) is strictly
positive. The interaction between the two layers is described by the erosion function
\(f\), which denotes the rate of mass being eroded or deposited per unit length in \(x\)
and per unit mass passing through. We assume that \(f\) depends only on the slope,
and we let \(w = U_x\) denote the slope. There is a critical slope, normalized to be 1,
where no interactions occur, i.e. \(f(1) = 0\). If the slope is bigger than 1, then \(f > 0\)
and we have erosion, so the moving layer grows. On the other hand, if the slope is
smaller than 1, then \(f < 0\) and part of the moving layer deposits onto the standing
bed.

2000 Mathematics Subject Classification. Primary: 35L65; Secondary: 35L67, 35Q70, 35L60,
35L03.

Key words and phrases. Slow Erosion, Granular Flow, Integro-differential Equation, Lipschitz
Semigroup, Wave Front Tracking, Conservation Laws.
The model (1) was first derived as the slow erosion limit of a $2 \times 2$ model for granular flow by Hadeler-Kuttler [8]. In [1] we prove that, as the thickness of the moving layer tends to 0, the solutions for the height of the standing profile in the $2 \times 2$ model satisfy the scalar integro-differential equation in (1), with the erosion function $f(w) = (w - 1)/w$. We remark that here $U(t, x)$ describes the strictly increasing asymptotic profile at $x$ after $t$ unit mass of material was slowly poured from an uphill location outside the interval of interest.

Wider classes of erosion functions are then studied. In general, the erosion function $f$ is a non-linear concave function, which satisfies the following assumptions:

$$f \in C^2([0, +\infty[; \mathbb{R}), \ f(1) = 0, \ f' \geq 0, \ f'' \leq 0, \ \lim_{w \to 0^+} f(w) = -\infty. \quad (2)$$

Due to the nonlinearity of the function $f$, solutions of (1) may well lose regularity. The types of singularities that could form in the solutions depend on the additional properties of the erosion function $f$ for large slope value $w$. We consider two different cases, that exhibit different types of singularities.

**Case 1:** We assume that the erosion function $f$ grows slowly for large slope, i.e.,

$$\lim_{w \to +\infty} f(w)/w = 0. \quad (3)$$

In this case, it is observed in [3] that the slope $w = U_x$ remains uniformly bounded in $t$, and no shocks would form in $U$. Thanks to this uniform bound, one can instead study the conservation law for the slope. Differentiating (1) in $x$, we get

$$w_t(t, x) + \left( f(w(t, x)) \exp \int_x^{+\infty} f(w(t, y)) \, dy \right)_x = 0. \quad (4)$$

The solutions $w(t, x)$ could develop discontinuities in finite time, which means that kinks could form in the profile $U(t, x)$. Section 2 is devoted to the existence, well-posedness and stability of BV solutions to (4) under conditions (2) and (3).

**Case 2:** If the erosion rate grows sufficiently fast for large slopes, the gradient of the solutions to (1) could blow up, leading to stronger singularities such as hyper-kinks and vertical jumps in the profile $U(t, x)$, see [9]. In this case, it would not be suitable to study (4) since the distributional derivative $w = U_x$ contains point masses. However, if we assume (2) and the additional assumption that $f$ approaches a linear asymptote for large $w$, i.e.,

$$\lim_{w \to +\infty} f'(w) > 0, \quad \lim_{w \to +\infty} \left( f(w) - w f'(w) \right) < +\infty, \quad (5)$$

one can establish the existence of global BV solutions, see [9].

We remark that several other examples of scalar conservation laws with integro-differential terms have been considered in the literature. Of particular relevance are the well-studied Camassa-Holm equation [5] and the variational wave equation (see for example in [4]). In both cases, the gradients of the solutions may blow up. But thanks to a balance law for the quantity $u_x^2$, the solution $u(t, \cdot)$ remains Hölder continuous at all times.

In contrast, solutions to our equation (1), in Case 2, may well develop jumps. Consequently, the definition of the integral term needs to be carefully chosen, to avoid evaluating a nonlinear function on a measure. An appropriate definition is given in [9], along with the proof that solutions to the Cauchy problem exist globally in time.
In Section 3, we establish continuous dependence for the solutions of (1) under conditions (2) and (5), showing the existence of a Lipschitz semigroup.

2. **Case 1: Solutions with Continuous Profile.** We now assume that \( f \) satisfies the assumptions (2) and (3), and study the equation (4). For notation simplicity, we use \( F(x;\nu) \) to denote the integral term, i.e.,

\[
F(x;\nu) = \exp \int_x^{+\infty} f[w(t,y)] \, dy.
\] (6)

Solutions of the Cauchy problem would be obtained within the class \( W \) consisting of all functions \( \nu : \mathbb{R} \mapsto \mathbb{R} \) satisfying the properties:

\[
\inf \nu(x) > 0, \quad \nu(\cdot) \in BV(\mathbb{R}), \quad \text{supp}(\nu(\cdot) - 1) \text{ is bounded}. \] (7)

Entropy weak solutions for (4) are defined through the weak formulation, with an additional Oleinik type entropy condition.

**Definition 2.1.** A function \( \nu : [0,T] \times \mathbb{R} \mapsto \mathbb{R} \) is called an entropy weak solution for (4) with initial data \( \bar{\nu}(x) \in W \) if

- \( [0,T] \ni t \mapsto \nu(t,\cdot) \in W \); is continuous in \( L^1_{loc} \) and \( \nu(0,\cdot) = \bar{\nu} \).
- For every test function \( \phi \in C^\infty_c(\mathbb{R}^2) \), one has the integral identity

\[
\int_0^T \int_{\mathbb{R}} \left( \nu\phi_t + f(\nu)F(x;\nu)\phi_x \right) \, dx \, dt = \int_{\mathbb{R}} \left( w(T,x)\phi(T,x) - w(0,x)\phi(0,x) \right) \, dx.
\]

- For a.e. \( x < y \), there exists some constant \( C \) (that does not depend on \( x, y \) or \( t \)), such that

\[
\nu(t,x+) - \nu(t,y+) \leq C \max \{1/t, 1\} (y-x).
\] (8)

The following existence and uniqueness result for solutions of (4) is proved in [2].

**Theorem 2.2.** [2, Theorem 1.2 & 1.3] Let \( T > 0 \) and an initial data \( \bar{\nu} \in W \) be given. Then the Cauchy problem for (4) admits a unique entropy weak solution \( \nu = \nu(t,x) \) defined for all \( t \in [0,T] \), that moreover satisfies

\[
\inf \nu(t,x) \geq \inf \bar{\nu}(x), \quad \nu(t,\cdot) \in BV(\mathbb{R}), \quad ||\nu(t,\cdot) - 1||_{L^1} \leq ||\bar{\nu} - 1||_{L^1}. \] (9)

Moreover, let \( \bar{f} \) be a function that satisfies (2) and (3) and let \( \nu \) be the entropy weak solution for the Cauchy problem

\[
v_t + \left( \bar{f}(\nu) \exp \left\{ \int_x^\infty \bar{f}(\nu(t,s)) \, ds \right\} \right)_x = 0, \quad \nu(0,x) = \bar{\nu}(x) \in W. \] (10)

Letting \( \kappa_0 > 0 \), \( M \) and \( m_0 \) be the constants that satisfy

\[
\inf \bar{\nu}(x) \geq \kappa_0, \quad TV \nu(t,\cdot) \leq M, \quad ||\bar{\nu} - 1||_{L^1} \leq m_0
\]

and the analogous bounds for \( \nu \), where \( M \) depends on \( T \), we have

\[
||\nu(t,\cdot) - \nu(t,\cdot)||_{L^1} \leq ||\bar{\nu} - \bar{\nu}||_{L^1} + C \int_0^t ||\nu(s,\cdot) - \nu(s,\cdot)||_{L^1} \, ds
\]

\[
+ Ct \left[ ||\bar{f} - \bar{f}||_{L^\infty} + ||\bar{f}' - \bar{f}'||_{L^\infty} \right], \] (11)

where \( C \) depends only on \( \kappa_0, M \) and \( m_0 \).
The existence of solution is established through piecewise constant approximate solutions generated by a modified front tracking algorithm, which we explain in some detail. Along characteristic lines \( t \mapsto x(t) \), we have
\[
\frac{d}{dt} x(t) = f'(w) F(x; w), \quad \frac{d}{dt} w(t, x(t)) = f^2(w) F(x; w). \tag{12}
\]
Note that the slope \( w \) increases in \( t \) along characteristics. This is different from standard conservation laws where the conserved quantity remains constant along characteristics, see \cite{7}. Such a difference is reflected in our modified front tracking algorithm. In addition to the propagation of the fronts of discontinuities, also the constant value between two fronts changes in \( t \).

The \( \epsilon \)-front tracking approximate solutions are generated by two sets of ODEs, with one set for the locations of all the fronts, and the other for the constant values between fronts. Let \( (x_i(t), w_i(t)) \) be the discrete data set, where \( (w_i(t), w_{i+1}(t)) \) are the left and right constant states for the front at \( x_i(t) \). The propagation speed of the front \( x_i \) satisfies the Rankine-Hugoniot condition,
\[
\dot{x}_i = F(x_i; w) \frac{f(w_i) - f(w_{i+1})}{w_i - w_{i+1}}. \tag{13}
\]
The evolution of constant state \( w_i(t) \) follows the ODE
\[
\dot{w}_i(t) = -f(w_i) \frac{F(x_i; w) - F(x_{i-1}; w)}{x_i - x_{i-1}}. \tag{14}
\]
Since the function \( f \) is concave, only upward jumps in \( w \) are admissible. The discrete data set \( (x_i, w_i) \) is initialized such that all non-admissible fronts are small. Furthermore, to ensure accuracy, the intervals \( (x_i, x_{i+1}) \) are chosen small enough such that the two neighboring integral terms \( F(x_i; w) \) and \( F(x_{i+1}; w) \) are sufficiently close.

Here we give a formal argument for some of the bounds. We first observe that \( w \equiv 1 \) is a trivial solution. The equation (4) could be written as
\[
(w - 1)_t + (f(w) F(x; w))_x = 0. \tag{15}
\]
Since \( f(w) \) and \( w - 1 \) have the same sign, then the \( L^1 \)-norm of \( w - 1 \) is non-increasing in \( t \), i.e.,
\[
\|w(t, \cdot) - 1\|_{L^1} \leq \|w(0, \cdot) - 1\|_{L^1} = m_0. \tag{16}
\]
As \( F > 0 \), from (12) we see that \( w \) increases along characteristics, so a lower bound follows. To obtain the uniform upper bound, we integrate the conservation law (15) over the region \( (t, y) \) with \( 0 \leq t \leq T \) and \( y \leq x(t) \) where \( t \mapsto x(t) \) is a characteristic. By divergence theorem, we have
\[
\left| \int_0^T [(w - 1) f'(w) - f(w)] F(x; w)_{x=x(t)} \ dt \right|
= \left| \int_{x(T)}^{x(0)} (w(T, x) - 1)dx - \int_{x(0)}^{x(0)} (w(0, x) - 1)dx \right| \leq 2m_0. \tag{17}
\]
Now, define an auxiliary function
\[
\alpha(w) = \frac{w - 1}{f(w)} \quad \text{if} \ w \neq 1, \quad \alpha(1) = 1/f'(1). \tag{18}
\]
Observe that, along the characteristic, we have
\[
\frac{d}{dt} \alpha(w(t, x(t))) = \alpha'(w) \frac{d}{dt} w(t, x(t)) = \left[ f(w) - (w - 1)f'(w) \right] F(x; w).
\] (19)
Combining (19) with (17), we get, for any \( T > 0 \),
\[
\alpha(w(T, x(T))) \leq \alpha(w(0, x(0)) + 2m_0.
\] (20)
From (3) we deduce that \( \alpha(w) \) diverges as \( w \to \infty \). Therefore, from (20), we achieve an upper bound for \( w \) uniformly in \( t \).

The discrete version of these estimates are proved for the front tracking approximate solutions. Furthermore, the bound on the total variation is derived. The existence of solutions is then achieved by a compactness argument.

The uniqueness of the solutions for (4) follows by the Oleinik type estimate (8), combined with an error formula for the \( L^1 \) distance of two different solutions, corresponding to the same initial data. For details of the proof, we refer to [2].

A somewhat similar front tracking algorithm is used in [9], where we construct piecewise polygonal (possibly discontinuous) approximate solutions for (1). The algorithm traces both the front \( x_i(t) \) and the value \( u_i(t) \), governed by two sets of coupled ODEs.

The stability estimate (11) is obtained by directly measuring the \( L^1 \) distance between the \( \epsilon \)-front tracking approximate solutions for \( w \) and \( v \), and pass to the limit as \( \epsilon \to 0^+ \), see [2]. By Gronwall’s Lemma, (11) gives continuous dependence.

We remark that the existence and uniqueness of BV solutions for (4) is also obtained in [3], where a fractional time step algorithm is used to generate approximate solutions.

Before moving on, we make an observation. Thanks to the bounds on \( w \), the inverse function \( x(t, u) \) is also conserved.

Indeed, equation (4) implies that the differential form \( w \, dx - \left( f(w) F(x; w) \right) \, dt \) is closed and then exact, and \((t, x) \mapsto u(t, x) \) could serve as a potential, with
\[
u_x = w, \quad u_t = -\left( f \left( w(t, x) \right) F(x; w) \right), \quad du = u_x \, dx + u_t \, dt.
\]
Since \( \frac{1}{w(t, x)} \in L^1_{loc} \), we can divide by \( w \) the differential form and get
\[
dx = (1/w) \, du + (f(w)/w) \, F(x; w) \, dt.
\] (21)
We switch to the coordinates \((t, u)\), and set \( z = x_u(t, u) = 1/w \). The integral changes to
\[
\int_x^{+\infty} f \left( U_z(t, y) \right) \, dy = \int_{U(t, x)}^{+\infty} g \left( z(t, v) \right) \, dv, \quad g(s) = s f(1/s) \quad (s > 0).
\] (22)
The cross-derivatives equation for (21) now gives a conservation law for \( z(t, u) \):
\[
\partial_t z - \partial_u \left( g(z) \exp \int_u^{+\infty} g \left( z(t, v) \right) \, dv \right) = 0.
\] (23)
Integrating (23), we obtain a conservation law for \( x(t, u) \):
\[
\partial_t x + \partial_u \left( \exp \int_u^{+\infty} g \left( z(t, v) \right) \, dv \right) = 0.
\] (24)
3. Case 2: Solutions with Discontinuous Profile. We now assume that \( f \) satisfies the assumption (2) and (5), and study equation (1). The presence of the measure \( U_z \) requires new treatment in the front tracking algorithm, as well as in the convergence analysis for the approximate solutions and their continuous dependence on data. To this end, we introduce the inverse function \( X = X(t, u) \) which is the graph completion of the inverse in space of \( U(t, x) \),

\[
X(t, u) = x \iff u \in \left[ U(t, x^-), U(t, x^+) \right].
\]

Note that if \( U(t, x) \) is right continuous, then \( u \leq U(t, X(t, u)) \) for all \( u \in \mathbb{R} \). Wherever \( U(t, x) \) has a jump in \( x \), the inverse function \( X(t, u) \) will remain constant over the interval of jump in \( U \). Furthermore, the lower bound on the gradient in (9) implies that the function \( X \) is Lipschitz continuous in \( u \). This allows us to define the function

\[
z(t, u) = X_u(t, u), \quad z \in L^\infty(\mathbb{R}; [0, +\infty[).
\]

In the presence of point masses in \( U_x \), the integral term in the flux is redefined as (22), see [9]. We are led to consider the following conservation law

\[
U_t(t, x) - \left( \exp \int_{U(t, x)}^{+\infty} g \left( z(t, v) \right) \, dv \right) = 0
\]

where \( z \) is treated as the unknown function. The assumptions (2) and (5) for \( f \) give the corresponding assumptions on \( g \):

\[
g \in C^2(0, +\infty; \mathbb{R}) \cap C^1(0, +\infty[; \mathbb{R}) \ , \ g(1) = 0 , \ \sup g'' < 0 , \ g(0) > 0 .
\]

Corresponding to (9), we seek \( \text{BV} \) solutions to the Cauchy problem for (26) in the set \( \mathcal{Z} \)

\[
\mathcal{Z} = \left\{ z \in \text{BV} \left( \mathbb{R}; [0, +\infty[ \right) : \ z \text{ is right-continuous, and } (z - 1) \in L^1 \left( \mathbb{R}; [0, +\infty[ \right) \right\}.
\]

The function \( U \) could be reconstructed from \( z \) as follows:

\[
X(t, u) = u + \int_{-\infty}^{u} (z(t, v) - 1) \, dv, \quad U(t, x) = \max \left\{ v \in \mathbb{R} : X(t, v) \leq x \right\}.
\]

In this setting, a Lipschitz semigroup of solutions can be constructed.

**Theorem 3.1.** [6, Theorem 2.1] Fix \( T > 0 \) and let \( \mathcal{Z} \) be as in (28). For any \( g \) satisfying (27), there exists a map \( S^g : [0, T] \times \mathcal{Z} \rightarrow \mathcal{Z} \) with the following properties:

1. \( S^g_0 = \text{Id} \) and for any \( t_1, t_2 \in [0, T] \) with \( t_1 + t_2 \in [0, T] \), the semigroup property holds: \( S^g_{t_1} \circ S^g_{t_2} = S^g_{t_1 + t_2} \).

2. For any \( z_0 \in \mathcal{Z} \), the orbit \( t \rightarrow S^g_t z_0 \) solves (26)–(29) in the sense of distributions.

3. Let \( z, \tilde{z} \in \mathcal{Z} \). There exists a constant \( L > 0 \), depending on \( \text{TV} \, z, \, \text{TV} \, \tilde{z} \) and \( T \), such that for any \( g, \tilde{g} \) satisfying (27), and for any \( t, \tilde{t} \in [0, T] \) with \( \tilde{t} \leq t \), one has

\[
\left\| S_{\tilde{t}}^g z - S_{t}^g \tilde{z} \right\|_{L^1} \leq L \left( t \| g - \tilde{g} \|_{W^{1, \infty}} + e^{Lt} \| z - \tilde{z} \|_{L^1} + | t - \tilde{t} | \right).
\]

Notice that (30) is the counterpart of (11) in Theorem 2.2.

Below we give some formal arguments. If the gradient \( U_x \) can blow up, the equation (23) should be equipped with a pointwise constraint \( z \geq 0 \). We have

\[
z_t(t, u) - \left( g(z) \exp \int_{u}^{+\infty} g(z(t, v)) \, dv \right)_u = \mu.
\]

\[\text{(31)}\]
Here the function \( g \) could be viewed as the erosion function in the new coordinates \((t, u)\), denoting the rate of erosion per unit distance travelled in \( u \) per unit mass passing through. The measure \( \mu \) on the right hand side of (31) yields the projection into the cone of non-negative functions. It satisfies the following property: for every \( t > 0 \) and \( a, b \in \mathbb{R} \) such that \( z(t, a), z(t, b) > 0 \), we have

\[
\mu([a, b]) = 0, \quad \int_a^b \mu([a, u]) \, du = 0. \tag{32}
\]

Note these two conditions are precisely the conservation laws for \( z \) in (23) and \( X \) in (24) over a jump, respectively.

To understand the effect of this projection on the \( L^1 \) distance between two solutions, consider \( z_1 \) and \( z_2 \) as in Figure 1, where \( z_1 < 0 \) on an interval and \( z_2 > 0 \). The property (32) implies the relation \( A_0 = A_1 + A_2 \) for the areas. Let \( z_1^+ \) be the projection of \( z_1 \). Then

\[
\|z_1^+ - z_2\|_{L^1} - \|z_1 - z_2\|_{L^1} \leq -A_0 + A_1 + A_2 \leq 0, \tag{33}
\]

formally proving that the projection does not increase the \( L^1 \) distance between two solutions.

![Figure 1. Effect of the measure \( \mu \) on the \( L^1 \) distance between two solutions.](image)

The approximate solutions are generated by a front tracking algorithm, similar to Case 1, but also adapted to the discontinuities in \( U(t, x) \). For \( z > 0 \), downward jumps (convex kinks in \( U \)) are admissible, and they travel with Rankine-Hugoniot speeds. Jumps in \( U(t, x) \) correspond to an interval \((u^-(t), u^+(t))\) where \( z(t) = 0 \). The propagation speeds of the fronts \( u^-, u^+ \) are determined by the two assumptions (32) on the measure \( \mu \), see (3.4) in [6] and the detailed derivation. It is interesting to observe that upward jumps in \( z \) could occur in the solutions, but only at the right front of a shock, i.e., at \( u^+(t) \).

By applying Lax entropy condition on the characteristic speeds around a shock, we obtain a restriction on the value \( z^+ \) for a given jump size \( u^+ - u^- \), i.e.

\[
g(z^+) - z^+ g'(z^+) \leq \exp \left( g(0) \frac{(u^+ - u^-)}{u^+ - u^-} \right) - 1. \tag{34}
\]

An equivalent condition on \( w^+ \) for jumps in \( U(t, x) \) is derived in [9].

Due to the presence of various types of singularities (shocks, hyper-kinks and kinks), wave interaction patterns for (1) are very complicated. Different types of waves could form and interact, and also get cancelled as they interact. In Figure 2 we show the numerical simulation for a case study, where one observes the formation
of a discontinuity in $U(t,x)$ (and $z(t,u) = 0$) from continuous initial data, and its cancellation by a rarefaction wave entering from the right.

We conclude with a final remark. Due to the presence of the jumps in $U(t,x)$ in Case 2, characteristic curves could exit from the right front of a shock in $U$, tangent to the shock curve, therefore an Oleinik type entropy condition for (1) such as (8) is not valid. The uniqueness of solutions for Case 2 is still open.

![Figure 2](image-url)

**Figure 2.** A cartoon sequence of plots of the solutions $U(t,x)$ (the upper 8 plots) and the corresponding plots of $z(t,u)$ (the lower 8 plots) as $t$ evolves. One observes the formation and the cancellation of a shock in $U$, and the corresponding behavior of $z(t,u)$.

**REFERENCES**


E-mail address: amadori@univaq.it
E-mail address: rinaldo@ing.unibs.it
E-mail address: graziano.guerra@unimib.it
E-mail address: shen_w@math.psu.edu
NUMERICAL SOLUTIONS TO THE RIEMANN PROBLEM FOR COMPRESSIBLE ISOTHERMAL EULER EQUATIONS FOR TWO PHASE FLOWS WITH AND WITHOUT PHASE TRANSITION

MAREN HANTKE
Institute for Analysis and Numerics
Otto-von-Guericke-University Magdeburg
PSF 4120, D-39106 Magdeburg, Germany
and
Institute for Applied Mathematics
Gottfried Wilhelm Leibniz University Hannover
PF 6009, D-30060 Hannover, Germany

Ferdinand Thein
Institute for Analysis and Numerics
Otto-von-Guericke-University Magdeburg
PSF 4120, D-39106 Magdeburg, Germany

ABSTRACT. We consider the isothermal Euler equations with and without phase transition between a liquid and a vapor phase. The mass transfer across the phase boundary is modeled by a kinetic relation. Existence and uniqueness results were proven in [2]. We present a method to obtain the numerical solution for associated Riemann problems. The calculated results will be compared to the exact solutions. Therefore we will highlight the major difficulties and propose possible strategies to overcome these problems.

1. Introduction. In this work we want to present a numerical procedure for solving Riemann problems for two phase flows with and without phase transition. Therefore we rely on the compressible isothermal Euler equations. We admit mass transfer across a sharp interface which is controlled by a kinetic relation. Furthermore we use a non-monotone pressure-density function. The complete model was presented and analytically solved by Hantke et al., see [2]. In [2] a detailed motivation for this problem with further related literature is given. The main point there is the use of the sharp interface and the kinetic relation which does not assume local equilibrium at the interface. Hence the work of Hantke et al. [2] can clearly be distinguished from other research on this field. For example Müller and Voss [4] used the van der Waals equation to model the fluid. Instead of a kinetic relation they applied the Liu entropy condition for obtaining uniqueness. For more information on this topic see [2]. Due to this new approach there is almost no literature on this topic and we want to present a method for solving this problem numerically. The outline of this work is as follows, in Section 2 to 4 we present the framework

2000 Mathematics Subject Classification. Primary: 35Q05, 76T10, 82B26.
Key words and phrases. Euler equations, two phase flow, phase transition, sharp interface model, numerical Riemann Solver.

The second author was supported by Landesgraduiertenstipendium Sachsen-Anhalt and is supported by the DFG grant HA 6471/2-1. The authors thankfully acknowledge the support.
containing the Euler equations, the equations of state and the exact solution presented in [2]. In Section 5 we discuss major challenges and present possible solution strategies. Numerical results are presented in Section 6.

2. Isothermal Euler Equations. We consider inviscid fluids under constant temperature \(T_0\), i.e. isothermal fluids. The phases are distinguished by the value of mass density \(\rho\) and furthermore we have the velocity \(v\) as a variable. The physical quantities depend on time \(t \in \mathbb{R}_{\geq 0}\) and space \(x \in \mathbb{R}\). In regular points of the phases we have the local mass conservation law and the balance law for momentum, i.e.

\[
\frac{\partial \rho}{\partial t} + \frac{\partial (\rho v)}{\partial x} = 0, \tag{1}
\]

\[
\frac{\partial (\rho v)}{\partial t} + \frac{\partial (\rho v^2 + p)}{\partial x} = 0. \tag{2}
\]

The additional quantity \(p\) is the pressure of the fluid and since it is not among the basic variables it is called constitutive quantity. The pressure depends on the mass density and the fluid under consideration. The corresponding equation of state (EOS) will be given in Section 3.

Across any of the appearing discontinuities we have the following jump conditions

\[
\rho(v - W) = 0, \tag{3}
\]

\[
\rho(v - W)[v] + [p] = 0. \tag{4}
\]

The quantity \(W\) denotes the propagation speed of the discontinuity and the mass flux \(Z\) across a discontinuity is given by

\[
Z = -\rho(v - W) \tag{5}
\]

with

\[
Z = \begin{cases} Q \text{ shock wave} & W = \begin{cases} S \text{ shock wave} \\ z \text{ phase boundary} & w \text{ phase boundary} \end{cases} \end{cases}
\]

Here we use for any physical quantity \(\Psi\) the bracket notation \([\Psi] = \Psi_{\text{right}} - \Psi_{\text{left}}\).

3. Equations of state. The equation of state links the pressure \(p\) to the mass density \(\rho\). The fluid considered here is water. Since we are dealing with two different phases, we have two different equations. For the vapor phase we use the ideal gas law

\[
p_v(\rho_v) = \rho_v \frac{kT_0}{m}. \tag{6}
\]

Here \(T_0\) is the given temperature, \(k\) the Boltzmann constant and \(m\) the mass of a single water molecule. The liquid phase is assumed to be compressible and the equation of state is

\[
p_l(\rho_l) = p_0 + K_0 \left( \frac{\rho_l}{\rho_0} - 1 \right). \tag{7}
\]

The quantities \(p_0, \rho_0\) and the modulus of compression \(K_0\) are arbitrary reference values. Here these quantities are chosen at the saturation state and hence depend on the temperature \(T_0\), see [6]. As done in [2] we characterize the two phases via the mass density. If \(0 \leq \rho \leq \tilde{\rho}\) we assume a vapor state and if \(\rho \geq \rho_m\) we assume a liquid state. The bounds \(\tilde{\rho}\) and \(\rho_m\) are constant parameters with \(\tilde{\rho} < \rho_m\). These constants are chosen as in [2].
4. Exact solution of the Riemann problem. In this work we consider the Riemann problem for the isothermal Euler equations given by (1), (2) and the equations of state (6) and (7). The initial Riemann data is given by

\[
\rho(x, 0) = \begin{cases} 
\rho_v, & x < 0 \\
\rho_l, & x > 0
\end{cases} \quad \text{and} \quad v(x, 0) = \begin{cases} 
v_v, & x < 0 \\
v_l, & x > 0
\end{cases}
\]

(8)

Here we assumed without any loss of generality that the left state is the vapor phase and the right state is the liquid phase respectively. The solution is denoted by \(\mathbf{W}\).

It was shown in [2] that the exact solution consists of four constant states which are separated by three waves. It was also shown that the second wave is always the phase boundary. To describe the phase transitions Hantke et al. used a kinetic relation, see [2]. This kinetic relation reads in our case

\[
z = \frac{p_v}{\sqrt{2\pi}} \left( \frac{m}{kT_0} \right)^\frac{3}{2} \left[ \frac{K_0}{\rho_0} \ln \frac{\rho_l}{\rho_0} - \frac{kT_0}{m} \ln \frac{p_v}{p_0} + \frac{1}{2} (v_l - w)^2 - \frac{1}{2} (v_v - w)^2 \right]
\]

(9)

and it describes the rate of change of mass across the interface. The phase transition is driven by the difference of the Gibbs free energies of the phases and therefore the kinetic relation is derived from thermodynamical considerations to describe this process. For further details see Dreyer et al. [1]. If the phases switch their sides, the kinetic relation just changes it sign. Together with the interface momentum balance (4) and the equations of state (6) and (7) the kinetic relation links the pressures in the star states

\[
[p] = -z^2 \left[ \frac{1}{\rho} \right].
\]

(10)

As stated in Lemma 5.4 in [2] the liquid pressure is uniquely defined for a given interface pressure \(p_v\) by (10). We denote the four constant states as follows

\[
\mathbf{W}_v = \left[ \begin{array}{c} \rho_v \\
v_v \end{array} \right], \quad \mathbf{W}_v^* = \left[ \begin{array}{c} \rho_v^* \\
v_v^* \end{array} \right], \quad \mathbf{W}_l = \left[ \begin{array}{c} \rho_l \\
v_l \end{array} \right].
\]

The pressures are \(p_v, p_v^*, p_l^*\) and \(p_l\) respectively. The exact solution is given by Theorem 6.5 in [2] for a temperature between 273.15\(K\) and 623.15\(K\).

**Theorem 4.1.** Let \(f_z(p_v^*, \mathbf{W}_v, \mathbf{W}_l)\) be given as

\[
f_z(p_v^*, \mathbf{W}_v, \mathbf{W}_l) = f_v(p_v^*, \mathbf{W}_v) + f_l(p_l^*(p_v^*), \mathbf{W}_l) + z \left[ \frac{1}{\rho} \right] + v_l - v_v
\]

(11)

where the functions \(f_v\) and \(f_l\) are given by

\[
f_v(p_v^*, \mathbf{W}_v) = \begin{cases} 
\frac{p_v^* - p_v}{\sqrt{p_v^* p_l}}, & p_v^* > p_v \text{ (shock)} \\
-a_v \ln p_v + a_v \ln p_v^*, & p_v^* \leq p_v \text{ (rarefraction)}
\end{cases}
\]

\[
f_l(p_l^*(p_v^*), \mathbf{W}_l) = \begin{cases} 
\frac{p_l^*(p_v^*) - p_l}{\sqrt{p_l^* p_v^*}}, & p_l^*(p_v^*) > p_l \text{ (shock)} \\
-a_l \ln p_v + a_l \ln \left( \frac{p_l^*(p_v^*) - p_v}{K_0} + 1 \right), & p_l^*(p_v^*) \leq p_l \text{ (rarefraction)}
\end{cases}
\]

The function \(p_l^*(p_v^*)\) is implicitly defined by (10) and \(z\) is given by (9). If the function \(f_z(p_v^*, \mathbf{W}_v, \mathbf{W}_l)\) has a root \(p^*\) with \(0 \leq p^* \leq \tilde{p} = \frac{\tilde{\rho} k T_0}{m}\) this root is unique and the exact solution for the vapor pressure in the star region.
For sufficient conditions, further details and the proof see [2]. In the case without phase transition we have $z = 0$ and hence we get from (10) that $p^*_v = p^*_l$. By $a_v$ and $a_l$ we denote the corresponding speeds of sound. If $p^*_v$ is determined the velocities can be calculated using (compare [2])

$$

v^*_v = v_v - f_v(p^*_v, W_v), \quad v^*_l = v_l + f_l(p^*_v, W_l).

$$

(12)

5. **Numerical solution of the Riemann problem.** Here we are only interested in the presence of a phase boundary. If two adjacent cells belong to the same phase one can use any standard Riemann solver. When this problem has to be solved numerically, some difficulties occur which we will highlight in the next section.

5.1. **Main challenges.** Beside the well known numerical problems of hyperbolic equations (e.g. [5]), we encounter further difficulties at a phase boundary. A common feature of finite-volume methods is the averaging inside the cells. This becomes more crucial, now that we have two phases. The phase boundary will lie inside a cell if $w \neq 0$. Now averaging can lead to unphysical states and hence to a wrong solution. For example you can obtain averaged densities which can not be identified as one of the phases, i.e. $\hat{\rho} < \rho < \rho_m$.

The next problem is the calculation of the quantities inside the star region. The quantities we need to determine are the values of the vapor phase $\rho^*_v, v^*_v, p^*_v(\rho^*_v)$, the values of the liquid phase $\rho^*_l, v^*_l, p^*_l(\rho^*_l)$ and the speed of the phase boundary $w$. The densities can be calculated using (6) and (7), if the pressures are known. The velocities can be calculated using (12). Furthermore we know that $p^*_l$ is implicitly defined by $p^*_v$ via (10). Finally $w$ can be calculated from the star state values. So it all comes down to $p^*_v$. If these quantity is known we can obtain the remaining ones.

Finally we need to determine the correct flux (since we have two different phases) and time step. Especially the time step is crucial, since we may have different cell sizes due to the motion of the phase boundary.

5.2. **Grid adaption.** When the phase boundary lies inside a cell, unphysical phase states can occur when values are averaged. Therefore we need the phase boundary to coincide with a cell boundary. For this purpose one can use different methods. We shift the cell boundary from the previous time step to the new position of the boundary. After that, we check and if necessary adjust the grid to make sure that the CFL condition is still satisfied. For this reason we have to calculate new cell values but now only from cells of the same phase. The drawback is, that $dx$ is not constant anymore. So we have to be careful whenever the size of the cells is used, as for example for the time step.

5.3. **Estimates for the interfacial vapor pressure $p^*_v$.** The next difficulty is to get a good estimate for the pressure $p^*_v$ in the vapor star state and therefore also for the speed of the phase boundary $w$. As mentioned above this is the quantity we need to determine all remaining values. We need this estimate for an iteration later on that is used to find sufficiently good values. At first we calculate an estimate for the speed of the phase boundary via (4) for the left (vapor) and the right (liquid) state, i.e.

$$

w = \frac{p_l - p_v + \rho_v v_v(S_L - v_v) - \rho_l v_l(S_R - v_l)}{\rho_v v_v(S_L - v_v) - \rho_l v_l(S_R - v_l)}.

$$

See for instance [5]. Now we can calculate an initial estimate for $p^*_v$ which we will denote by $p^*_v$. For this estimate we assume a discontinuity between the left vapor
and left vapor star state. Furthermore we assume \( v_v^* = w \) and \( S_L = v_v - a_v \). If we also use \( p_{\text{est}} = \frac{p_v}{\sigma^3} \) we can apply (4) and solve it for \( p_{\text{est}} \) and obtain

\[
p_{\text{est}} = \frac{p_v}{\sigma^3} (w - S_L)(w - v_v) + 1.
\]

In the next step we perform an ANRS routine for two phases analog to the one presented in [5] or in [3] to improve the estimate for the vapor pressure. The role of the PVRS approximation of the pressure is now played by \( p_{\text{est}} \).

### 5.4. Simplified calculation of the corresponding interfacial liquid pressure

\( p_L^* \). An important point throughout the computation is to determine the liquid pressure corresponding to a vapor pressure. We need this routine very often, but this pressure is only implicitly defined by (10) and hence an iteration procedure is needed. To avoid an iteration we approximate (10) by a polynomial in \( p_l \). The function which describes the liquid pressure implicitly can be written as

\[
f(p_v, p_l) = \left[ p_l + z^2 \left( \frac{1}{p_l} \right) \right] = \left[ p_l + \frac{p_v^2}{2\pi} \left( \frac{m}{kT_0} \right)^3 \left[ K_0 \ln \frac{p_l}{p_0} - \frac{kT_0}{m} \ln \frac{p_v}{p_0} - \frac{1}{2} \left[ p_l - \frac{1}{2} \left( \frac{1}{p_l} + \frac{1}{p_v} \right) \right] \right] \right]^2 - \left[ \frac{1}{p_l} \right].
\]

We approximate this function by

\[
\hat{f}(p_v, p_l) = f(p_v, p_0) + \partial_{p_l} f(p_v, p_0)(p_l - p_0) + \frac{1}{2} \partial_{p_l}^2 f(p_v, p_0)(p_l - p_0)^2.
\]

Now we can easily find \( p_l(p_v) \) as a root of a second order polynomial. It is important to note that basically three estimates are needed to verify the correctness and quality of the solution. At first we need to estimate the remainder of the Taylor expansion

\[
\frac{1}{6} \partial_{p_l}^3 f(p_v, \hat{p})(p_l - p_0)^3
\]

from below and above to know the approximation error (or simply find an upper bound for the absolute value). The second estimate is for \( \partial_{p_l}^2 f(p_v, p_l) \) since this function must not be zero. Finally we need the discriminant to be positive, i.e.

\[
d(p_v) := (\partial_{p_l} f(p_v, p_0))^2 - 2f(p_v, p_0)\partial_{p_l}^2 f(p_v, p_0) \geq 0.
\]

We will not present these calculations in detail but give a brief idea of the way we used. We have for the pressure (for both phases) and temperature

\[(p, T) \in [0, 2p_0] \times [273.15 K, 623.15 K] .
\]

The temperature interval is due to Theorem 4.1, see [2]. For every temperature given above we can determine the maximum pressure and this is always smaller than \( 2p_0 \), so this is no restriction (compare [2]). Next we calculate the derivatives of \( f \) that we have to deal with. As in [2] we write

\[
f(p_v, p_l) = \left[ p_l + \frac{p_v^2 h(p_v, p_l)^2}{2 \pi} \left( \frac{1}{p_l} \right) \right],
\]

\[
h(p_v, p_l) = \frac{1}{\sqrt{2\pi}} \left( \frac{m}{kT_0} \right)^2 \left[ K_0 \ln \frac{p_l}{p_0} - \frac{kT_0}{m} \ln \frac{p_v}{p_0} - \frac{1}{2} \left[ p_l - \frac{1}{2} \left( \frac{1}{p_l} + \frac{1}{p_v} \right) \right] \right].
\]

Hence we have to determine lower and upper bounds for \( h(p_v, p_l) \) and its derivatives with respect to \( p_l \). This has to be done with great care. For example, we do not directly obtain bounds for \( \partial_{p_l} h \) but we get bounds for \( p_v \partial_{p_l} h \). So when we finally
calculate bounds for the terms related to \( f(p_v, p_l) \) noted above, we have to keep that in mind. In the end we have bounds depending on the (monotonically temperature dependant) constants \( \rho_0, K_0 \) and \( p_0 \). That means that the bounds differ in their absolute value for different temperatures (for examples see Table 1 below) but the qualitative behavior remains unchanged. As seen exemplarily in Table 1 we could show with the calculated bounds that \( \partial_{p_l}^2 f(p_v, p_l) \) is strictly negative. The fact that it is negative helps to find the correct root. To verify which root to choose we use the condition \( p_l(p_l) = p_0 \) which is pointed out in [2] in the proof of Theorem 4.1. Together with \( f(p_v, p_0) = 0, \partial_{p_v} f(p_0, p_0) = 1 \) and \( \partial_{p_l}^2 f(p_v, p_l) < 0 \) we determine the correct solution

\[
p_l(p_v) = \frac{\partial_{p_v} f(p_v, p_0) - \partial_{p_l}^2 f(p_v, p_0) p_0}{\partial_{p_l} f(p_v, p_0)} - \frac{(\partial_{p_l} f(p_v, p_0))^2 - 2f(p_v, p_0)\partial_{p_v}^2 f(p_v, p_0)}{(\partial_{p_l} f(p_v, p_0))^3}.
\]

(14)

Now we are able to calculate the liquid pressure without any iteration. When studying convergence one has to keep in mind that we basically replaced the correct kinetic relation by an (very good) approximation and hence we solve a slightly different Riemann problem.

5.5. Iteration procedure. We now want to improve our values for \( p_v^* \) and \( p_l^* \) since they are crucial for computing the remaining values. Therefore we perform a Newton iteration using the pressure function \( f_z \) defined in Theorem 4.1. Omitting this iteration will often lead to oscillations, smearing and wrong values, not only at the phase boundary (see Figure 2). We start with the estimated vapor pressure and calculate the corresponding liquid pressure. With these two we can calculate \( h, \partial_{p_v} h \) and \( \partial_{p_l} h \) from (13). Furthermore we need the derivative \( \frac{\partial w}{\partial p_v} \). This derivative cannot be calculated exactly, hence we use standard numerical differentiation. This is not very expensive since we can calculate any liquid pressure directly using (14). Finally we iterate until a desired accuracy is achieved. In our examples it usually does not take more than five iterations with \( \epsilon = 10^{-6} \). Usually this iteration converges very fast which leads to errors even smaller than epsilon. Hence a qualitative good and fast calculation is obtained. We further get as an extra output from the previous calculation of \( p_v^* \) and \( p_l^* \) the quantities \( z, v_v^* \) and \( v_l^* \). Hence we can also update \( w \) due to (5).

5.6. Calculation of the fluxes. Now it remains to determine the correct flux and then we can calculate the next time step. We cannot use for example the HLLC flux or even the Roe flux since these fluxes do not work for our problem with two phases with mass transfer. To determine the flux we take a look at the jump conditions (3) and (4). They basically determine the flux up to the sign. So finally the flux is

\[
F = \begin{bmatrix} -z \\ -z v^* + p^* \end{bmatrix}.
\]
Now one has the freedom to choose whether to take the vapor or the liquid values.

5.7. **Time Step.** For the time step we have to keep in mind, that the cell sizes are not necessarily equal. Hence we have for all cells

$$dx_n^L U_n^L + dx_R^L U_R^L = dx_{n+1}^L U_{n+1}^L + dx_R^{n+1} U_{n+1}^R + dt(F_R - F_L).$$

Therefore we obtain for the time step

$$U_n^{n+1} = \frac{dx^n}{dx^{n+1}} U_n^n - \frac{dt}{dx^{n+1}}(F_R - F_L).$$

One can clearly see that the standard method is obtained for a constant cell size \( dx \).

6. **Numerical results.** Here we present one example and compare it to the exact solution given in [2] (see Example 1). The initial data is given in Table 2. The computation was performed with the following parameters

\( dt = 5.7019 \cdot 10^{-7}, \quad dx_0 = 0.001, \quad C_{CFL} = 0.9, \quad t_{end} = 0.001s \quad \text{and} \quad x \in [-2, 2]. \)

The solution is pictured in Figure 1 and the numerical values at the phase boundary are compared with the exact values in Table 3. In Figure 2 we present a zoom plot of

<table>
<thead>
<tr>
<th>( p_0 )</th>
<th>( v_0 )</th>
<th>( v_l )</th>
<th>( p_l )</th>
<th>( T_0 )</th>
<th>( p_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2300 Pa</td>
<td>-100 m/s</td>
<td>1000 Pa</td>
<td>100 m/s</td>
<td>293.15 K</td>
<td>2339 Pa</td>
</tr>
</tbody>
</table>

Table 2. Initial Data

Table 3. Results at the phase boundary

<table>
<thead>
<tr>
<th></th>
<th>( p_v^* )</th>
<th>( v_v^* )</th>
<th>( p_l^* )</th>
<th>( v_l^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>numerical</td>
<td>1560.8888 Pa</td>
<td>42.4483 m/s</td>
<td>1599.4904 Pa</td>
<td>100.0004 m/s</td>
</tr>
<tr>
<td>exact</td>
<td>1561.3074 Pa</td>
<td>42.4852 m/s</td>
<td>1599.4847 Pa</td>
<td>100.0004 m/s</td>
</tr>
<tr>
<td>relative error</td>
<td>0.00026809</td>
<td>0.00086828</td>
<td>3.5374 \cdot 10^{-6}</td>
<td>5.6130 \cdot 10^{-9}</td>
</tr>
</tbody>
</table>

Figure 1. Numerical (blue) and exact (magenta) solution
the pressure for the same initial data and parameters. But now no Newton iteration was performed and the solution starts to oscillate. There are more examples with stronger effects but the presentation and discussion would be beyond the scope of this work.

![Oscillating solution in the case without iteration](image)

**Figure 2.** Oscillating solution in the case without iteration

7. **Summary.** In this work we presented and briefly discussed a numerical method to solve Riemann problems for compressible isothermal Euler equations for two phase flows with and without phase transition. The analysis for the exact solution was done by Hantke et al. in [2]. We highlighted major difficulties and proposed solutions to overcome these difficulties. Future work will include convergence and stability studies. Furthermore an extension to mixtures, nucleation and cavitation will be of interest. Another crucial point will be the extension to higher dimensions and non isothermal problems.

**REFERENCES**


*E-mail address: maren.hantke@ovgu.de, hantke@ifam.uni-hannover.de*

*E-mail address: ferdinand.thein@st.ovgu.de*
ASYMPTOTIC BEHAVIOR OF SOLUTIONS FOR DAMPED WAVE EQUATIONS WITH NON-CONVEX CONVECTION TERM ON THE HALF LINE

Itsuko Hashimoto
Toyama national college of technology
Toyama 939-8630, Japan

Abstract. In this article, we report the asymptotic stability of nonlinear waves for damped wave equations with a non-convex convection term on the half line. We show the asymptotic stability of stationary wave by using weighted energy method. Furthermore, we present the result of the time convergence rate which is polynomially fast if the initial perturbation decays polynomially as $x$ goes up to infinity. This article is based on a joint work with Professor Yoshihiro Ueda of Kobe University.

1. Introduction. This article is based on a joint work with Professor Yoshihiro Ueda of Kobe University. We discuss the initial-boundary value problem on the half line for damped wave equations with a nonlinear convection term:
\begin{align}
\begin{cases}
  u_{tt} - u_{xx} + u_t + f(u)_x = 0, & x > 0, \ t > 0, \\
  u(0, t) = u_-, & t > 0, \\
  \lim_{x \to \infty} u(x, t) = 0, & t > 0, \\
  u(x, 0) = u_0(x), \ u_t(x, 0) = u_1(x), & x > 0,
\end{cases}
\end{align}

where the function $f = f(u)$ is a given smooth function satisfying $f(0) = 0$ and $u_-$ is a given constant with $u_- < 0$.

Throughout this article, we impose that the convection term satisfies the following condition:
\begin{align}
|f'(0)| < 1, \ f(u) > f(0) (= 0) \ \text{for} \ u \in [u_-, 0].
\end{align}

The first condition in (2) is so-called sub-characteristic condition which also be delt by Ueda-Kawashima [12] for the research of the solution of (1) with convex convection term. In [10], it is considered the case where $f(u)$ of (1) satisfies
\begin{align}
f''(u) > 0, \ |f'(u)| < 1 \ \text{for} \ u \in [u_-, 0],
\end{align}
which is the case that the solution of corresponding viscous conservation laws tends toward a stationary solution $\phi = \phi(x)$ which is defined by the solution of the
stationary problem corresponding to (1):
\[
\begin{cases}
  f(\phi) = \phi_x, & x > 0, \\
  \phi(0) = u_-, \quad \lim_{x \to \infty} \phi(x) = 0.
\end{cases}
\] (4)

As a direct expansion, Hashimoto-Ueda [2] considered the case that the convection term satisfies
\[f''(0) > 0, \quad |f'(0)| < 1, \quad f(u) > 0 \quad \text{for} \quad u \in [u_-, 0).\] (5)

In [2], we showed asymptotic stability of stationary wave by using weighted energy method:

**Theorem 1.1** ([2]). Suppose that \(u_- < 0 < u_+ = 0\) and (5) hold. Let \(\phi(x)\) be the stationary solution satisfying the problem (4). Assume that \(u_0 - \phi \in H^1\) and \(u_1 \in L^2\). Then there exists a positive constant \(\varepsilon_0\) such that if \(\|u_0 - \phi\|_{H^1} + \|u_1\|_{L^2} \leq \varepsilon_0\), then the initial-boundary value problem (1) has a unique global solution in time \(u\) satisfying
\[u - \phi \in C^0([0, \infty); H^1), \quad u_x, u_t \in L^2(0, \infty; L^2),\]
and the asymptotic behavior
\[
\lim_{t \to \infty} \sup_{x > 0} |u(x, t) - \phi(x)| = 0.
\]

In this paper, we showed the asymptotic stability for the stationary wave for the damped wave equations with convection term which is not necessarily convex at the origin. To investigate the main theorem of this article, we assume that \(u_0 - \phi\) and \(u_1\) are integrable:
\[
z_0(x) = -\int_{x}^{\infty} (u_0(y) - \phi(y)) \, dy \in L^2, \quad z_1(x) = -\int_{x}^{\infty} u_1(y) \, dy \in L^2.
\]

By using these functions, main theorem is stated as follows:

**Theorem 1.2.** ([3]) Suppose that \(f(u)\) satisfies (2) and \(\phi(x)\) be the stationary solution satisfying the problem (4). Then it holds the following results.

Assume that \(z_0 \in H^2\) and \(z_1 \in H^1\). Then there is a positive constant \(\varepsilon_0\) such that if \(\|z_0\|_{H^2} + \|z_1\|_{H^1} \leq \varepsilon_0\), then the initial-boundary value problem (1) has a unique global solution \(u(x, t)\) satisfying \(u - \phi \in C^0([0, \infty); H^1) \cap C^1([0, \infty); L^2)\) and the asymptotic behavior:
\[
\lim_{t \to \infty} \sup_{x > 0} |u(x, t) - \phi(x)| = 0. \quad (6)
\]

Convergence rates of the solution \(u\) toward the stationary wave \(\phi\) is stated as follows.

**Theorem 1.3.** ([3]) Suppose that the same assumptions as in Theorem 1.2 hold true. Assume that \(z_0 \in H^2_{\alpha}\) and \(z_1 \in H^1_{\alpha}\) for \(\alpha > 0\). Let \(u(x, t)\) be the global solution to the problem (1), which is constructed in Theorem 1.2. Then it holds that
\[
\|u(t) - \phi\|_{H^1} \leq CE_{\alpha}(1 + t)^{-\alpha/2},
\]
for \(t \geq 0\), where \(C\) is a positive constant and \(E_{\alpha} = \|z_0\|_{H^2_{\alpha}} + \|z_1\|_{H^1_{\alpha}}\).
Theorem 1.4. ([3]) Suppose that the same assumptions as in Theorem 1.2 hold true. Assume that \( z_0 \in H^2_{\alpha,exp} \) and \( z_1 \in H^1_{\alpha,exp} \) for \( \alpha > 0 \). Let \( u(x,t) \) be the global solution to the problem (1), which is constructed in Theorem 1.2. Then it holds that

\[
\|u(t) - \phi\|_{H^s} \leq C E_{\alpha,exp} e^{-\beta t},
\]

for \( t \geq 0 \), where \( \beta \) is a positive constant depending on \( \alpha \), \( C \) is a positive constant and \( E_{\alpha,exp} = \|z_0\|_{H^2_{\alpha,exp}} + \|z_1\|_{H^1_{\alpha,exp}} \).

This article is the survey of our recent paper [3].

Notation. \( L^2 = L^2(\mathbb{R}_+) \) and \( H^s = H^s(\mathbb{R}_+) \) denote the usual Lebesgue space of square integrable functions and \( s \)-th order Sobolev space on the half line \((0,\infty)\) with norms \( \|\cdot\|_{L^2} \) and \( \|\cdot\|_{H^s} \), respectively.

For \( \alpha > 0 \), \( L^2_{\alpha} = L^2_{\alpha}(\mathbb{R}_+) \) denotes the polynomially weighted \( L^2 \) space with the norm

\[
\|u\|_{L^2_{\alpha}} = \left( \int_0^\infty (1+x)^\alpha |u(x)|^2 \, dx \right)^{1/2},
\]

while \( L^2_{\alpha,exp} = L^2_{\alpha,exp}(\mathbb{R}_+) \) denotes the exponentially weighted \( L^2 \) space with the norm

\[
\|u\|_{L^2_{\alpha,exp}} = \left( \int_0^\infty e^{\alpha x} |u(x)|^2 \, dx \right)^{1/2}.
\]

Similarly, we define the corresponding weighted Sobolev spaces \( H^s_{\alpha} = H^s_{\alpha}(\mathbb{R}_+) \) and \( H^s_{\alpha,exp} = H^s_{\alpha,exp}(\mathbb{R}_+) \) with a positive integer \( s \). Especially, we note that \( H^0_{\alpha} = H^0 \).

For an interval \( I \) and a Banach space \( X \), \( C^s(I;X) \) denotes the space of \( s \)-times continuously differentiable functions on the interval \( I \) with values in \( X \). Finally, letters \( C \) and \( c \) in this paper are defined as positive generic constants unless they need to be distinguished.

2. Global existence of solution. In order to derive the existence of the global solution in time, which is described in Theorem 1.2, we need to rewrite our original problem (1). Let \( \phi(x) \) be the stationary solution satisfying (4) and introduce a new unknown function \( v(x,t) \) and its integral \( z(x,t) \) by

\[
v(x,t) = u(x,t) - \phi(x), \quad z(x,t) = \int_x^\infty v(y,t) \, dy.
\]

Here, we assume the integrability of \( z(x,t) \) over \( \mathbb{R}_+ \). This transformation is motivated by the argument in Liu-Nishihara [6]. By using (1) and (4), we reformulate (1) in terms of \( v(x,t) \) as

\[
\begin{cases}
  v_{tt} - v_{xx} + v_t + \{f(\phi + v) - f(\phi)\}_x = 0, & x > 0, \quad t > 0, \\
  v(0, t) = 0, & t > 0, \\
  v(x, 0) = v_0(x), & v_t(x, 0) = v_1(x), & x > 0,
\end{cases}
\]

where we put \( v_0(x) = u_0(x) - \phi(x) \) and \( v_1(x) = u_1(x) \). We also reformulate (7) in terms of \( z(x,t) \) as

\[
\begin{cases}
  z_{tt} - z_{xx} + z_t + \{f(\phi + z_x) - f(\phi)\} = 0, & x > 0, \quad t > 0, \\
  z_x(0, t) = 0, & t > 0, \\
  z(x, 0) = z_0(x), & z_t(x, 0) = z_1(x), & x > 0.
\end{cases}
\]
where we put \( z_0(x) = -\int_x^\infty (u_0(y) - \phi(y))dy \) and \( z_1(x) = -\int_x^\infty u_1(y)dy \). To derive the existence of the global solution in time, we need to construct the local existence theorem and a priori estimate. For this purpose, we define the solution space for an arbitrary \( T > 0, \ M > 0 \) and \( k = 0, 1 \) that

\[
X_{k,M}(T) := \{ z \in C^0((0,T]; H_k^2); \ z_t \in C^0((0,T]; H_k^1), \ N(T) \leq M \},
\]

where we define that \( N(T) = \sup_{0 \leq t \leq T} (\| z(t) \|_{H^2} + \| z_t(t) \|_{H^1}) \). Local existence theorem can be proved by using a standard iterative method and we omit the proof. We derive the following a priori estimate of solutions \( z \) for (8) in the Sobolev space \( H^2 \).

**Proposition 1 (a priori estimate).** Suppose that the same assumptions as in Theorem 1.2 hold true. Let \( k = 0, 1 \). Then there exists a positive constant \( \varepsilon \) such that if \( z \in X_{k,\varepsilon}([0,T]) \) is the solution of the problem (8) for some \( T > 0 \), then the following a priori estimates hold.

(i) For the non-degenerate case: \( f'(0) < 0 \), it holds that

\[
\| z(t) \|_{H_k^2}^2 + \| z_t(t) \|_{H_k^1}^2 + \int_0^t (\| (z_t, z_z)(\tau) \|_{H_k^1}^2 + \| \sqrt{\varphi} z(\tau) \|_{L^2}^2) \ d\tau 
\]

\[
\leq C(\| z_0 \|_{H_k^2}^2 + \| z_1 \|_{H_k^1}^2),
\]

with \( k = 0 \), where \( t \in [0,T] \) and \( C \) is a positive constant which independent of \( T \).

(ii) For the degenerate case: \( f'(0) = 0 \), the a priori estimate (9) with \( k = 1 \) holds true.

To show Proposition 1, we use weighted energy method. We define a weight function as

\[
w(u) = (-e^{Au} + 1)/f(u) \quad \text{for} \quad u \in [u_-, 0],
\]

where \( A \) is a positive constant. Now, we apply Taylor expansion for the nonlinear term \( f(u) \) in (8). Since the nonlinear term \( f(u) \) is smooth, it is described as

\[
f(u) = a_{q+1}(0)(-u)^{q+1} + a_{q+2}(\theta u)(-u)^{q+2},
\]

for some constant \( \theta \) with \( 0 < \theta < 1 \), where \( q \) is a non-negative integer. For the weight function defined in (10), the following lemma is important.

**Lemma 2.1.** ([3]) Suppose that \( f(u) \) satisfies (2). Let \( w(u) \) be the weight function defined in (10). Then there exists a positive constant \( \delta \) such that if \( A \geq \delta \), then \( w(u) \) satisfies the following conditions:

(i) \( c(-u)^{-q} < w(u) < C(-u)^{-q} \),

(ii) \( \{(fw)'(u)\}^2 < w^2(u) \),

(iii) \( (fw)''(u) < 0 \),

(iv) \( (fw)'(u_-) < 0 \),

for \( u \in [u_-, 0] \), where non-negative integer \( q \) is the degeneracy exponent defined by (11), and \( C \) and \( c \) are some positive constants.

By using Lemma 2.1, we give the summary of proof of Proposition 1. The whole proof is given in [3].

**Outline of proof of Proposition 1.** Throughout this proof, we use the weighted \( L^2 \) norm:

\[
\| f \|_{L^2} := \left( \int_0^\infty w(\phi(x)) |f(x)|^2 \ dx \right)^{1/2},
\]
where $w$ is the weight function defined by (10). The first equality of (8) is rewritten as

$$z_{tt} - z_{xx} + z_t + f'(\phi)z_x = g,$$

where

$$g = -\{f(\phi + z_x) - f(\phi) - f'(\phi)z_x\} = O(|z_x|^2).$$

Multiplying (12) by $w(\phi)z$, then we get

$$\left\{\frac{1}{2}w(\phi)z^2 + w(\phi)zz_t\right\}_t + \left\{\frac{1}{2}f'(\phi)w(\phi)z^2 - w(\phi)zz_x + \frac{1}{2}w'(\phi)\phi_xz^2\right\}_x$$

$$- w(\phi)z_t^2 + w(\phi)z_x^2 = \frac{1}{2}(f w)'(\phi)\phi_xz^2 = w(\phi)z g.\tag{14}$$

We next multiply (12) by $w(\phi)z_t$. Then this yields

$$\frac{1}{2}\{w(\phi)z_t^2 + w(\phi)z_x^2\}_t - \{w(\phi)zz_t\}_x + w(\phi)z_t^2 + (fw)'(\phi)zz_x = w(\phi)z_t g.\tag{15}$$

Therefore, making a combination (14) + $2 \times$ (15), we obtain the following energy equality

$$E_t + D + F_x = R,\tag{16}$$

where we define that

$$E = w(\phi)\left(\frac{1}{2}z^2 + z_t^2 + z_x^2 + zz_t\right), \quad R = w(\phi)(z + 2z_t)g,$$

$$D = w(\phi)z_t^2 + w(\phi)z_x^2 + 2(f w)'(\phi)zz_x - \frac{1}{2}(f w)''(\phi)\phi_xz^2,\tag{17}$$

$$F = \frac{1}{2}(f w)'(\phi)z^2 - w(\phi)zz_x - 2w(\phi)z_t z_x.$$

Since $E$ is quadratic form in terms of $z$, $z_x$ and $z_t$, then calculating the discriminant of $E$, we have the following condition

$$E \sim w(\phi)(z^2 + z_t^2 + z_x^2).\tag{18}$$

In a similar way, since $D$ is quadratic form in terms of $z$, $z_x$, $z_t$ and $\sqrt{\phi_x}z$, then by using Lemma 2.1-(ii), we have

$$w(\phi)z_t^2 + w(\phi)z_x^2 + 2(f w)'(\phi)zz_x = w(\phi)\left\{z_t^2 + z_x^2 + 2\frac{(f w)'}{w}(\phi)z_t z_x\right\}$$

$$\geq c w(\phi)(z_t^2 + z_x^2).$$

Thus, by using this inequality and Lemma 2.1-(iii), this yields

$$D \geq c w(\phi)(z_t^2 + z_x^2) + c \phi_x z^2.\tag{19}$$

Now, integrating (16) in terms of $x$ and $t$, we have

$$\int_0^\infty E(x, t)\, dx + \int_0^t \int_0^\infty D(x, \tau)\, dx d\tau - \frac{1}{2}(f w)'(u_-) \int_0^t z(0, \tau)^2\, d\tau$$

$$\leq \int_0^\infty E(x, 0)\, dx + \int_0^t \int_0^\infty R(x, \tau)\, dx d\tau.$$
Therefore, applying (13), (18), (19) and Lemma 2.1-(iv), we get
\[
\|(z, z_t, z_{x_t})(t)\|_{L^2}^2 + \int_0^t \left( \|z(t, z_x)(\tau)\|_{L^2}^2 + \|\sqrt{\phi_x} z(\tau)\|_{L^2}^2 \right) d\tau + \int_0^t z(0, \tau)^2 d\tau \\
\leq C\|(z_0, z_1, z_{0, x})\|_{L^2}^2 + C\|(z, z_t)(t)\|_{L^\infty} \int_0^t \|z_x(\tau)\|_{L^2}^2 d\tau.
\]
Finally, using the relation \(\|(z, z_t)(t)\|_{L^\infty} \leq CN(t)\) and letting \(N(t)\) suitably small, we arrive at the energy estimate for the lower order derivatives as follows.
\[
\|(z, z_t, z_{x_t})(t)\|_{L^2}^2 + \int_0^t \left( \|z(t, z_x)(\tau)\|_{L^2}^2 + \|\sqrt{\phi_x} z(\tau)\|_{L^2}^2 \right) d\tau \\
\leq C\|(z_0, z_1, z_{0, x})\|_{L^2}^2.
\]
We reduce the higher order estimate in the similar way, so please refer to [3] for details.

3. Convergence rates of stationary solutions. In this section, we prove Theorems 1.3 and 1.4. The main ideas of the proofs are due to Ueda [10]. We use the space-time weighted energy method introduced in Kawashima-Matsumura [4].

Outline of proof of Theorem 1.3. We start with the energy equality (16). By using the property of the stationary wave \(\phi\), we have the following estimate:
\[
-F = -\frac{1}{2}(fw)'(\phi)^2 + w(\phi)(zz_x + 2z_t z_x) \\
\geq cw^2 - C(z_x^2 + z_t^2),
\]
where \(c\) and \(C\) are positive constants.

Let \(\gamma\) and \(\beta\) be any positive constants satisfying \(0 \leq \gamma, \beta \leq \alpha\). We multiply the equality (16) by \((1 + t)^\gamma (1 + x)^\beta\), obtaining
\[
\{(1 + t)^\gamma (1 + x)^\beta E\}t - \gamma (1 + t)^{\gamma-1}(1 + x)^\beta E + (1 + t)^\gamma (1 + x)^\beta D \\
+ \{(1 + t)^\gamma (1 + x)^\beta F\}x - \beta (1 + t)^{\gamma-1}(1 + x)^\beta F = (1 + t)^\gamma (1 + x)^\beta R.
\]
Integrating (21) over \((0, \infty) \times (0, t)\), substituting (18), (19) and (20) into the resultant inequality, and letting \(\sup_{0 \leq t < \infty} \|z(t)\|_{H^1}\) suitably small, we have
\[
(1 + t)^\gamma \|(z, z_t, z_{x_t})(t)\|_{L^2}^2 + \int_0^t (1 + t)^\gamma \|z(t, z_x, \sqrt{\phi_x} z)(\tau)\|_{L^2}^2 + \beta \|z(\tau)\|_{L^{\beta-1}}^2 d\tau \\
\leq CE_\beta^2 + C\int_0^t (1 + t)^{\gamma-1}(\|z(t, z_x)(\tau)\|_{L^2}^2 \|z(\tau)\|_{L^{\beta-1}}^2 + \beta C \int_0^t (1 + t)^\gamma \|(z_t, z_{x_t})(\tau)\|_{L^{\beta-1}}^2 d\tau,
\]
for an arbitrary \(\gamma\) and \(\beta\) with \(0 \leq \gamma, \beta \leq \alpha\), where \(C\) is a constant independent of \(\gamma\) and \(\beta\).

We can derive the higher order inequality in a similar way, and reduce the desired inequality in the same strategy as in [10].

In the rest of this paper, we prove Theorem 1.4 by using the space-time weighted energy method.

Outline of proof of Theorem 1.4. Let \(\alpha, \beta > 0\). Multiplying (16) by \(e^{\beta t} e^{\alpha x}\), we obtain
\[
\{e^{\beta t} e^{\alpha x} E\}_t - \beta e^{\beta t} e^{\alpha x} E + e^{\beta t} e^{\alpha x} D + \{e^{\beta t} e^{\alpha x} F\}_x - \alpha e^{\beta t} e^{\alpha x} F = e^{\beta t} e^{\alpha x} R.
\]
Then, integrating (23) over \((0, \infty) \times (0, t)\), substituting (18), (19) and (20) into the resultant inequality, and letting \(\sup_{0 \leq t < \infty} \|z(t)\|_{H^1}\) suitably small, we get

\[
e^{\beta t} \|(z, z_x, z_t)(t)\|_{L^2_{\alpha, \exp}}^2 + \int_0^t e^{\beta \tau} \|(z_x, z_t, \sqrt{\phi_x} z)(\tau)\|_{L^2_{\alpha, \exp}}^2 d\tau + \alpha \int_0^t e^{\beta \tau} \|z(\tau)\|_{L^2_{\alpha, \exp}}^2 d\tau \leq CE^{2\alpha, \exp} + \beta C \int_0^t e^{\beta \tau} \|(z, z_x, z_t)(\tau)\|_{L^2_{\alpha, \exp}}^2 d\tau + \alpha C \int_0^t e^{\beta \tau} \|(z_x, z_t)(\tau)\|_{L^2_{\alpha, \exp}}^2 d\tau,
\]

where \(C\) is positive constant independent of \(\alpha\) and \(\beta\).

We can derive the higher order inequality in a similar way, and reduce the desired inequality in the same strategy as in \([10]\). Please refer to \([3]\) for detail. \(\square\)

REFERENCES


E-mail address: itsuko@nc-toyama.ac.jp
POROUS MEDIA EQUATIONS, FAST DIFFUSIONS EQUATIONS AND THE EXISTENCE OF GLOBAL WEAK SOLUTION FOR THE QUASI-SOLUTIONS OF COMPRESSIBLE NAVIER-STOKES EQUATIONS

Boris Haspot
Ceremade UMR CNRS 7534 Université de Paris Dauphine
Place du Maréchal DeLattre De Tassigny
75775 Paris Cedex 16, France

Abstract. In [3, 4, 5], we have developed a new tool called quasi solutions which approximate in some sense the compressible Navier-Stokes equation. In particular it allows to obtain global strong solution for the compressible Navier-Stokes equations with large initial data on the irrotational part of the velocity (large in the sense that the smallness assumption is subcritical in terms of scaling, it turns out that in this framework we are able to obtain large initial data in the energy space in dimension $N = 2$). In this paper we are interested in studying in details this notion of quasi solution and in particular proving global weak solution, we also observe that for some choice of initial data (irrotational) we obtain some quasi solutions verifying the porous medium equation, the heat equation or the fast diffusion equation in function of the structure of the viscosity coefficients. Finally we show the convergence of the global weak solution of compressible Navier-Stokes equations to the quasi solutions when the pressure vanishing.

1. Introduction. The motion of a general barotropic compressible fluid is described by the following system:

$$
\begin{align*}
\partial_t \rho + \text{div}(\rho u) &= 0, \\
\partial_t (\rho u) + \text{div}(\rho u \otimes u) - \text{div}(\mu(\rho)D(u)) - \nabla(\lambda(\rho)\text{div}(u)) + \nabla P(\rho) &= \rho f, \\
(\rho, u)|_{t=0} &= (\rho_0, u_0).
\end{align*}
$$

(1)

Here $u = u(t, x) \in \mathbb{R}^N$ stands for the velocity field, $\rho = \rho(t, x) \in \mathbb{R}^+$ is the density and $D(u) = \frac{1}{2}(\nabla u + (\nabla u)^t)$. The pressure $P$ is such that $P(\rho) = a\rho^\gamma$ with $\gamma \geq 1$. We denote by $\mu(\rho) > 0$ and $2\mu(\rho) + N\lambda(\rho) > 0$ the viscosity coefficients of the fluid. Throughout the paper, we assume that the space variable $x \in \mathbb{R}^N$.

In this paper we are interested in studying the notion of quasi-solution developed in [4, 5, 3] for general viscosity coefficients following the algebraic equality discovered by Bresch and Desjardin in [1]:

$$
\lambda(\rho) = 2\rho \mu'(\rho) - 2\mu(\rho).
$$

(2)

We shall define in the sequel the function $\varphi(\rho)$ by $\varphi'(\rho) = \frac{2\mu'(\rho)}{\rho}$. With this choice of viscosity coefficients Bresch and Desjardin have obtained a new entropy giving a $L^2$ control on the gradient of the density. It has permit to Mellet and Vasseur in
to prove the stability of the global weak solution for compressible Navier Stokes equations with such viscosity coefficients and with $\gamma$ law pressure $P(\rho) = \rho^\gamma$ with $a > 0$ and $\gamma > 1$. In the sequel we will work only with such viscosity coefficients verifying the relation (2).

This paper is devoted to prove the existence of quasi solutions for compressible Navier-Stokes equations with degenerate viscosity coefficients. Let recall the definition of quasi solutions introduced in [3, 4, 5].

**Definition 1.1.** We say that $(\rho, u)$ is a quasi solution if $(\rho, u)$ verifies in distribution sense:

$$
\begin{aligned}
&\frac{\partial}{\partial t} \rho + \text{div}(\rho u) = 0, \\
&\frac{\partial}{\partial t} (\rho u) + \text{div}(\rho u \otimes u) - \text{div}(2\mu(\rho) \text{Du}) - \nabla(\lambda(\rho) \text{div} u) = 0,
\end{aligned}
$$

(3)

More precisely $(\rho, u)$ is a weak solution of (3) on $[0, T] \times \mathbb{R}^N$ with:

$$
\begin{aligned}
&\rho_0 L^1(\mathbb{R}^N), \quad \sqrt{\rho_0} \nabla \varphi(\rho_0) \in L^2(\mathbb{R}^N), \quad \rho_0 \geq 0, \\
&\sqrt{\rho_0 |u_0| \ln(1 + |u_0|^2)} \in L^2(\mathbb{R}^N).
\end{aligned}
$$

(4)

if

- $\rho \in L^\infty_t(L^1(\mathbb{R}^N)), \sqrt{\rho} \nabla \varphi(\rho) \in L^\infty_t(L^2(\mathbb{R}^N)), \mu(\rho) \nabla u \in L^2((0, T) \times \mathbb{R}^N), \sqrt{\rho(t)} \nabla \varphi \ln(1 + |u|^2) \in L^2_t(L^2(\mathbb{R}^N))$.

with $\rho \geq 0$ and $(\rho, \sqrt{\rho} u)$ satisfying in distribution sense on $[0, T] \times \mathbb{R}^N$:

$$
\begin{aligned}
&\partial_t \rho + \text{div}(\sqrt{\rho} u) = 0, \\
&\rho(0, x) = \rho_0(x).
\end{aligned}
$$

and if the following equality holds for all $\varphi(t, x)$ smooth test function with compact support such that $\varphi(T, \cdot) = 0$:

$$
\int_{\mathbb{R}^N} (\rho u)_0 \cdot \varphi(0, \cdot) dx + \int_0^T \int_{\mathbb{R}^N} \sqrt{\rho} \sqrt{\rho u} \partial_t \varphi + \sqrt{\rho} \otimes \sqrt{\rho u} : \nabla \varphi dx dt
\begin{aligned}
&- < 2\mu(\rho) \text{Du}, \sqrt{\rho} \nabla \varphi > - < \lambda(\rho) \text{div} u, \text{div} \varphi > = 0,
\end{aligned}
$$

(5)

where we give sense to the diffusion terms by rewriting him according to $\sqrt{\rho}$ and $\sqrt{\rho} u$:

$$
< 2\mu(\rho) \text{Du}, \sqrt{\rho} \nabla \varphi > = - \int \frac{\mu(\rho)}{\sqrt{\rho}} (\sqrt{\rho} u_j) \partial_i \varphi_j dx dt - \int 2(\sqrt{\rho} u_j) \mu(\rho) \partial_i \sqrt{\rho} \partial_i \varphi_j dx dt
\begin{aligned}
&- \int \frac{\mu(\rho)}{\sqrt{\rho}} (\sqrt{\rho} u_j) \partial_i \varphi_j dx dt - \int 2(\sqrt{\rho} u_j) \mu(\rho) \partial_i \sqrt{\rho} \partial_i \varphi_j dx dt
\end{aligned}
$$

$$
< \lambda(\rho) \text{div} u, \text{div} \varphi > = - \int \frac{\lambda(\rho)}{\sqrt{\rho}} (\sqrt{\rho} u) \partial_i \varphi_j dx dt - \int 2(\sqrt{\rho} u_j) \lambda(\rho) \partial_i \sqrt{\rho} \partial_i \varphi_j dx dt
\begin{aligned}
&- \int \frac{\lambda(\rho)}{\sqrt{\rho}} (\sqrt{\rho} u) \partial_i \varphi_j dx dt - \int 2(\sqrt{\rho} u_j) \lambda(\rho) \partial_i \sqrt{\rho} \partial_i \varphi_j dx dt
\end{aligned}
$$

We assume also the same extra assumption on the viscosity coefficient 8-12 than in [6].

**Remark 1.** Here $\lambda$ and $\mu$ verifies the condition (2), in particular we have classical energy estimates by multiplying the momentum equation by $u$ except that we have no information on the density of the type $\rho^\gamma \in L^\infty(\mathbb{R}^+, L^1(\mathbb{R}^N))$ as for compressible
Navier-Stokes equation when \( P(\rho) = \rho^\gamma \) with \( \gamma \geq 1 \) because here \( P(\rho) = 0 \). However using the entropy discovered in [1] we can prove that \( \sqrt{\rho} \nabla \varphi(\rho) \) belongs in \( L^\infty(\mathbb{R}^+, L^2(\mathbb{R}^N)) \) with \( \varphi'(\rho) = \frac{2 \mu'(\rho)}{\rho} \) and that \( \rho \) belongs in \( L^\infty(\mathbb{R}^+, L^1(\mathbb{R}^N)) \) by conservation of the mass. It will be sufficient to prove the stability of global weak solution and it explains the assumption of the definition 1.1.

**Remark 2.** Let us remark that by using this notion of quasi solution in [3] we obtain global strong solution with initial data small in subcritical space for the scaling of the equations. In this sense quasi solutions are good approximate in order to obtain global strong solution with large initial data in terms of scaling (in particular in dimension \( N = 2 \) we can choose large initial data in energy space).

We now are going to investigate the existence of such quasi solution for the viscosity coefficients verifying (2). More precisely as in [3] we are going to search in a first time irrotational solution under the form \( u(t, x) = \nabla \psi(t, x) \). Let us assume now that:

\[
\mu(\rho) = \mu \rho^\alpha \quad \text{with} \quad \alpha > 0 \quad \text{and} \quad \lambda(\rho) = 2(\alpha - 1)\mu \rho^\alpha,
\]

with \( \alpha > 1 - \frac{1}{N} \) in order to insure the relation \( 2\mu(\rho) + N\lambda(\rho) > 0 \). Furthermore we observe that \( \mu(\rho) \) and \( \lambda(\rho) \) verify the relation (2).

Let us now briefly recall the so-called porous medium and fast diffusion equations before explaining the link between the quasi-solutions and these solutions. More precisely the solutions of the following the nonlinear Cauchy problem:

\[
\begin{align*}
\frac{\partial \rho}{\partial t} - \mu \Delta \rho^\alpha &= 0, \\
\rho(0, \cdot) &= \rho_0,
\end{align*}
\]

where \( \alpha \) is a positive number which we assume different from one are solutions of the porous media equation or fast diffusion equations, here we assume that \( \rho_0 \in L^1(\mathbb{R}^N) \) is nonnegative. The case \( \alpha > 1 \) (the porous media equations) arises as a model of slow diffusion of a gas inside a porous container. Unlike the heat equation \( \alpha = 1 \), this equation exhibits finite speed of propagation in the sense that solutions associated to compactly supported initial data remain compactly supported in space variable at all times (see [7]). When \( 0 < \alpha < 1 \), the opposite happens. Infinite speed of propagation occurs and solutions may even vanish in finite time. This problem is usually referred to as the fast diffusion equation.

Let us recall the notion of global strong solution for the equation (7) of the porous medium equation \( (\alpha > 1) \) and of the fast diffusion equation \( (0 < \alpha < 1) \) (see [7] chapter 9 for more details and [8]).

**Definition 1.2.** We say that a function \( \rho \in C([0, +\infty), L^1(\mathbb{R}^N)) \) positive is a strong \( L^1 \) solution of problem (7) if:

\[
\begin{itemize}
\item \( \rho^\alpha \in L^1_{loc}(0, +\infty, L^1(\mathbb{R}^N)) \) and \( \rho_t, \Delta \rho^\alpha \in L^1_{loc}((0, +\infty) \times \mathbb{R}^N) \)
\item \( \rho_t = \mu \Delta \rho^\alpha \) in distribution sense.
\item \( u(t) \rightarrow \rho_0 \) as \( t \rightarrow 0 \) in \( L^1(\mathbb{R}^N) \).
\end{itemize}

Let us mention (see [7]) that we have the following theorem:

**Theorem 1.3.** Let \( \alpha > 0 \). For every \( \rho_0 \in L^1(\mathbb{R}^N) \) positive there exists a unique global strong solution \( \rho \) positive of problem (7) such that \( \rho \in C([0, +\infty), L^1(\mathbb{R}^N)) \cap L^\infty((\tau, +\infty) \times \mathbb{R}^N) \) for every \( \tau > 0 \).

Let us recall that there exists global solution which are not classical it means not \( C^\infty \) even if the initial data is \( C^\infty \) (see a example due to Aronson in the problem 5.7 of [7]).
Remark 3. Let us mention that if the initial data is non-negative then the unique
global weak solution solution of the system (3) of the form
\[ \alpha \]

when \( \alpha > 1 \) and of the fast diffusion when \( 0 < \alpha < 1 \) that is why we assume that \( u = 0 \) on the vacuum set. In other

We refer for more details on the
to the books of J-L Vázquez (see [7, 8]).

Proof of Theorem 1.4. Let us assume in a first time that the solution \((\rho, u)\) of system
(3) are classical, we are going to search solution under the form: \((\rho, -\frac{2\mu\alpha}{\alpha - 1} \nabla \rho^{\alpha - 1})\).

The mass equation give us:
\[\begin{align*}
\partial_t \rho - 2\mu \Delta \rho^\alpha &= 0, \\
\rho(0, \cdot) &= \rho_0.
\end{align*}\]  

(8)

When \( \alpha = 1 \) with \((\rho_0, u_0) = -2\mu \nabla \ln \rho_0\), similarly we have particular global weak
unique solution solution of the system (3) of the form \((\rho, u = -2\mu \nabla \ln \rho)\) solving
the heat equation:
\[\begin{align*}
\partial_t \rho - 2\mu \Delta \rho &= 0, \\
\rho(0, \cdot) &= \rho_0.
\end{align*}\]  

(9)

Remark 4. Let us point out that any solution of (7) such that \( \rho \) is in \( C^3((0, +\infty) \times \mathbb{R}^N) \) is a classical solution of (3). In the case where \( \rho = 0 \) the velocity is not defined
when \( 0 < \alpha < 1 \) is that why we assume that \( u = 0 \) on the vacuum set. In other
we could give sense to \( \rho u \) as in [6].

Remark 5. We observe as in [2] that if we consider the compressible Navier-Stokes equation with a friction term \( a \rho u \) and a pressure of the form \( 2\mu \rho^{\alpha} \) then the
same solution than theorem 1.4 verify a such system.

Remark 6. We recognize here the so called equation of the porous medium when
\( \alpha > 1 \) and of the fast diffusion when \( 0 < \alpha < 1 \). We refer for more details on the
the theory to the books of J-L Vázquez (see [7, 8]).

\[ \partial_t \rho - 2\mu \Delta \rho^\alpha = 0 \]  

(10)

Let us check that the second equation is compatible with the first and keep an
irrotational structure. First we have:

\[ \partial_t (\rho u) = -\frac{2\mu\alpha}{\alpha - 1} \partial_t (\rho \nabla \rho^{\alpha - 1}) = -2 \partial_t \nabla \rho^\alpha. \]

\[ \text{div}(\rho u \otimes u) = \frac{4\mu^2 \alpha^2}{(\alpha - \frac{1}{2})^2} (\Delta \rho^{\alpha - \frac{1}{2}} \nabla \rho^{\alpha - \frac{1}{2}} + \frac{1}{2} \nabla |\nabla \rho^{\alpha - \frac{1}{2}}|^2). \]

(11)

Indeed we have:

\[ \text{div}(\rho u \otimes u)_j = \frac{4\mu^2 \alpha^2}{(\alpha - \frac{1}{2})^2} \sum_i \partial_i (\rho \partial_j \rho^{\alpha - 1} \partial_j \rho^{\alpha - 1}) = 4\alpha^2 \mu^2 \sum_i \partial_i (\rho^{2\alpha - 3} \partial_i \rho \partial_j \rho) \]

\[ = \frac{4\mu^2 \alpha^2}{(\alpha - \frac{1}{2})^2} \sum_i \partial_i (\partial_j \rho^{\alpha - \frac{1}{2}} \partial_j \rho^{\alpha - \frac{1}{2}}) = \frac{4\mu^2 \alpha^2}{(\alpha - \frac{1}{2})^2} (\Delta \rho^{\alpha - \frac{1}{2}} \partial_j \rho^{\alpha - \frac{1}{2}} + \frac{1}{2} \partial_j |\nabla \rho^{\alpha - \frac{1}{2}}|^2). \]
Finally we have from (11):
\[-\nabla(\lambda(\rho)\nabla u) = 2(\alpha - 1)\mu^2 \nabla(\rho^\alpha \nabla(\rho^\alpha - 1)) = 4\alpha \mu^2 \nabla(\rho^\alpha \Delta \rho^\alpha - 1).\] (12)

Finally by combining (12) and (13) we obtain:
\[\text{div}(\rho u \otimes u) = \frac{4\alpha \mu^2}{(\alpha - 1)} \Delta \rho^\alpha - 1 \nabla \rho^\alpha + \frac{2\alpha^2 \mu^2}{(\alpha - \frac{1}{2})^2} \rho^{\alpha - \frac{1}{2}} |\nabla \rho|^2 \nabla \rho^{\alpha - \frac{1}{2}} + \frac{2\alpha^2 \mu^2}{(\alpha - \frac{1}{2})^2} \nabla |\nabla \rho^{\alpha - \frac{1}{2}}|^2.\] (13)

by using the fact that:
\[\Delta \rho^\alpha - \frac{1}{2} = \sum_i \partial_i \left( \frac{\alpha - 1}{\alpha - 1} \rho_1 \rho^{\alpha - 1} \rho_i \right) = \frac{\alpha - 1}{\alpha - 1} \rho^\alpha \Delta \rho^\alpha - 1 + \frac{1}{2} (\alpha - 1) \rho^{\alpha - \frac{1}{2}} |\nabla \rho|^2.\]

Finally by combining (12) and (13) we obtain:
\[\text{div}(\rho u \otimes u) - \text{div}(2\rho^\alpha Du) = \frac{4\alpha \mu^2}{(\alpha - 1)} \nabla(\rho^\alpha \Delta \rho^\alpha - 1) + \frac{4\alpha \mu^2}{(\alpha - 1)} \nabla \rho^\alpha \cdot \nabla \nabla \rho^\alpha - 1 + \frac{2\alpha^2 \mu^2}{(\alpha - \frac{1}{2})^2} \rho^{\alpha - \frac{1}{2}} |\nabla \rho|^2 \nabla \rho^{\alpha - \frac{1}{2}} + \frac{2\alpha^2 \mu^2}{(\alpha - \frac{1}{2})^2} \nabla |\nabla \rho^{\alpha - \frac{1}{2}}|^2.\] (14)

Now since we have:
\[\nabla \rho^\alpha \cdot \nabla \nabla \rho^\alpha - 1 = \frac{\alpha (\alpha - 1)}{2(\alpha - \frac{1}{2})^2} |\nabla \rho^{\alpha - \frac{1}{2}}|^2 - \frac{\alpha (\alpha - 1)}{2(\alpha - \frac{1}{2})} \rho^{\alpha - \frac{1}{2}} |\nabla \rho|^2 \nabla \rho^{\alpha - \frac{1}{2}}\]

indeed it is due to the following calculus:
\[(\nabla \rho^\alpha \cdot \nabla \nabla \rho^\alpha - 1)_j = \sum_i \partial_i \rho^\alpha \partial_{ij} \rho^{\alpha - 1}\]
\[= \sum_i \frac{\alpha}{\alpha - \frac{1}{2}} \rho^\alpha \partial_i \rho_1 \rho^{\alpha - \frac{1}{2}} \partial_i \left( \frac{\alpha - 1}{\alpha - \frac{1}{2}} \rho_1 \rho^{\alpha - \frac{1}{2}} \partial_j \rho^{\alpha - \frac{1}{2}} \right)\]
\[= \frac{\alpha (\alpha - 1)}{(\alpha - \frac{1}{2})^2} \sum_i \partial_i \rho^{\alpha - \frac{1}{2}} \partial_{ij} \rho^{\alpha - \frac{1}{2}} \left( \frac{\alpha - 1}{\alpha - \frac{1}{2}} \rho^{\alpha - \frac{1}{2}} \partial_j \rho^{\alpha - \frac{1}{2}} \right) - \frac{1}{2} \rho^{\alpha - \frac{1}{2}} \partial_j \rho^{\alpha - \frac{1}{2}} \partial_i \rho^{\alpha - \frac{1}{2}} \partial_i \rho\]

we finally reduce (14) to the following equation:
\[\nabla \rho^\alpha \cdot \nabla \nabla \rho^\alpha - 1 = \frac{\alpha (\alpha - 1)}{2(\alpha - \frac{1}{2})^2} |\nabla \rho^{\alpha - \frac{1}{2}}|^2 - \frac{\alpha (\alpha - 1)}{2(\alpha - \frac{1}{2})} \rho^{\alpha - \frac{1}{2}} |\nabla \rho|^2 \nabla \rho^{\alpha - \frac{1}{2}}\]

We finally have:
\[\text{div}(\rho u \otimes u) - \text{div}(2\rho^\alpha Du) = \frac{4\mu^2 \alpha}{(\alpha - 1)} \nabla(\rho^\alpha \Delta \rho^\alpha - 1) + \frac{4\mu^2 \alpha^2}{(\alpha - \frac{1}{2})^2} \nabla |\nabla \rho^{\alpha - \frac{1}{2}}|^2.\] (15)
Finally using (11), (12) and (15) we obtain:
\[
\frac{\partial}{\partial t}(\rho u) + \text{div}(\rho u \otimes u) - \text{div}(2\rho^\alpha Du) - \nabla(\rho \text{div} u) = \\
-2\mu \nabla(\partial_\rho \rho^\alpha - \frac{2\mu}{\alpha - 1})\rho^\alpha \Delta \rho^{\alpha - 1} - \frac{2\mu\alpha^2}{(\alpha - \frac{1}{2})^2} |\nabla \rho^{\alpha - \frac{1}{2}}|^2 - 2\mu \rho^\alpha \Delta \rho^{\alpha - 1},
\]
\[
= -2\alpha \mu \nabla(\rho^{\alpha - 1}(\partial_\rho - \frac{2\mu}{\alpha - 1}) \rho \Delta \rho^{\alpha - 1} - 2\mu \rho^\alpha |\nabla \rho|^2 - 2\mu \rho \Delta \rho^{\alpha - 1})
\]
\[
= -2\alpha \mu \nabla(\rho^{\alpha - 1}(\partial_\rho - 2\mu \Delta \rho^\alpha)).
\]
This concludes the proof inasmuch as via the above equation the momentum equation is compatible to the mass equation and verify the system (7). But when we take initial density in \(L^1\) non negative, we know via the remark 3 that the unique global solution of (7) is classical and non negative. It justify in particular all the previous formal calculus and prove that \((\rho, u) = -\frac{2\mu}{\alpha - 1} \nabla \rho^{\alpha - 1}\) is a classical solution of (3) with \(\rho\) verifying (7). It concludes the proof. \(\square\)

**Remark 7.** More generally we have solution of the form \((\rho, -\nabla \varphi(\rho))\) with \(\rho\) verifying the more general porous media equation (we refer to [7] for such equations):
\[
\partial_\rho \rho - \text{div}(\rho \nabla \varphi(\rho)) = 0.
\]

**Remark 8.** Let us mention that when \(\alpha\) is in the interval \((0, m_c)\) with \(m_c = \max(0, \frac{N}{2N - 2})\) then it can appears a phenomena of extinction of the solution in finite time, in particular it implies a lost of the initial mass when \(\rho_0\) is in \(L^1\) (it is not the case in our framework because \(1 - \frac{1}{N} > m_c\). Let us recall that when \(\alpha > 1\) and \(\rho_0\) belongs in \(L^1\), it exists global unique weak solution and that the solution converges asymptotically to the so called Barrenblatt solution
\[
U_m(t, x) = t^{-\gamma}F\left(\frac{x}{t}\right) \quad \text{with} \quad F(x) = (C - \frac{\alpha - 1}{2\alpha} |x|^2 + \frac{1}{\alpha} x),
\]
which are self similar.

Finally we obtain the following theorems.

**Theorem 1.5.** Assume that we have a sequence \((\rho_n, u_n)\) of global weak solutions of system (3) satisfying the entropies of [6]) with initial data \(\rho_0^\alpha\) and \(u_0^\alpha\) such that:
\[
\rho_0^\alpha \geq 0, \quad \rho_0^\alpha \to \rho_0 \quad \text{in} \quad L^1(\mathbb{R}^N), \quad \rho_0^\alpha u_0^\alpha \to \rho_0 u_0 \quad \text{in} \quad L^1(\mathbb{R}^N),
\]
and satisfy the following bounds (with \(C\) constant independent on \(n\)):
\[
\int_{\mathbb{R}^N} \rho_0^\alpha \frac{|u_0^\alpha|^2}{2} < C, \quad \int_{\mathbb{R}^N} \sqrt{\rho_0^\alpha} |\nabla \varphi(\rho_0^\alpha)|^2 dx < C \quad \text{and} \quad \int_{\mathbb{R}^N} \rho_0^\alpha \frac{1 + |u_0^\alpha|^2}{2} \ln(1 + |u_0^\alpha|^2) dx < C.
\]
Then, up to a subsequence, \((\rho_n, \sqrt{\rho_n} u_n)\) converges strongly to a weak solution \((\rho, \sqrt{\rho} u)\) of (3) satisfying entropy inequalities of [6].

Furthermore when we choose \((\rho_0, u_0) = (\rho_0, -\frac{2\mu}{\alpha - 1} \nabla \rho_0^{\alpha - 1})\) with \(\alpha \neq 1\) it exists global weak quasi solutions in the sense of the definition 1.1.
Theorem 1.6. Assume that there exists global weak solution \((\rho_\varepsilon,u_\varepsilon)\) verifying the definition of [6] with the conditions (8) – (12) on \(\mu\) and \(\lambda\) of [6] and with the same restriction on \(\gamma\) as in [6] (see also theorem 2.1 in [6]) of the system:

\[
\begin{align*}
\partial_t \rho_\varepsilon + \nabla \cdot (\rho_\varepsilon u_\varepsilon) &= 0, \\
\partial_t (\rho_\varepsilon u_\varepsilon) + \nabla \cdot (\rho_\varepsilon u_\varepsilon \otimes u_\varepsilon) - \nabla (2\mu(\rho_\varepsilon)D(u_\varepsilon)) - \nabla (\lambda(\rho_\varepsilon) \text{div} u_\varepsilon) + \varepsilon \nabla \rho_\varepsilon^\gamma &= 0, \\
(\rho_\varepsilon, u_\varepsilon)_{t=0} &= (\rho_0, u_0).
\end{align*}
\]

(18)

then \((\rho_\varepsilon,u_\varepsilon)\) converges in distribution sense to a quasi-solution \((\rho,u)\) when \(\varepsilon\) goes to 0 with initial data \((\rho_0, u_0)\). (Here \((\rho_0, u_0)\) verifies the entropies of [6]).

Proof of Theorem 1.5 and 1.6. Concerning the stability of global weak solution, assume the existence of a sequel \((\rho_n,u_n)_{n\in\mathbb{N}}\) of global weak solution in the sense of [6] then it suffices to observe that we have as in [1, 6] the following uniform bounds in \(\rho\) \(n\):

\[
\rho_0 \left| u_n(t,x) \right|^2 (t,x) dx + \int_0^t \int_{\mathbb{R}^N} \mu(\rho_n)|D u_n|^2 dxdt + \int_0^t \int_{\mathbb{R}^N} \lambda(\rho_n) |\text{div} u_n|^2 dxdt \leq \int_{\mathbb{R}^N} \rho_0^n |u_0^n|^2(x) dx.
\]

Finally as in [6] it remains to obtain a gain of integrability on the velocity \(u_n\), it is obvious by multiplying the momentum equation by \((1 + \ln(1 + |u_n|^2))u_n\) and by bootstrap argument assuming the condition (11) as in [6]. In particular we show that \(\rho_n^{\frac{1}{\alpha} + \frac{|u_n|^2}{2}} \ln(1 + |u_n|^2)\) is uniformly bounded in \(L^\infty((0,T),L^1(\mathbb{R}^N))\) for any \(T > 0\). And by the conservation of the mass \(\rho_n\) is uniform bounded in \(L^\infty(\mathbb{R}^N)\) for any \(T > 0\). By using these different entropies as in [6] we can easily show the convergence in distribution sense of \((\rho_n, u_n)\) to \((\rho, u)\) via compactness argument.

We now want to prove the global existence of weak solution when \((\rho_0, u_0) = (\rho_0, -\frac{2u_0}{\alpha^2} \nabla \rho_0^{-1})\) with \(\alpha \neq 1\). To do this it only suffices to construct a sequel of regular global weak solution \((\rho_n, u_n)\) verifying uniformly in \(n\) the entropies of theorem 1.5 and to use the previous result. Indeed \(\rho_n, \sqrt{\rho_n} u_n\) must converge in distribution sense to a global weak solution \((\rho, \sqrt{\rho} u)\). Let us define \((\rho_n, u_n)\) as the solutions of the theorem 1.4 with initial data \(\rho_0^n = \rho_0 + \frac{1}{n} f\) where \(f\) is in \(L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N) \cap W^{1,1}(\mathbb{R}^N)\) (for \(\alpha \geq 1\)), and is non-negative and we have \(u_0^n = -\frac{2u_0}{\alpha^2} \nabla (\rho_0^n)^{\alpha-1}\). Then by theorem 1.4 it exists global regular weak solution \((\rho_n, u_n)\) and \((\rho^n_0, \rho^n_0 u^n_0)\) converges to \((\rho_0, u_0)\) which concludes the proof of theorem 1.5.

We are now going to prove that if we have some global weak solution \((\rho_\varepsilon, u_\varepsilon)\) for the system (18) in the sense of the definition in [6], then these global weak solution converge in distribution sense to a quasi-solution with initial data \((\rho_0, u_0)\). It suffices then to obtain the same uniform entropies in \(\varepsilon\) than in [6] for the sequel.
Similarly we have:

\[
\int_{\mathbb{R}^N} \left[ \rho_\varepsilon |u_\varepsilon(t,x)|^2(t,x) + \frac{\varepsilon}{\gamma - 1} |\rho_\varepsilon|^2 \right] dx + \int_0^t \int_{\mathbb{R}^N} \mu(\rho_\varepsilon) |Du_\varepsilon|^2 dxdt \\
+ \int_0^t \int_{\mathbb{R}^N} \lambda(\rho_\varepsilon) \text{div} u_\varepsilon^2 dxdt \leq \int_{\mathbb{R}^N} \left[ |\rho_0| u_0^2(x) + \frac{\varepsilon}{\gamma - 1} |\rho_0|^2 \right] dx.
\]

\[
\int_{\mathbb{R}^N} \left[ \rho_\varepsilon |u_\varepsilon(t,x)|^2 + \rho_\varepsilon |\nabla \varphi(\rho_\varepsilon)|^2 \right] dx + \varepsilon \int_0^t \int_{\mathbb{R}^N} \nabla \varphi(\rho_\varepsilon) \cdot \nabla \rho_\varepsilon^2 dxdt \\
\leq C\left( \int_{\mathbb{R}^N} \left( |\rho_0| u_0^2(x) + \rho_0 |\nabla \varphi(\rho_0)|^2 \right) dx + \frac{\varepsilon}{\gamma - 1} \rho_0^2(x) \right) dx.
\]

(20)

By using the lemma 3.2 of [6] with \( \forall \delta \in (0, 2) \), we have that:

\[
\int_{\mathbb{R}^N} \frac{1 + |u_\varepsilon|^2}{2} \ln(1 + |u_\varepsilon|^2)(t,x) dx \\
+ \nu \int_0^t \int_{\mathbb{R}^N} \mu(\rho_\varepsilon)(1 + \ln(1 + |u_\varepsilon|^2)) |Du_\varepsilon|^2 dxdt \\
\leq C \int_0^t \int_{\mathbb{R}^N} \mu(\rho_\varepsilon) |\nabla u_\varepsilon|^2 dxdt + C_d \int_0^t \left( \int_{\mathbb{R}^N} \frac{\rho_\varepsilon^{2\gamma - \frac{2}{\gamma}}}{\mu(\rho_\varepsilon)} dx dt \right). 
\]

(21)

We can easily observe that via the energy estimates the right hand side of (21) is uniformly bounded in \( \varepsilon \). The last step corresponds to use the same compactness argument than in [6] to show that \((\rho_\varepsilon, u_\varepsilon)\) converges in distribution sense to a quasi-solution \((\rho, u)\) when \( \varepsilon \) goes to 0 with initial data \((\rho_0, u_0)\). Let us point out that \( \varepsilon \rho_\varepsilon^2 \) goes to 0 in distribution sense. Indeed it suffice to observe that \( \rho_\varepsilon \) and \( \sqrt{\rho_\varepsilon \nabla \varphi(\rho_\varepsilon)} \) are uniformly bounded in \( \varepsilon \) respectively in \( L^\infty((0,T), L^1(\mathbb{R}^N)) \) and \( L^\infty((0,T), L^2(\mathbb{R}^N)) \) for any \( T > 0 \). In particular by Sobolev embedding and interpolation (for \( \gamma \) not so large as in [6]) we obtain that \( \rho_\varepsilon^\alpha \) with \( \alpha > 0 \) is uniformly bounded in \( L^1_{\text{loc}} \) what means that \( \varepsilon \rho_\varepsilon^2 \) goes to 0. \( \square \)

REFERENCES


E-mail address: haspot@ceremade.dauphine.fr
CARTESIAN GRID EMBEDDED BOUNDARY METHODS
FOR HYPERBOLIC PROBLEMS

CHRISTIANE HELZEL
Department of Mathematics
Ruhr-University Bochum
Universitätsstr. 150
44780 Bochum, Germany

MARSHA J. BERGER
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012, USA

Abstract. We describe our recent and current efforts towards the construction of Cartesian grid embedded boundary methods for hyperbolic pdes.

1. Introduction. We are developing finite volume methods for hyperbolic pdes on Cartesian grids with embedded boundaries. Embedded boundary methods are very attractive for several reasons: The grid generation is simple even in the presence of complicated geometries. Furthermore, such an approach allows the use of regular Cartesian grid methods away from the embedded boundary, which are much simpler to construct and more accurate than unstructured grid methods. In embedded boundary grids with cut cells adjacent to the boundary, the cut cell volumes can be orders of magnitude smaller than a regular Cartesian grid cell volume. The use of standard difference procedures would lead to an unacceptably small integration timestep. Both accuracy and stability are issues that need to be addressed at these highly irregular cut cells adjacent to solid bodies. The goal is to construct a method which is stable for time steps that are appropriate for the regular part of the mesh and at the same time do not lead to a loss of accuracy.

To overcome the small cell problem in an embedded boundary approach, Colella et al. [3] proposed a flux redistribution method. In this approach a fraction of the flux balance, proportional to the size of the small cell, is used to update the conserved quantities in the small cell. The rest of the flux balance is distributed to the neighboring cells. While flux redistribution has been used for several complex applications, it leads to a small loss of accuracy in the cut cells. With our approach, we want to avoid a loss of accuracy in the cut cells. Several other authors use some form of cell merging where a cut cell is combined with one or more neighboring...
cells, see for example [5, 6, 9]. While this is conceptually simple, it is in practice difficult to implement, especially in 3d.

Recently, cut cell methods find increased interest in the meteorological community, see [8, 10]. It is expected that such methods will lead to more accurate simulations for flow in mountainous terrains.

Here we give a brief overview of our recent and current work in this direction. Our approach is based on constructing fluxes at cut cells in such a way that a certain cancellation property is satisfied, see [1]. This means that the flux differences used to update the cell average of the conserved quantity in the small cell must be of the order of the size of the small grid cell. Such a property can be obtained by introducing $h$-boxes (i.e., boxes of the length of a regular grid cell) at cut cell interfaces as explained in more detail below. More recently, see [2], we have constructed embedded boundary methods in the context of the method of lines which led to several simplifications compared to the previously proposed MUSCL-type method [7]. Our current goal is to construct cut cell methods which are higher than second order accurate. Furthermore, we are interested in exploring further possible simplifications of cut cell methods. Preliminary results in this direction, still focusing on second-order methods, will be presented in Section 4.

2. A one–dimensional model problem. To illustrate the basic idea behind the construction of our Cartesian grid embedded boundary methods, we first consider a one–dimensional model problem. Here we consider approximations of equation

$$\partial_t q(x, t) + \partial_x f(q(x, t)) = 0, \quad -\infty < x < \infty$$

with initial values $q(x, 0) = q_0(x)$ of compact support. We use an almost equidistant finite volume mesh with grid cells of width $h$ except for one small cell (with index $k$) which has mesh width $\alpha h$, $0 < \alpha \leq 1$.

We consider finite volume methods of the the form

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x_i} \left( F_i^{n+\frac{1}{2}} - F_i^{n-\frac{1}{2}} \right)$$

or of the semi–discrete form

$$Q_i'(t) = -\frac{1}{\Delta x_i} \left( F_i^{n+\frac{1}{2}}(t) - F_i^{n-\frac{1}{2}}(t) \right).$$

Our goal is to construct explicit methods which are stable for time steps appropriate for the regular part of the grid. This requires a special treatment of the small cell.
we use numerical fluxes of the form
\[ F_{i+\frac{1}{2}}^n = \mathcal{F}(Q_{i-\frac{1}{2}}^n, Q_{i+\frac{1}{2}}^n, Q_{i+1}^n, Q_{i+2}^n), \]
where \( \mathcal{F} \) is a consistent numerical flux function. For the update of the small cell \( k \), we use numerical fluxes of the form
\[ F_{i-\frac{1}{2}}^n = \mathcal{F}(Q_{k-\frac{1}{2}}^n, Q_{k-\frac{3}{2}}^n, \ldots, Q_{k-1}^n), \]
\[ F_{i+\frac{1}{2}}^n = \mathcal{F}(Q_{k+\frac{1}{2}}^n, Q_{k+\frac{3}{2}}^n, \ldots, Q_{k+1}^n), \]
where \( Q_{k-\frac{1}{2}}, Q_{k-\frac{3}{2}}, \ldots, Q_{k-1} \) and \( Q_{k+\frac{1}{2}}, Q_{k+\frac{3}{2}}, \ldots, Q_{k+1} \) are values of the conserved quantity assigned to what we call \( h \)-boxes. These are grid cells of width \( h \), which are constructed to the left and to the right of the small grid cell, i.e. at the interfaces \( k - \frac{1}{2} \) and \( k + \frac{1}{2} \), compare with Figure 1. In [1], we showed that versions of the resulting finite volume method are stable for time steps corresponding to \( CFL \leq 1 \) for the regular part of the grid, independently of the size of the small grid cell.

In order to obtain methods for which the accuracy of the approximation in the small cell is comparable with the accuracy of the regular finite volume method, the computation of the \( h \)-box values, i.e. of the conserved quantities which are assigned to the \( h \)-boxes, must be accurate enough. We compute \( h \)-box values by integrating piecewise polynomial reconstructions of the conserved quantities over the \( h \)-box length. For a second order accurate method for example, we integrate over piecewise linear reconstructions of \( q \).

The \( h \)-box values constructed to update the small cell are computed using
\[ Q_{k-\frac{1}{2}} = Q_{k-\frac{3}{2}}, \quad Q_{k-\frac{3}{2}} = Q_{k-1}^n, \]
\[ Q_{k-\frac{1}{2}} = \alpha Q_{k}^n + \frac{1}{h} \int_{x_{k-\frac{1}{2}}^{+}}^{x_{k+\frac{1}{2}}^{-}} p_{k+1}(x) dx \]
\[ Q_{k-\frac{1}{2}} = \frac{1}{h} \left( \int_{x_{k-\frac{1}{2}}^{+}}^{x_{k+\frac{1}{2}}^{-}} p_{k+1}(x) dx + \int_{x_{k+\frac{1}{2}}^{-}}^{x_{k+\frac{3}{2}}^{-}} p_{k+2}(x) dx \right) \]
and similar formulas for \( Q_{k+\frac{1}{2}}, Q_{k+\frac{3}{2}}, \ldots, Q_{k+1} \). Here \( p_{k+1}(x) \) and \( p_{k+2}(x) \) are polynomial reconstructions of the conserved quantities in grid cells \( k+1 \) and \( k+2 \), respectively. For problems with discontinuous solutions we use limiters in the reconstruction of these polynomials.

For our one-dimensional model problem it is now easy to consider different \( h \)-box methods, which are accurate and have the required stability property. However, we are in particular interested in methods which can easily be extended to the spatially two-dimensional case. In our recent work [2], we have studied \( h \)-box methods using the method of lines and found that they are simpler to implement than \( h \)-box methods, which are based on a MUSCL-type method. For the model problem the update of the small cell has the semi-discrete form
\[ Q'(t) = -\frac{1}{\alpha h} \left( \mathcal{F} \left( Q_{k+\frac{1}{2}}^{-}(t), Q_{k+\frac{1}{2}}^{+}(t) \right) - \mathcal{F} \left( Q_{k-\frac{1}{2}}^{-}(t), Q_{k-\frac{1}{2}}^{+}(t) \right) \right), \]
with
\[ Q_{k+\frac{1}{2}}^{-} = Q_{k+\frac{1}{2},2} + \frac{h}{2} \nabla Q_{k+\frac{1}{2},2}, \quad Q_{k+\frac{1}{2}}^{+} = Q_{k+\frac{1}{2},3} - \frac{h}{2} \nabla Q_{k+\frac{1}{2},3}. \]
Here, \( \nabla Q \) is a gradient assigned to the \( h \)-box. While a limited gradient \( \nabla Q_{k+\frac{1}{2}, 2} \) and \( \nabla Q_{k+\frac{1}{2}, 3} \) can be computed using the four \( h \)-box values \( Q_{k+\frac{1}{2}, 1}, \ldots, Q_{k+\frac{1}{2}, 4} \), we can alternatively compute the gradient as an area weighted average of gradients from the original finite volume grid. With such an approach, we only need to construct two \( h \)-boxes at each cut cell interface. Such an approach simplifies the construction of two-dimensional \( h \)-box methods considerably. Based on this observation, we presented in [2] an \( h \)-box method which is much simpler than our previous method [7].

3. Two dimensional embedded boundary methods using rotated grid \( h \)-box methods. We assume that the embedded boundary is described by a piecewise linear segment that cuts through each boundary cell once. Each irregular grid cell is a polygon with at most five sides, and the finite volume method can be written in the semi-discrete form

\[
Q_{i,j}^t(t) = -\frac{1}{\kappa_{i,j} \Delta x \Delta y} \left( h_{i+\frac{1}{2}, j} F_{i+\frac{1}{2}, j}^t(t) - h_{i-\frac{1}{2}, j} F_{i-\frac{1}{2}, j}^t(t) ight) + h b F_b(t),
\]

where \( h_{i-\frac{1}{2}, j}, h_{i,j-\frac{1}{2}} \) and \( h_b \) denote the length of the cut cell edges and of the boundary segment. For cut cells, some of the edge lengths might vanish. The area of the cut cell with index \((i, j)\) is \( \kappa_{i,j} \Delta x \Delta y \), with \( 0 < \kappa_{i,j} \leq 1 \). To obtain a stable cut cell method, the fluxes at cut cell interfaces are computed using a rotated grid approach, where we compute fluxes in directions \( \xi = (\alpha, \beta)^T \) (tangential to the boundary segment) and \( \eta = (-\beta, \alpha)^T \) (normal to the boundary segment). For a triangular shaped grid cell with index \((i, j)\) of the form indicated in Figure 2, the flux computation has the form

\[
F_{i-\frac{1}{2}, j} = \alpha F_{i-\frac{1}{2}, j}^\xi - \beta F_{i-\frac{1}{2}, j}^\eta \\
F_{i,j+\frac{1}{2}} = \beta F_{i,j+\frac{1}{2}}^\xi + \alpha F_{i,j+\frac{1}{2}}^\eta \\
F_b = F_b^\eta
\]

The fluxes in the rotated directions are obtained using values of the conserved quantities assigned to \( h \)-boxes as indicated in Figure 2

\[
F^\xi = \mathcal{F}(Q_\xi^+, Q_\xi^-), \quad F^\eta = \mathcal{F}(Q_\eta^+, Q_\eta^-), \quad F_b = \mathcal{F}(Q_b^-, Q_b^+),
\]
using reconstructed values of the conserved quantities at both sides of the considered interface. The simplest way to obtain the values $Q_{\xi,\eta}^{\pm}$ is by reconstructing the $h$-box value using a gradient which is obtained as an area weighted gradient from the underlying Cartesian cut cell grid. These cut cell gradients themselves are computed using a standard least square approach. With such an approach we only need two cut cells in each direction of the rotated grid at each cut cell interface, which reduces the stencil size compared to the MUSCL-type approach considered previously.

Note that by using a method of lines approach, we only require left and right values of the conserved quantities at grid cell interfaces at the current time level. In a second order accurate MUSCL-type method we require these values at an intermediate time $t^{n+\frac{1}{2}}$ instead. Let us for instance consider the computation of $Q_{\eta}^{\pm}$ required to compute the flux $F_{i-\frac{1}{2},j}$ in a cut cell of the form considered in Figure 2. Using the method of lines we reconstruct this values using

$$Q_{\eta}^{\pm} = Q_{L,R}^{L,R} \eta^{\pm} \frac{h}{2} \nabla Q_{L,R}^{L,R},$$

(11)

where $h$ is the length of the $h$-box.

In a MUSCL-type approach we instead need a reconstructed value at the half time level, which can be obtained using

$$Q_{\eta}^{\pm} = Q_{L,R}^{L,R} \eta^{\pm} \frac{h}{2} \nabla Q_{L,R}^{L,R} + \Delta t \frac{q}{2}.$$

(12)

Now $q_t$ must be replaced by fluxes in the $\xi$ and the $\eta$-directions using a rotated version of the pde, see [7].

4. New cut cell methods for the approximation of linear problems. In our current work we want to construct cut cell methods which do not rely on the construction of $h$-boxes and in particular not on a rotated grid approximation of cut cells. Here we consider only approximations of smooth solutions of the advection equation.

4.1. The one–dimensional case. On an equidistant one–dimensional grid a fourth order accurate method for the advection equation $q_t + q_x = 0$ is given by

$$Q'_i(t) = -\frac{1}{\Delta x}(Q_{i+\frac{1}{2}}(t) - Q_{i-\frac{1}{2}}(t)),$$

(13)

with

$$Q_{i+\frac{1}{2}}(t) = \frac{7}{12}(Q_i(t) + Q_{i+1}(t)) - \frac{1}{12}(Q_{i-1}(t) + Q_{i+2}(t)),$$

(14)

where $Q_i(t)$ is the cell average of the conserved quantity in grid cell $i$ and $Q_{i+\frac{1}{2}}$ is a fourth order accurate reconstruction of the value of $q$ at the grid cell interface, see [4]. If the resulting system of ordinary differential equations is approximated using a fourth order accurate time stepping scheme (such as the standard RK4 method), then we obtain a fourth order accurate approximation. The method can easily be extended to non-equidistant grids using interface values as also presented in [4]. This irregular version of the method will now be used on a grid with only one small grid cell as shown in Figure 1.

To study the stability of the resulting methods, we write the semi–discrete method for a periodic problem in the form

$$Q'(t) = AQ(t),$$
where \( Q(t) \) is the grid function of the numerical solution at time \( t \). The discrete problem using RK4 can then be written in the form
\[
Q^{n+1} = B(k)Q^n,
\]
where \( Q^n \) is the grid function of the numerical solution at time \( t_n \), \( k \) is the time step, and
\[
B(k) = I + kA + \frac{1}{2}k^2AA + \frac{1}{6}f^3AAAA + \frac{1}{24}k^4AAAA.
\]
The method is Lax-Richtmeyer stable if \( \|B(k)\|_2 \leq 1 \), i.e. if all eigenvalues of \( B(k) \) are inside the unite circle. In Figure 3 we show the spectrum of \( B(k) \) for different versions of the method using 501 grid cells. On a regular grid the method is stable for both cases considered, i.e. CFL=1.6 and CFL=1.7. If we have one small cell with \( \alpha = 10^{-10} \), the method is only stable for CFL=1.6 but not for CFL=1.7. Note that in the case of one small grid cell, there are always two eigenvalues which are a bit separated from the other eigenvalues. By increasing the time steps these two eigenvalues will first trigger the instability. All the CFL numbers were computed with respect to the regular grid cells. Our tests have confirmed that the irregular grid method was stable for \( CFL \leq 1.6 \) and all values of \( \alpha \). Thus, the presence of a small cell requires a smaller time step compared to the regular grid method but this relatively small reduction of the time step is independent of the size of \( \alpha \). Furthermore, note that we need to use at least a third order accurate Runge-Kutta method in order to stabilize the centered spatial discretization. A second order RK method would not be stable, since its stability region does not contain an interval on the imaginary axis.

Now we consider second order accurate spatial discretizations on a grid of the form considered in Figure 1. This serves as a preliminary step for the development of two-dimensional cut cell methods. For the update of the advection equation we now use a flux of the form \( F_{i+\frac{1}{2}}(t) = Q_{i+\frac{1}{2}}(t) \), with
\[
Q_{i+\frac{1}{2}}(t) = (Q_i(t) + Q_{i+1}(t))/2 \quad \text{for} \ i \neq k - 1, k
\]
\[
Q_{k-\frac{1}{2}} = (\alpha Q_{k-1} + Q_k)/(1 + \alpha)
\]
\[
Q_{k+\frac{1}{2}} = (\alpha Q_{k+1} + Q_k)/(1 + \alpha).
\]
In Figure 4 we plot the spectrum of the matrix \( B(k) \) for the resulting method using RK4 as time stepping scheme. As in the case of the fourth order accurate
spatial discretization, we observe that the small grid cell reduces the admissible CFL number compared to the regular grid case. However, this reduction of the time step is relatively small and independent of $\alpha$ as $\alpha \to 0$. Thus, it seems that these centered discretizations do not suffer from the small cell problem in the same way as an upwind discretization does.

### 4.2. The two–dimensional case.

For smooth solutions of the two-dimensional advection equation

$$q_t + q_x + q_y = 0$$

we consider a Cartesian grid method of the form

$$Q_{i,j}'(t) = -\frac{1}{\Delta x_i} \left( F_{i+\frac{1}{2},j}(t) - F_{i-\frac{1}{2},j}(t) \right) - \frac{1}{\Delta y_j} \left( F_{i,j+\frac{1}{2}}(t) - F_{i,j-\frac{1}{2}}(t) \right),$$

(17)

with fluxes of the form

$$F_{i+\frac{1}{2},j}(t) = \Delta y_j Q_{i+\frac{1}{2},j}(t), \quad F_{i,j+\frac{1}{2}}(t) = \Delta x_i Q_{i,j+\frac{1}{2}}(t)$$

(18)

and with $Q_{i+\frac{1}{2},j}(t)$, $Q_{i,j+\frac{1}{2}}(t)$ computed using face neighbors analogously to the one-dimensional method in (15). First we consider the stability of the two-dimensional method on a grid which is obtained as the tensor product of two one dimensional grids with one small grid cell. Such a grid has one row and one column with grid cells of size $\alpha \Delta x \Delta y$ and one grid cell of size $\alpha^2 \Delta x \Delta y$. Here we assume $\Delta x = \Delta y$.

In Figure 5 we show the spectrum of the resulting two-dimensional second order accurate method using RK4. As in the one-dimensional case, we observe that the presence of the small cells reduces the stability. However, the reduction is again independent of $\alpha$ as $\alpha \to 0$.

Finally we use such a centered spatial discretization for a cut cell method in a simple configuration. The update of cut cells has the form (8) with $F_b(t) = 0$. For cut cell interfaces the flux is computed using linear interpolation of the conserved quantity evaluated at the midpoint of the interface. For a triangular shaped cut cell we use the same three cell average values to compute the linear reconstructed polynomial. For this we use the value from the small triangular shaped cell as well as the value from the two face neighbors. For the temporal discretization we use the classical RK4 method. In Figure 6 we show the numerical result for the propagation of a smooth wave in a channel. In this computation the smallest grid cell was about $7.5 \cdot 10^{-5}$ times the size of a regular cell. This did not cause accuracy or stability issues.

Figure 4. Spectrum of $B(k)$ for different versions of the second order accurate method (15) using 501 grid cells. For the two left plots CFL=1.9 and for the two right plots CFL=2.05 was used. In both cases we show results for the regular grid method and for the method on a grid with one small cell with $\alpha = 10^{-10}$. 

4.2. The two–dimensional case. For smooth solutions of the two-dimensional advection equation
Figure 5. Spectrum of $B(k)$ for different versions of the two-dimensional second order accurate method using $51 \times 51$ grid cells. For the two left plots $CFL=0.95$ and for the two right plots $CFL=1.01$ was used. In both cases we show results for the regular grid method and for the method on a grid with one row and one column of small cells with $\alpha = 10^{-10}$.

Figure 6. Numerical results for the propagation of a smooth wave in a channel. The smallest cut cell was $7.5 \times 10^{-5}$ times the size of a regular cell, $CFL = 0.95$ corresponding to the regular grid cells.

problems for time steps suggested by the eigenvalue analysis presented in Figure 5. A numerical convergence study confirms second order accuracy for this approach. If we replace the regular cell update away from the embedded boundary by the fourth order flux (14), then the resulting method is third order accurate (as expected with a second order boundary treatment).

In the future we plan to further investigate the stability of such cut cell methods for linear problems in more complicated cut cell configurations.

REFERENCES


E-mail address: christiane.helzel@rub.de
E-mail address: berger@cims.nyu.edu
TIME AND SPACE DISCRETE SCHEME TO SUPPRESS NUMERICAL SOLUTION OSCILLATION FOR THE NEUTRON TRANSPORT EQUATIONS

ZHENYING HONG, GUANGWEI YUAN AND XUEDONG FU

Institute of Applied Physics and Computational Mathematics
Department 6, NO.2. Fenghaidong Road, Haidian District
Beijing, 100094, China

Abstract. In this paper, the numerical solution oscillation for time-dependent neutron transport equations which are hyperbolic partial differential equations is investigated. The influence of time discrete scheme and space discrete scheme on this oscillating phenomenon is analyzed for neutron transport equations. The second-order time evolution scheme for time variable is very suit to adaptive time step problem and suppresses the numerical solution oscillation. Moreover, the linear discontinuous finite element method for space variable is an asymptotic preserving scheme and restricts the oscillation problem which exits in exponential method and diamond difference.

1. Introduction. The neutron transport equations which are hyperbolic partial differential equations used in radiation shielding and nuclear reaction system, as well as medicine realm, are linearized version of the equation originally developed by Boltzmann for the kinetic theory of gases. Up to now attention has been focused on the time-independent particle transport problems. But, with the increase of practical requirements, the physics problems discussed now become more sophisticated than before, for example, complex structure, non-uniform medium, anisotropic property, broad-energy region, sophisticated physical condition, movement station etc. Therefore, it is essentially necessary to study more advanced numerical method, and pay more attention on time-dependent transport equations. Moreover, the discretization methods for angular variable should be studied for curvilinear geometry transport equations (Lathrop, 2000).

Moreover, a good numerical method should rationally simulate physical properties of mathematical models, and do not introduce non-physical deviation. One of the focus for solving nuclear radiation problems is to obtain accurate estimates of certain physical quantities at some spatial points within the physical system considered. In particular, for photonics problems accurate estimates of the exiting partial current on the outer boundary of the system are required.

2000 Mathematics Subject Classification. Primary: 65M08; Secondary: 82D75.

Key words and phrases. Transport equations, numerical solution oscillation, second-order time evolution scheme, linear discontinuous finite element method, asymptotic preserving.

Project 2011B0202039 supported by CAEP, Supported by the National Natural Science Foundation(11126135, 11201034), the Defense Industrial Technology Development Program B1520110011 and the National Basic Research Program under the Grant 2011CB309705.
The exponential method and diamond difference method (FVM) are typical discrete schemes for finite volume method to solving transport equations. The diamond difference method is simple and it is easy to be applied to solve transport equations. However, a disadvantage of the DD method is that it can produce non-physical oscillating solutions (Klose and Larsen, 2006). The exponential method is more sophisticated than diamond difference method and is a nonlinear, positive numerical scheme.

But we focus on differential property of flux with respect to time in different numerical schemes. Discontinuous finite element methods have been applied to transport equations in a variety of different settings (Wareing and et al, 2001; Adams, 2001; Morel and Wareing, 1996; Machorro, 2007), and we hope to choose a suitable numerical method with higher order accuracy. We consider partial variable finite element method, namely apply linear discontinuous finite element method (LD) to spatial variable and diamond difference method to other variables. As for angular variable, we apply the SN method.

In this paper, we construct time discrete scheme for multi-media complex time-dependent progress and apply second-order time evolution (SOTE) to discrete ordinates equation for one-dimensional spherical geometry particle transport equation. To space discrete scheme, we consider firstly one kind of particular transport equations with analytic solution and then spherical geometry transport equations. Moreover, we analyze the oscillating phenomenon about numerical solution. The remainder of this paper is organized as follows. In Section 2, we introduce the time-dependent neutron transport equations. In Section 3, the typical time discrete scheme is presented. In Section 4, we construct the second-order time evolution scheme. In Section 5, we apply space linear discontinuous finite element to time-dependent problem. In Section 6, we provide numerical results for neutron transport equation to study the time and space discrete scheme of effect for suppress numerical solution oscillation. In the final section, we offer a summary with some concluding remarks.

2. Time-dependent neutron transport equations. The time-dependent particle transport equation may be written as follows in multi-group form:

\[ \frac{1}{v_g} \frac{\partial \varphi_g}{\partial t} + \Omega \cdot \nabla \varphi_g + \Sigma_{tr} \varphi_g = Q_g, \quad g = 1, \ldots, G. \]

where \( \varphi_g \) is the angular flux of \( g \)'th group particle, \( v_g \) is the velocity of \( g \)'th group particle, \( \Sigma_{tr} \) is the total macroscopic cross section of \( g \)'th group particle, and \( Q_g \) is the total source (including scatter source \( Q_s^g \), fission source \( Q_f^g \) and external source \( S_g \)).

In this paper, we focus on conservative equation for 1-D spherical geometry transport equations in the multi-group form:

\[ \frac{1}{v_g} \frac{\partial \varphi_g}{\partial t} + \frac{\mu}{r^2} \frac{\partial (r^2 \varphi_g)}{\partial r} + \frac{1}{r} \frac{\partial \left[ (1-\mu^2) \varphi_g \right]}{\partial \mu} + \Sigma_{tr} \varphi_g = Q_g. \]

With the following initial and boundary conditions:

\[ \varphi_g(r, \mu, 0) = \varphi^{(0)}_g(r, \mu), \quad t = 0, \quad 0 \leq r \leq r_J. \]

\[ \varphi_g(r_J, \mu, t) = 0, \quad \mu \leq 0. \]

Where \( r_J \) is the outermost boundary point.
There exist numerical solution oscillation for typical discrete scheme when solving multi-group multi-media sophisticated time-dependent neutron transport equations which brings difficulty for mathematics and physics analysis. Especially for pivotal physical quantity, we cannot take the key physical progress for the numerical solution oscillation. Therefore, we will study the time discrete scheme and space discrete scheme for sophisticated problem. The motivation of our research is to construct robust numerical scheme to provide reasonable physical quantity cures and to provide means for simulating the neutron transport equations.

3. **Time discrete scheme.** The time step can be large at flat stage and be small at strenuous stage for physical progress. Therefore, the adaptive time step is adopted in numerical calculation for practical physical problem.

If taking exponential extrapolation or diamond extrapolation for time variable, the extrapolation flux can exit deviation when the time step has great change. The modified time discrete scheme can be adopted (Hong, Yuan and Fu, Yang, 2010).

To consider the time step change in the whole physical progress adequately, we apply the second-order time evolution (SOTE) scheme to time-dependent spherical geometry particle transport equation by discrete ordinates (Sn) method. The SOTE consider the case of adaptive time step for the whole physical progress and need not to introduce exponential extrapolation or diamond extrapolation for time variable. The SOTE take three-level backward difference and the specific equation is as followed (Olson, 2007; Olson, 2009).

We should introduce the extrapolation relation for angular variable and space variable. In this paper, we introduce the exponential extrapolation and diamond extrapolation to angular and space variable which are named SOTE.EM, SOTE.DD respectively.

Through rearrangement, we get (SOTE.EM):

\[
a \left( \frac{\varphi^{n+1}}{g_{m,k}+\frac{1}{2}} \right)^2 + b \frac{\varphi^{n+1}}{g_{m,k}+\frac{1}{2}} + c = 0.
\]

The parameters for a, b, c are described by different case (\(-1 < \mu_m < 0, \mu_m > 0\) and \(\mu_m = -1\)). The derivation for SOTE.DD is similar to SOTE.EM.

4. **Space Linear Discontinuous Finite Element Method.** We will study the phenomenon of oscillation of numerical solution, and the performance of different numerical scheme to transport equations with a small perturbation. To make statement clear, we consider at first the following particular transport problems:

\[
\frac{\partial \varphi}{\partial t} + \frac{\partial \varphi}{\partial x} = \alpha \varphi \quad x \in [0, 1], \ t > 0
\]

\[
\varphi(x, 0) = f(x) \quad x \geq 0
\]

\[
\varphi(0, t) = g(t) \quad t \geq 0
\]

where \(\alpha\) is a given constant, \(f(x)\) and \(g(t)\) are given functions satisfying consistency condition \(f(0) = g(0)\).

The analytic solution of the above problem is:

\[
\varphi(x, t) = \begin{cases} 
e^{\alpha t}f(x - t) & x \geq t \\ e^{\alpha x}g(t - x) & x \leq t \end{cases}
\]

Let the initial condition be continuous, but the boundary condition be discontinuous:

\[
f(x) = e^x,
\]
angular flux of cell center is written as boundary for adjacent cell. (Yuan, 2009).

Through similar derivation, the difference scheme of EM and DD method for the 1-D spherical geometry transport problem Eq.2–4. Through the expansion of basis function of finite element for angular flux, the linear discontinuous finite element method is implied to spatial variable. Different basis function and weighted function will give different specific discrete form. We will take square method in this paper.

The basis function of finite element is as follows:

\[ f_k(x) = \begin{cases} \frac{x_{k+1} - x}{\Delta x_k}, & x_k \leq x \leq x_{k+1} \\ 0, & \text{else} \end{cases} \]  

\[ f_{k+1}(x) = \begin{cases} \frac{x - x_k}{\Delta x_k}, & x_k \leq x \leq x_{k+1} \\ 0, & \text{else} \end{cases} \]  

The weighted function is given as follows:

\[ w_1(x) = 1, \quad w_2(x) = \begin{cases} x - x_k, & \mu_m > 0 \\ x_{k+1} - x, & \mu_m < 0. \end{cases} \]  

For Eq.6, the diamond difference method is implied to time variable and the linear discontinuous finite element method is implied to spatial variable. By a direct derivation, the system of linear algebraic equations of flux about cell-edge are given as follows:

\[
\begin{align*}
\frac{(2V_{k+1}^{n+1})}{\Delta t} - \alpha V_{k}^{n+1} & \varphi_{k}^{n+1} + \frac{(2V_{k+1}^{n+1})}{\Delta t} + s_1 - \alpha V_{k+1}^{n+1} \varphi_{k+1}^{n+1} \\
= & \frac{2}{\Delta t} (\varphi_{k}^{n+1} + \varphi_{k+1}^{n+1}) + s_1 \varphi_{k}^{(l),n+1}
\end{align*}
\]

\[
\begin{align*}
\frac{2s_3}{\Delta t} + s_4 - s_3 \alpha \varphi_{k}^{n+1} + \frac{2s_3}{\Delta t} - s_4 - s_2 \alpha \varphi_{k+1}^{n+1} & = \frac{2}{\Delta t} (s_3 \varphi_{k}^{n+1} + s_2 \varphi_{k+1}^{n+1})
\end{align*}
\]

where: \( V_{k}^{n+1} = \frac{1}{12} x_k^3 + \frac{1}{4} x_k^4 - \frac{1}{6} x_k^3 x_{k+1}, V_{k+1}^{n+1} = \frac{1}{2} x_k^4 + \frac{1}{12} x_k^4 + \frac{1}{12} x_k x_{k+1}, s_1 = \frac{1}{3} x_k^3 + \frac{1}{3} x_k^3, s_2 = \frac{1}{20} x_k^5 - \frac{1}{20} x_k^5 - \frac{1}{12} x_k^4 + \frac{1}{12} x_k^4 x_{k+1}, s_3 = \frac{1}{30} x_k^5 + \frac{1}{5} x_k^5 - \frac{1}{3} x_k^3 x_{k+1} + \frac{1}{3} x_k x_{k+1}, s_4 = -\frac{1}{3} x_k^3 x_{k+1} + \frac{1}{4} x_k^4 + \frac{1}{12} x_k^4 x_{k+1} \varphi_{k}^{(l),n+1} \) is the flux of exit boundary for adjacent cell.

Through the expansion of basis function of finite element for angular flux, the angular flux of cell center is written as:

\[
\varphi_{k+\frac{1}{2}}^{n+1}(x) = \frac{\varphi_{k+1}^{n+1} + \varphi_{k+1}^{n+1}}{2}
\]

Now let’s come back to the 1-D spherical geometry transport problem Eq.2–4. Through similar derivation, the difference scheme of EM and DD method for time-dependent spherical geometry transport equations can be obtained (Hong and Yuan, 2009).

The convergence criterion employed in this work was

\[
\max_{g,k,m} \left| 1 - \frac{\varphi_{g,k,m}^{(l+1)}}{\varphi_{g,k,m}^{(l)}} \right| < \varepsilon
\]

where \( l + 1, l \) denote the present and the previous iterations, respectively. The prescribed tolerance \( \varepsilon \) is set to be \( 10^{-7} \).
5. Numerical results. We define the undermentioned physical quantity to describe different differential curve for time variable. \[ J = \int_{\Omega, n > 0} d\Omega \varphi(r_J, \Omega, E, t) \]

This physical quantity \( J \) gives the information about outflux at outermost boundary, which denotes the outflux current of system particle. The influx is zero for taking void boundary condition. \( \lambda_{\text{edge}} = \frac{dJ}{dt} \), denotes the derivative of outflux current (\( J \)).

The results presented and discussed in this section are organized into time discrete scheme and space discrete scheme.

5.1. Tests for time discrete scheme. The problem discussed in this paper is about spherical geometry multi-group time-dependent problem including two media. The isotropic scattering source is employed. The discrete angular takes \( S_4 \) and the end time is 0.1 \( \mu s \). The self adaptive time step is showed in table 1. We adopt the typical \( EM, DD \) and the modified time discrete scheme and second-order time evolution scheme. To study the computing effectiveness, we also take constant time step \( (10^{-4} \mu s) \) (\( EM \)) to this problem.

<table>
<thead>
<tr>
<th>Table 1. Self adaptive time step(( \mu s ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>time interval</td>
</tr>
<tr>
<td>time step</td>
</tr>
</tbody>
</table>

**Figure 1.** \( \lambda_{\text{edge}} \) for \( EM, MEM \) and \( SOTE\_EM \)

**Figure 2.** Iterative number for \( EM, MEM \) and \( SOTE\_EM \)
5.2. Test for space discrete scheme. The results presented and discussed in this section are organized into three subsections, according to the form of transport equation. Firstly, we analyze the time-independent transport equation. Secondly, a kind of particular transport equation with a small perturbation is studied. Thirdly, 1-D spherical geometry multi-group time-dependent transport equation is studied, and anisotropic scattering source with $P_5$ spherical harmonic expansion is considered.

5.2.1. Transport problems with a small perturbation. The first problem has analytic solution as mentioned in part 2. We apply $EM, DD, LD$ to this problem to investigate the computational performance with a small perturbation.

The radius is $0.5cm$, and boundary condition is

$$g(t) = e^{-t} + \varepsilon \delta(t - t_0).$$

The value of small perturbation is $10^{-3}$. The end time is $1.0\mu s$. The time step is $10^{-3}\mu s$ and the space step is $10^{-3}cm$. The $\alpha$ is $10$ and $t_0$ is $0.3\mu s$. We research the physical quantity $\lambda_{edge}$ which describes the differential curve of time variable for outflux at boundary. The numerical results for $\lambda_{edge}$ are showed in Fig.3.

![Figure 3. $\lambda_{edge}$ by different numerical scheme](image)

Given a small perturbation, the numerical solution of $LD$ method is approximate to the analytic solution, but the numerical solutions of $EM$ and $LD$ method have major difference to the analytic solution. According to $\lambda_{edge}$, the $EM$ shows violent oscillatory motion and the $DD$ method has inferior oscillatory motion. However, the solution of $LD$ method has no oscillatory motion and the curve of $\lambda_{edge}$ is very smooth.

5.2.2. Spherical geometry multi-group time-dependent transport problem. This test is about spherical geometry multi-group time-dependent problem including two media. The four-group cross sections are considered and the anisotropic scattering source with $P_5$ is employed. There is no analytic solution for this problem, therefore the numerical solution of exponential method by fine cell($S_{16}$, $\Delta x = 0.1cm$, $\Delta t = 5 \times 10^{-5}\mu s$) is used by reference solution. The solutions of coarse cell($S_4$, $\Delta x = 0.5cm$, $\Delta t = 2 \times 10^{-5}\mu s$) is used by reference solution.
max $\Delta x = 0.97\text{cm}$, $\Delta t = 2 \times 10^{-4}\text{µs}$) for different scheme are contrasted with that of fine cell. The solution of $\lambda_{edge}$ is showed in Fig.4,

![Figure 4. $\lambda_{edge}$ for mutigroup problem](image)

![Figure 5. iteration number for mutigroup problem](image)

Therefore, to non-uniform medium time-dependent neutron transport problems, we can take $LD$ method, which can give more accurate numerical solution for coarse cells. Also the differential curve about time variable of $LD$ method is very smooth, which avoid the violent oscillatory motion of $EM$ and $DD$ method. The weakness of $LD$ is that the computing time is larger than that of $EM,DD$.

6. **Summary.** We study the numerical solution oscillation from the aspect of time discrete and space discrete scheme. The second-order time evolution scheme associated exponential method($SOTE\_EM$) has some good properties. The differential curve including $\lambda_{edge}$ on time about particle current is more smoother than that of exponential method and diamond difference and modified time discrete scheme.

In conclusion, the $LD$ method yields more accurate results, especially for the flux on edge of cell, and can reduce the oscillatory phenomenon effectively. Therefore the $LD$ method can provide accurate numerical solutions for time-dependent neutron transport equations.

The shortcoming of $SOTE\_EM$ and $LD$ is that the iterative number is more than other schemes and we will take acceleration method such as taking effective iterative initial value to decrease the iterative number.
REFERENCES


E-mail address: zyhong@iapcm.ac.cn
E-mail address: yuan_guangwei@iapcm.ac.cn
E-mail address: xdfu@iapcm.ac.cn
NUMERICAL SIMULATION OF A HYPERBOLIC MODEL FOR CHEMOTAXIS AFTER BLOW UP

FRANÇOIS JAMES
Mathématiques – Analyse, Probabilités, Modélisation – Orléans (MAPMO)
Université d’Orléans & CNRS UMR 6628
Fédération Denis Poisson, Université d’Orléans & CNRS FR 2964
45067 Orléans Cedex 2, France

NICOLAS VAUCHELET
UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions
CNRS, UMR 7598
Laboratoire Jacques-Louis Lions and INRIA Paris-Rocquencourt
EPI BANG, F-75005, Paris, France

ABSTRACT. A macroscopic system modelling the chemotactic motion of bacteria is considered. This model has been obtained in a previous work as the hydrodynamic limit of a kinetic system. Existence and uniqueness of measure solutions for this system using the concept of duality solution have been proved by the authors in [7]. In this paper, we investigate the numerical discretization of this system. A scheme based on a finite volume approach is proposed and the convergence of the numerical solution towards the unique duality solution is stated. Numerical simulations are provided that shows the behaviour of solutions after blow up.

1. Introduction. Chemotaxis is the phenomenon in which a population of cells rearranges its structure according to the behaviour of some chemical present in the environment. In this work, we focus on the following macroscopic model for chemotaxis:

\[ \partial_t \rho + \partial_x (a(\partial_x S)\rho) = 0, \]
\[ a(\partial_x S) = -c \phi(c \partial_x S), \]
\[ -\partial_{xx} S + S = \rho. \]

In this system, \( \rho(t, x) \geq 0 \) denotes the density of cells at time \( t \geq 0 \) and position \( x \in \mathbb{R} \), and \( S(t, x) \) is the concentration of the chemical, which is here produced by the bacteria at a rate proportional to their density, and diffuses in the system. The positive constant \( c \) corresponds to the individual velocity of cells, whereas the real-valued function \( \phi \) describes the influence of the chemical concentration \( S \) on the global motion of bacteria. The quantity \( a(\partial_x S) \) is the macroscopic velocity. When \( \phi \) is nonincreasing, hence \( a \) nondecreasing, cells attract each other, we are in the case of positive chemotaxis, the chemical is then called the chemoattractant.

2000 Mathematics Subject Classification. Primary: 35Q92, 35L60, 65M08; Secondary: 35D30.

Key words and phrases. Chemotaxis, non-local conservation equations, numerical analysis, finite volume scheme, measure-valued solutions.
otherwise, we are in the repulsive case and speak about the chemorepellant. We complement this system with the boundary conditions
\[ \rho(t = 0, x) = \rho_{ini}(x), \quad \lim_{x \to \pm \infty} \rho(t, x) = 0, \quad \lim_{x \to \pm \infty} S(t, x) = 0. \] (4)
Introducing the elementary interaction kernel \( K \) solving \(-\partial_{xx}K + K = \delta_0\), i.e. \( K = \frac{1}{2} e^{-|x|} \), the latter system reduces to the nonlocal scalar conservation law
\[ \partial_t \rho + \partial_x (a(\partial_x K \ast \rho)) \rho = 0. \] (5)
When \( a \) is nondecreasing, this equation is known as the aggregation equation. It is now classical that regular solutions blow up in finite time when \( K \) is not smooth (see e.g. [1]). Thus measure solutions have to be considered, together with a suitable definition of the product \( a(\partial_x K \ast \rho) \rho \), when \( \rho \) is a measure.
System (1)–(3) can be obtained thanks to a hydrodynamic limit of the so-called Othmer-Dunbar-Alt model. This latter system has been introduced to take care of the run and tumble process: the motion of cells is due to the alternance of a swim phase in a given direction (run) and a reorientation phase during which cells take a new direction for the next run (tumble). Denoting \( f(t, x, v) \) the distribution function at time \( t \), position \( x \in \mathbb{R} \) and velocity \( v \in \{-c, c\} \), the kinetic system of equations writes in one space dimension (see e.g. [5])
\[ \partial_t f_\varepsilon + v \partial_x f_\varepsilon = \frac{1}{\varepsilon} ((1 + \phi(\varepsilon v \partial_x S_\varepsilon)) f_\varepsilon(-v) - (1 + \phi(\varepsilon v \partial_x S_\varepsilon)) f_\varepsilon(v)), \] (6)
\[ -\partial_{xx} S_\varepsilon + S_\varepsilon = \rho_\varepsilon = f_\varepsilon(v) + f_\varepsilon(-v). \] (7)
The real constant \( c \) represents the velocity of cells. The left-hand side describes the run phase whereas the right-hand side models the tumble phase. The operator \( T[S] := \frac{1}{2} (1 + \phi(\varepsilon v \partial_x S)) \) is called the turning rate and corresponds to the rate of reorientation of cells. Obviously we must have \( 0 \leq T[S] \leq 1 \). The parameter \( \varepsilon \) is a scaling factor which is assumed here to be very small (\( \varepsilon \ll 1 \)). This corresponds to the phenomenon of dominant taxis. Other scalings are possible, that lead eventually to drift-diffusion equations such as the classical Keller-Segel model.
The hydrodynamic limit consists in letting \( \varepsilon \to 0 \), and it is easily checked that formally the solution \((\rho_\varepsilon, S_\varepsilon)\) of the kinetic system converges to some solution \((\rho, S)\) of system (1)–(3). A complete proof of this result is given in [7], together with a global-in-time existence and uniqueness of duality solutions for system (1)–(3). The main drawback of this method is that, since duality solutions are presently defined only in one space dimension, we are limited to \( x \in \mathbb{R} \). On the other hand, existence of measure-valued solutions for the aggregation equation (5) has been obtained in [4] in any space dimension. The authors make use of optimal transport technique: in this geometric approach, the solution appears as a gradient flow for the interaction energy. However, this latter tool is not so convenient to build numerical schemes and to obtain numerical simulations.
The aim of this paper is therefore to propose a numerical scheme based on a finite volume approach to get numerical simulations of solutions to the macroscopic model (1)–(3). The main ingredient is a proper definition of the flux \( J = a(\partial_x S) \rho \).

The outline is the following. In the next Section we recall some definitions and the existence result stated in [7]. In Section 3 we present the scheme and state some of its main properties in the attractive case. Finally Section 4 provides numerical simulations, both in the attractive and the repulsive case.
2. Existence of duality solutions. Duality solutions have been introduced in [2] to solve scalar conservation laws with discontinuous coefficients. More precisely, it gives sense to measure valued solutions of the scalar conservation law

$$\partial_t \rho(t, x) + \partial_x (b(t, x)\rho(t, x)) = 0,$$

where $b \in L^\infty((0, T) \times \mathbb{R})$ satisfies the so-called one-sided Lipschitz (OSL) condition

$$\partial_x b(t, \cdot) \leq \beta(t) \quad \text{for } \beta \in L^1(0, T),$$

in the distributional sense. (8)

We refer to [2] for the precise definition and general properties of these solutions. Let us denote by $\mathcal{M}_b(\mathbb{R})$ the set of bounded Radon measures and by $\mathbf{P}_1(\mathbb{R})$ the set of nonnegative measures in $\mathcal{M}_b(\mathbb{R})$ with finite first moment, that is $\int |x| d\mu(x) < \infty$. In this framework we define duality solutions of the studied system in the spirit of [3]:

**Definition 2.1.** We say that $(\rho, S) \in C([0, T]; \mathcal{M}_b(\mathbb{R})) \times C([0, T]; W^{1, \infty}(\mathbb{R}))$ is a duality solution to (1)–(3) if there exists $b \in L^\infty((0, T) \times \mathbb{R})$ and $\alpha \in L^{1}_{loc}(0, T)$ satisfying $\partial_x b \leq \alpha$ in $D'(\mathbb{R})$, such that

(i) for all $0 < t_1 < t_2 < T$

$$\partial_t \rho + \partial_x (b \rho) = 0 \quad \text{in the sense of duality on } |t_1, t_2|,$$

(ii) equation (2) is satisfied in the weak sense:

$$\forall \psi \in C^1(\mathbb{R}), \forall t \in [0, T], \quad \int_{\mathbb{R}} (\partial_x S \partial_x \psi + S \psi)(t, x) \, dx = \int_{\mathbb{R}} \psi(x) \rho(t, dx),$$

(iii)

$$b = a(\partial_x S) \quad \text{a.e.}$$

The OSL estimate suggests that the velocity field $a(\partial_x S)$ has to be compressive, which naturally corresponds to positive chemotaxis. Indeed we have a natural one-sided estimate on the potential $S$: from (3), when $\rho$ is nonnegative, we get $\partial_{xx} S \leq S$. This estimate can be considered as an entropy estimate for the scalar conservation law (1) which is crucial for the proof of uniqueness of solutions (see Theorem 5.1 of [7]). If we assume to be in the case of positive chemotaxis, $a$ is nondecreasing, and a straightforward computation leads to the OSL estimate. More precisely we assume

$$a \in C^1(\mathbb{R}), \quad 0 \leq a' \in L^\infty(\mathbb{R}), \quad \forall x \in \mathbb{R} \quad |a(x)| \leq c. \quad (9)$$

The latter estimate means that the collective displacement of cells should not be faster than the velocity of each individual cell; it is a direct consequence of the fact that the turning rate $T[S]$ satisfies $0 \leq T[S] \leq 1$.

From now on, we will denote by $A$ the antiderivative of $a$ which vanishes at 0. We have from the chain rule that, when it is defined, the product

$$a(\partial_x S) \rho = a(\partial_x S) (-\partial_{xx} S + S) = -\partial_x (A(\partial_x S)) + a(\partial_x S) S.$$

Then a natural definition of the flux is given by

$$J = -\partial_x (A(\partial_x S)) + a(\partial_x S) S. \quad (10)$$

We are now in position to state the existence and uniqueness result of [7].

**Theorem 2.2.** Let us assume that $\rho^{ini}$ is given in $\mathbf{P}_1(\mathbb{R})$ and that (9) is satisfied. Then, for all $T > 0$ there exists a unique duality solution $(\rho, S)$ with $0 \leq \rho \in \mathbf{P}_1(\mathbb{R})$ of (1)–(3) which satisfies in the distributional sense:

$$\partial_t \rho + \partial_x J = 0, \quad (11)$$
where \( J \) is the flux defined in (10). Moreover, there exists a universal representative, denoted \( \tilde{a} \), such that \( \tilde{a} = a(\partial_x S) \) a.e. and
\[
\tilde{a} \rho = J, \quad \text{in the sense of measures.}
\]
Then, we have \( \rho = X_# \rho_{\text{ini}} \) where \( X \) is the backward flow corresponding to \( a(\partial_x S) \).

**Remark 1.** This theorem is proved in [7] with a weaker assumption than (9). However, this stronger assumption is needed here for the numerical analysis.

### 3. Numerical scheme.

#### 3.1. Discretization.

Let us consider a uniform space discretization with step \( \delta x \) and denote by \( \delta t \) the time step, and set \( \lambda = \delta t / \delta x \). Then \( t^n = n \delta t \) and \( x_i = x_0 + i \delta x \). We assume that \( (\rho^n_i)_{0 \leq i \leq N} \) is an approximation of \( (\rho(t_n, x_i))_{0 \leq i \leq N} \). We obtain an approximation \( \rho^{n+1}_i \) of \( \rho(t_{n+1}, x_i) \) by using the following Lax-Friedrichs discretization of equations (10)–(11):
\[
\rho^{n+1}_i = \rho^n_i (1 - \lambda c) + \frac{\lambda}{2} (\rho^n_{i-1} + \rho^n_{i+1}) + \frac{\lambda}{2} (J^n_{i-1/2} - J^n_{i+1/2}), \tag{12}
\]
\[
J^n_{i+1/2} = -\frac{A(\partial_x S^n_{i+1}) - A(\partial_x S^n_{i})}{\delta x} + a^n_{i+1/2} \frac{S^n_{i+1} + S^n_{i}}{2}, \tag{13}
\]
where we recall that \( c \) denotes the constant modulus of the velocity of cells. The velocity \( a(\partial_x S) \) is discretized as
\[
a^n_{i+1/2} = \begin{cases} 
0 & \text{if } \partial_x S^n_{i+1} = \partial_x S^n_{i}, \\
\frac{A(\partial_x S^n_{i+1}) - A(\partial_x S^n_{i})}{\partial_x S^n_{i+1} - \partial_x S^n_{i}} & \text{otherwise},
\end{cases} \tag{14}
\]
complemented with the standard centered finite difference \( \partial_x S^n_{i+1} = \frac{S^n_{i+2} - S^n_{i}}{2 \delta x} \).

We couple this equation with the following standard discretization of the equation for the chemoattractant (3)
\[
-\frac{S^n_{i+1} - 2S^n_{i} + S^n_{i-1}}{\delta x^2} + S^n_{i} = \rho^n_i. \tag{15}
\]
In order to avoid the treatment of boundary conditions, we assume that the solutions are compactly supported in the computational domain. Then from now on, we have that \( \rho^n_0 = S^n_0 = S^n_i = J^n_{i-1/2} = 0 \) and \( \rho^n_N = S^n_N = J^n_{N+1/2} = 0 \).

#### 3.2. Numerical analysis.

Before stating and proving our convergence result, we start by a Lemma which proves a CFL-like condition for the scheme:

**Lemma 3.1.** Assume that (9) holds and that the CFL condition
\[
\lambda := \frac{\delta t}{\delta x} \leq \frac{2}{3c}, \tag{16}
\]
is satisfied. Then the scheme defined in (12)–(15) is nonnegative.

**Proof.** Assume that \( \rho^n_i \geq 0 \). Using (14)–(15), we can rewrite (13) as
\[
J^n_{i+1/2} = a^n_{i+1/2} \frac{\rho^n_{i+1} + \rho^n_{i}}{2}. \tag{17}
\]
Thus we can rewrite (12)

$$\rho_{i+1}^n = \rho_i^n (1 - \lambda c + \frac{\lambda}{4} (a_i^{n-1/2} - a_i^{n+1/2})) + \frac{\lambda}{2} \left( c + \frac{a_i^{n-1/2}}{2} \right) \rho_i^n + \frac{\lambda}{2} \left( c - \frac{a_i^{n+1/2}}{2} \right) \rho_i^+.$$

Moreover, by assumption (9) we have $|a_i^{n+1/2}| \leq c$ for all $i = 0, \ldots, N$. Therefore, if $\lambda \leq \frac{2c}{a_i^{n+1/2}}$, all the coefficients in front of $\rho_i^n$, $\rho_i^n$ and $\rho_i^{n+1}$ are nonnegative. We conclude that the scheme is nonnegative.

Let us introduce piecewise constant functions

$$\rho_\delta(t, x) = \sum_{n \in \mathbb{N}} \sum_{i=0}^N \rho_i^n \mathbf{1}_{[n \delta t, (n+1) \delta t) \times [x_i, x_{i+1})}(t, x),$$

$S_\delta$, $\partial_x S_\delta$, $J_\delta$ and $a_\delta$, which are defined in a similar way using $(S_i^n)_i$, $(\partial_x S_i^n)_i$, $(J_{i+1/2}^n)_i$, and $(a_{i+1/2}^n)_i$.

**Theorem 3.2.** Assume we are given $\rho^{i+1} \in P_1(\mathbb{R})$ and define $\rho_i^0 = \mathbf{1}_{x_i} \rho^{i+1}(dx) \geq 0$. Under assumption (9), if (16) is satisfied, then the discretization $(\rho_\delta, S_\delta)$ converges towards the solution $(\rho, S)$ of Theorem 2.2 as $\delta t$ and $\delta x$ go to 0.

**Proof.** We first notice that since the matrix of the linear system (15) is a $M$–matrix, if $(\rho_i^n)_i$ is nonnegative, then $S_i^n \geq 0$ for all $i = 0, \ldots, N$.

Let us define $M_i^n = \delta x \sum_{j=0}^i \rho_j^n$ and $M_i^{n+1} = \delta x \sum_{j=0}^i \rho_j^{n+1}$. Since the scheme (12) is conservative, we have $M_i^n = M_i^0$. Clearly, $\rho_i^n = (M_i^n - M_i^{n-1})/\delta x$ and from (17) we have $J_{i+1/2}^n = a_{i+1/2}^n (M_{i+1}^n - M_i^n)/2 \delta x$. Then we deduce from (12) that

$$M_i^{n+1} = (1 - \lambda c) M_i^n + \frac{\lambda}{2} \left( c - \frac{a_i^{n+1/2}}{2} \right) M_i^n + \frac{\lambda}{2} \left( c + \frac{a_i^{n+1/2}}{2} \right) M_i^{n+1}.$$

If $\lambda$ satisfies (16), we have $\lambda c < 1$ and, by assumption (9), $|a_i^{n+1/2}| \leq c$. Thus $M_i^{n+1}$ is a convex combination of $M_i^{n-1}$, $M_i^n$, and $M_i^{n+1}$. We deduce that provided condition (16) is satisfied, we have $0 \leq \rho_i^n = (M_i^n - M_i^{n-1})/\delta x$ and equation (18) implies a $BV(\mathbb{R})$ estimate on $(M_i^n)_i$. Moreover, for all $i = 0, \ldots, N$, we have $M_i^n \leq M_i^N = M_i^0$ which provides a $L^\infty$ estimate on $(M_i^n)_i$.

Summing (15), we get

$$M_i^n = \delta x \sum_{j=0}^i S_j^n - 2 \partial_x S_i^n.$$

We deduce that $0 \leq \delta x \sum_{j=0}^N S_j^n \leq M_i^N$. Then we have a $L^\infty \cap BV(\mathbb{R})$ estimate on $(\partial_x S_i^n)_i$.

Let us define

$$M_\delta(t, x) = \sum_{n \in \mathbb{N}} \sum_{i=0}^N M_i^n \mathbf{1}_{[n \delta t, (n+1) \delta t) \times [x_i, x_{i+1})}(t, x).$$

Using standard arguments, we have a $L^\infty \cap BV((0, T) \times \mathbb{R})$ estimate on $M_\delta$. It implies the convergence, up to a subsequence, of $M_\delta$ in $L^1_{loc}(\mathbb{R}^+ \times \mathbb{R})$ towards a function $M \in L^\infty \cap BV((0, T) \times \mathbb{R})$ when $\delta t$ and $\delta x$ go to 0 and satisfy (16). By the same token we have the strong convergence in $L^1_{loc}(\mathbb{R}^+ \times \mathbb{R})$ of $\partial_x S_\delta$ towards $\partial_x S$ and by definition of $\partial_x S_i$ we have the convergence in $L^1_{loc}(\mathbb{R}^+, W^{1,1}_{loc}(\mathbb{R}))$ of $S_\delta$ towards $S$. Let us define, in the weak sense, $\rho = \partial_x M \in \mathcal{M}_b(\mathbb{R})$. Obviously, noting
that \( \rho^n = (M^n_i - M^n_{i-1})/\delta x \), we deduce that \( \rho \) is the limit in the distributional sense of \( \rho_\delta \). Passing to the limit in the equation (20) we deduce that \( S \) is a weak solution of equation (3).

Moreover we have

\[
\frac{A(\partial_x S^n_{i+1}) - A(\partial_x S^n_i)}{\partial_x S^n_{i+1} - \partial_x S^n_i} = a(\theta^n_i), \quad \theta^n_i \in (\partial_x S^n_i, \partial_x S^n_{i+1}).
\]

Using assumption (9), we deduce that the sequence \((a_\delta)\) is bounded in \( L^\infty \), thus we can extract a subsequence converging in \( L^\infty \) weak* towards \( \tilde{a} \). From the \( L^1 \) convergence of \((\partial_x S_\delta)\), we deduce that \( \tilde{a} = a(\partial_x S) \) a.e. Then, from (13), we have the convergence in the sense of distributions of \( J_\delta \) towards \( J = -\partial_x (A(\partial_x S)) + a(\partial_x S) \) a.e. Finally, taking the limit in the distributional sense of equation (12) we deduce that \( \rho \) is a distributional solution to (11)-(10). By uniqueness of this solution, we deduce that \((\rho, S)\) is the unique duality solution of Theorem 2.2.

4. Numerical simulations. We present in this section some numerical results obtained with the scheme (12)-(15). Even though the scheme was designed to ensure convergence in the attractive case, we evidence that it gives interesting results in the repulsive case as well.

In all the following, the computational domain is assumed to be \([-2.5, 2.5]\) and the velocity \( c \) is normalized to 1.

4.1. Attractive case. Let us consider the function \( a(x) = 2/\pi \arctan(10x) \), which clearly satisfies (9). In Figure 1, we plot the dynamics for the smooth initial data:

\[
\rho^{ini}(x) = e^{-10(x-1.25)^2} + 0.8e^{-20x^2} + e^{-10(x+1)^2}.
\]

We notice that the blow-up occurs fastly. Then after a small time, solutions are formed by 3 peaks which can be considered as numerical Dirac masses. Eventually the Dirac masses move and collapse in finite time. This behaviour is very similar to the one observed in [7] where a particle method has been implemented.

4.2. Repulsive case. It is interesting to observe the numerical results obtained by this scheme in the repulsive case, i.e. when the function \( a \) is non-increasing. In Figure 2, we present the result for \( a(x) = -2/\pi \arctan(10x) \) (left) and \( a(x) = -2/\pi \arctan(50x) \). The initial data is \( \rho^{ini}(x) = e^{-10x^2} \). In this case, the velocity \( x \mapsto a(\partial_x S) \) does not satisfy the one-sided Lipschitz estimate (8). Therefore we cannot define measure solutions in the sense of duality. However, we can prove, using the arguments in e.g. [8] that if \( \rho^{ini} \in L^1 \cap W^{1,\infty}(\mathbb{R}) \), we have global-in-time existence of solutions in \( L^1 \cap W^{1,\infty}(\mathbb{R}) \).

We observe in Figure 2 that support of the solution increases in time. It corresponds to the fact that cells repel one another. Finally, Figure 3 displays the dynamics of the cell density in the repulsive case \( a(x) = -2/\pi \arctan(10x) \) and for the initial data \( \rho^{ini}(x) = e^{-10(x-0.7)^2} + e^{-10(x+0.7)^2} \).

The particle scheme proposed in [7] does not allow to get satisfying numerical results in the repulsive case. In fact in this latter scheme, we approximate the solution by a finite sum of Dirac masses. However the OSL condition is not satisfied, so that, as noticed above, duality solutions can not be defined in this case, so that the particle scheme cannot be properly defined.
Figure 1. Dynamics of the cell density for an initial data given by a sum of 3 regular bumps. We notice the fast blow up of solutions, then the obtained aggregates collapse together.

Figure 2. Dynamics of the cell density in the repulsive case, i.e. for a nonincreasing function $a(x) = -2/\pi \text{Atan}(kx)$. Left: $k = 10$; Right: $k = 50$. 
Figure 3. Dynamics of the cell density for an initial data given by a sum of 2 regular bumps in the repulsive case for $a(x) = -2/\pi \text{Atan}(10x)$.

REFERENCES


E-mail address: francois.james@univ-orleans.fr
E-mail address: vauchelet@ann.jussieu.fr
GLOBAL WELL-POSEDNESS OF 2D COMPRESSIBLE
NAVIER-STOKES EQUATIONS WITH LARGE
DATA AND VACUUM

QUANSEN JIU
School of Mathematical Sciences
Capital Normal University
Beijing 100048, China

YI WANG
Institute of Applied Mathematics, AMSS
and Hua Loo-Keng Key Laboratory of Mathematics, CAS
Beijing 100190, China

ZHOUPING XIN
The Institute of Mathematical Science
Chinese University of Hong Kong, Hong Kong, China

Abstract. In this note, we present main results of the global well-posedness
of the 2D compressible Navier-Stokes equations with large initial data and
vacuum. It is proved that if the shear viscosity $\mu$ is a positive constant and the
bulk viscosity $\lambda$ is the power function of the density, that is, $\lambda(\rho) = \rho^\beta$ with
$\beta > 3$, then the 2D compressible Navier-Stokes equations with the periodic
boundary conditions on the torus $T^2$ admit a unique global classical solution
$(\rho, u)$ which may contain vacuums in an open set of $T^2$. Note that the initial
data can be arbitrarily large to contain vacuum states. Moreover, the Cauchy
problem is also discussed and the main results are presented.

1. Introduction. In this note, we consider the following compressible and isen-
tropic Navier-Stokes equations with density-dependent viscosities

$$
\begin{align*}
\partial_t \rho + \text{div}(\rho u) &= 0, \\
\partial_t (\rho u) + \text{div}(\rho u \otimes u) + \nabla P(\rho) &= \mu \Delta u + \nabla ((\mu + \lambda(\rho)) \text{div} u), \quad x \in T^2, \quad t > 0,
\end{align*}
$$

where $\rho(t, x) \geq 0$, $u(t, x) = (u_1, u_2)(t, x)$ represent the density and the velocity of
the fluid, respectively. And $T^2$ is the 2-dimensional torus $[0, 1] \times [0, 1]$ and $t \in [0, T]$ for any fixed $T > 0$. We denote the right hand side of (1) by

$${\mathcal L}_\rho u = \mu \Delta u + \nabla ((\mu + \lambda(\rho)) \text{div} u).$$

2000 Mathematics Subject Classification. Primary: 35L65, 35Q30; Secondary: 76N10.

Key words and phrases. Compressible Navier-Stokes equations, density-dependent viscosity,
global well-posedness, vacuum.

Quansen Jiu is partially supported by National Natural Sciences Foundation of China (No. 11171229 & No.11228102) and Project of Beijing Education Committee.Yi Wang is partially supported by Hong Kong RGC Earmarked Research Grant CUHK4042/08P and CUHK4041/11P, and National Natural Sciences Foundation of China (No. 11171326). Zhouping Xin is partially supported by Zheng Ge Ru Funds, Hong Kong RGC Earmarked Research Grant CUHK4042/08P and CUHK4041/11P, and a Focus Area Grant at The Chinese University of Hong Kong.

701
Here, it is assumed that
\[ \mu = \text{const.} > 0, \quad \lambda(\rho) = \rho^\beta, \quad \beta > 3, \]
(2)
such that the operator \( \mathcal{L}_\rho \) is strictly elliptic.

Let the pressure function be given by
\[ P(\rho) = A\rho^\gamma, \]
(3)
where \( \gamma > 1 \) denotes the adiabatic exponent and \( A > 0 \) is the constant. Without loss of generality, \( A \) is normalized to be 1. The initial values are given by
\[ (\rho, u)(t = 0, x) = (\rho_0, u_0)(x). \]
(4)

Here the periodic boundary conditions on the unit torus \( \mathbb{T}^2 \) on \( (\rho, u)(t, x) \) are imposed to the system (1). This model problem, (1)-(4), was first proposed by Vaigant-Kazhikhov in [38] where they showed the well-posedness of the classical solution to this problem provided the initial density is uniformly away from vacuum. In this paper, we study the global well-posedness of the classical solution to this problem (1)-(4) with general nonnegative initial densities.

There are extensive studies on global well-posedness of the compressible Navier-Stokes equations in the case that both the shear and the bulk viscosity are positive constants satisfying the physical restrictions, see [15, 30, 26, 27, 35, 17, 37, 7, 32, 33, 13, 8] and the references therein. For weak solutions with large data, it is referred to [29, 11, 20]. However, the uniqueness and regularity of those weak solutions remain completely open in general. It should be pointed out that this important question is a very difficult and subtle issue since, in general, one would not expect a positive answer to this question due to the finite time blow-up results of Xin in [39], where it is shown that in the case that the initial density has compact support, any smooth solution to the Cauchy problem of the CNS without heat conduction blows up in finite time for any space dimension, see also the recent generalizations to the case for non-compact but rapidly decreasing (at far fields) initial density [36]. More recently, Huang-Li-Xin [16] proved the global well-posedness of classical solutions with small energy but large oscillations which can contain vacuums to 3D isentropic compressible Navier-Stokes equations.

The case that the viscosity coefficients depend on the density and vanish at the vacuum has received a lot attention recently, see [1, 2, 3, 4, 8, 12, 18, 19, 20, 21, 22, 23, 28, 31, 34, 40, 41, 42] and the references therein. Liu, Xin and Yang first proposed in [31] some models of the compressible Navier-Stokes equations with density-dependent viscosities to investigate the dynamics of the vacuum. On the other hand, when deriving by Chapman-Enskog expansions from the Boltzmann equation, the viscosity of the compressible Navier-Stokes equations depends on the temperature and thus on the density for isentropic flows. Also, the viscous Saint-Venant system for the shallow water is kind of important compressible Navier-Stokes equations with density-dependent viscosity. For the special case, (2), the global well-posedness result of Vaigant-Kazhikhov [38] is the first important surprising result for general large initial data with the only constraint that it is initially away from vacuum. However, in the presence of vacuum, there appear new mathematical challenges in dealing with such systems. In particular, these systems become highly degenerate. The velocity cannot even be defined in the presence of vacuum and hence it is difficult to get uniform estimates for the velocity near vacuum.

We will first investigate the global existence of the classical solution to 2-dimensional Vaigant-Kazhikhov model [38], that is, CNS system (1)-(4) with periodic
boundary condition and general nonnegative initial density. It should be noted that for the 2-dimensional problem, the basic reformulation of Vaigant-Kazhikhov [38] and the formulation in terms of the material derivative used in [13, 16] are equivalent. Following some of the key ideas developed by Vaigant-Kazhikhov [38], we are able to derive the uniform upper bound of the density under the assumptions that the initial density is nonnegative. Then we can derive the higher order estimates to the solution to guarantee the existence of the global classical solution. Finally, we will discuss the Cauchy problem to the CNS system (1)-(4) \( \mathbb{T}^2 \) replaced by \( \mathbb{R}^2 \). If the density vanishes at the far field, then the elementary energy estimates only gives the \( L^2 \)-integrability of \( \nabla u \) both in space and time, which presents no integrability on the velocity \( u \) itself. Similar situation also occurs in higher order estimates. Therefore, suitable weighted estimates in each step should be carried out in the a priori estimates. By various Caffarelli-Kohn-Nirenberg inequality and the estimates. Therefore, suitable weighted estimates in each step should be carried out in the a priori estimates. By various Caffarelli-Kohn-Nirenberg inequality and the weighted estimates, we successfully prove the global well-posedness of the classical solution to 2D Cauchy problem with large data and vacuum. The case that the density keeps a positive constant (no vacuum) at far fields will be discussed later on.

The main results can be stated in the following.

**Theorem 1.1.** If the initial values \( (\rho_0, u_0)(x) \) satisfy that

\[
0 \leq (\rho_0(x), P(\rho_0)(x)) \in W^{2,q}(\mathbb{T}^2) \times W^{2,q}(\mathbb{T}^2), \quad u_0(x) \in H^2(\mathbb{T}^2), \quad \int_{\mathbb{T}^2} \rho_0(x)dx > 0
\]

for some \( q > 2 \) and the compatibility condition

\[
\mathcal{L}_{\rho_0} u_0 - \nabla P(\rho_0) = \sqrt{\rho_0} g(x)
\]

with some \( g \in L^2(\mathbb{T}^2) \), then there exists a unique global classical solution \( (\rho, u)(t,x) \) to the compressible Navier-Stokes equations (1)-(4) with

\[
0 \leq \rho(t,x) \leq C, \quad \forall (t,x) \in [0,T] \times \mathbb{T}^2, \quad (\rho, P(\rho))(t,x) \in C([0,T]; W^{2,q}(\mathbb{T}^2)),
\]

\[
u \in C([0,T]; H^2(\mathbb{T}^2)) \cap L^2(0,T; H^3(\mathbb{T}^2)), \quad \nabla u \in L^\infty(0,T; H^3(\mathbb{T}^2)),
\]

\[
\nabla u_t \in L^\infty(0,T; W^{2,q}(\mathbb{T}^2)), \quad u_t \in L^2(0,T; H^1(\mathbb{T}^2)),
\]

\[
\nabla^2 u \in L^\infty(0,T; L^2(\mathbb{T}^2)), \quad t \nabla u_{tt} \in L^\infty(0,T; L^2(\mathbb{T}^2)),
\]

\[
\nabla u_{tt} \in L^2(0,T; H^2(\mathbb{T}^2)).
\]

(7)

**Remark 1.** From the regularity of the solution \( (\rho, u)(t,x) \), it can be shown that \( (\rho, u) \) is a classical solution of the system (1) in \([0,T] \times \mathbb{T}^2\).

If the initial values are much more regular, based on Theorem 1.1, we can prove

**Theorem 1.2.** If the initial values \( (\rho_0, u_0)(x) \) satisfy that

\[
0 \leq (\rho_0(x), P(\rho_0)(x)) \in H^3(\mathbb{T}^2) \times H^3(\mathbb{T}^2), \quad u_0(x) \in H^3(\mathbb{T}^2), \quad \int_{\mathbb{T}^2} \rho_0(x)dx > 0
\]

and the compatibility condition (6), then there exists a unique global classical solution \( (\rho, u)(t,x) \) to the compressible Navier-Stokes equations (1)-(4) satisfying all the properties listed in (7) in Theorem 1.1 with any \( 2 < q < \infty \). Furthermore, it holds that

\[
u \in L^2(0,T; H^4(\mathbb{T}^2)), \quad (\rho, P(\rho)) \in C([0,T]; H^3(\mathbb{T}^2)),
\]

\[
\rho u \in C([0,T]; H^3(\mathbb{T}^2)), \quad \sqrt{\rho} \nabla^3 u \in C([0,T]; L^2(\mathbb{T}^2)).
\]

(9)
Remark 2. In Theorem 1.4, it is not clear whether or not \( u \in C([0, T]; H^3(T^2)) \) even though one has \( \rho u \in C([0, T]; H^3(T^2)) \).

To the Cauchy problem (1)-(4) with \( T^2 \) replaced by \( \mathbb{R}^2 \), we have the following main results:

**Theorem 1.3.** Suppose that the initial values \( (\rho_0, u_0) \) satisfy

\[
0 \leq (\rho_0(x), P(\rho_0)(x)) \in W^{2,q}(\mathbb{R}^2) \times W^{2,q}(\mathbb{R}^2), \quad u_0(x) \in D^1 \cap D^2(\mathbb{R}^2), \quad \rho_0(1 + |x|^{\alpha_1}) \in L^1(\mathbb{R}^2), \quad \sqrt{\rho_0} u_0 (1 + |x|^2) \in L^2(\mathbb{R}^2), \quad \nabla u_0 |x|^{\frac{\alpha}{2}} \in L^2(\mathbb{R}^2),
\]

(10)

for some \( q > 2 \) and the weights \( 0 < \alpha < 2\sqrt{2} - 1, \alpha < \alpha_1 \), and the compatibility condition

\[
\mathcal{L}_{\rho_0} u_0 - \nabla P(\rho_0) = \sqrt{\rho_0} g(x)
\]

(11)

with some \( g \) satisfying \( g(1 + |x|^2) \in L^2(\mathbb{R}^2) \). If one of the following restrictions holds:

1) \( 1 < \alpha < 2\sqrt{2} - 1, \beta > 3, \gamma > 1, \)

(12)

2) \( 0 < \alpha \leq 1, \beta > 3, 1 < \gamma \leq 2\beta, \)

(13)

then there exists a unique global classical solution \( (\rho, u)(t, x) \) to the Cauchy problem (1)-(4) with

\[
0 \leq \rho \leq C, \quad (\rho, P(\rho))(t, x) \in C([0, T]; W^{2,q}(\mathbb{R}^2)), \quad \rho(1 + |x|^{\alpha_1}) \in C([0, T]; L^1(\mathbb{R}^2)), \quad \sqrt{\rho_0} u(1 + |x|^2), \sqrt{\rho_0} u(1 + |x|^2), \nabla u |x|^{\frac{\alpha}{2}} \in C([0, T]; L^2(\mathbb{R}^2)), \quad u \in C([0, T]; L^\infty(\mathbb{R}^2) \cap D^2(\mathbb{R}^2) \cap D^3(\mathbb{R}^2)), \quad \sqrt{t} u \in L^\infty(0, T; L^\frac{3}{2}(\mathbb{R}^2)), \quad \sqrt{t} u_t \in L^\infty(0, T; L^\frac{3}{2}(\mathbb{R}^2) \cap D^3(\mathbb{R}^2)), \quad \sqrt{t} \nabla u_t \in L^\infty(0, T; L^2(\mathbb{R}^2)), \quad \sqrt{t} \nabla u_{tt} \in L^\infty(0, T; L^2(\mathbb{R}^2)), \quad \sqrt{t} \nabla u_{ttt} \in L^\infty(0, T; L^2(\mathbb{R}^2)), \quad \sqrt{t} \nabla u_{ttt} \in L^\infty(0, T; L^2(\mathbb{R}^2)), \quad \sqrt{t} \nabla u_{ttt} \in L^\infty(0, T; L^2(\mathbb{R}^2)),
\]

(14)

where \( u \) is the material derivative of \( u \).

**Remark 3.** From the regularity of the solution \( (\rho, u)(t, x) \), it can be shown that \( (\rho, u) \) is a classical solution of the system (1) in \([0, T] \times \mathbb{R}^2 \).

**Remark 4.** If the initial data contains vacuum, then the compatibility condition (11) is necessary for the existence of the classical solution, just as the case of constant viscosity coefficients in [7].

If the initial values are more regular, we have

**Theorem 1.4.** Under assumptions of (10)-(13), assume further that

\[
0 \leq (\rho_0(x), P(\rho_0)(x)) \in H^3(\mathbb{R}^2) \times H^3(\mathbb{R}^2), \quad u_0(x) \in D^1 \cap D^3(\mathbb{R}^2)
\]

(15)

and the compatibility condition (11), then there exists a unique global classical solution \( (\rho, u)(t, x) \) to the Cauchy problem (1)-(4) satisfying all the properties listed in (14) in Theorem 1.2 with any \( 2 < q < \infty \). Furthermore, it holds that

\[
u \in L^2(0, T; D^4(\mathbb{R}^2)), \quad (\rho, P(\rho)) \in C([0, T]; H^3(\mathbb{R}^2)), \quad \rho u \in C([0, T]; D^1 \cap D^3(\mathbb{R}^2)), \quad \sqrt{\rho} \nabla^3 u \in C([0, T]; L^2(\mathbb{R}^2)).
\]

(16)

**Remark 5.** In Theorem 1.4, it is not clear whether or not \( u \in C([0, T]; D^3(\mathbb{R}^2)) \) even though one has \( \rho u \in C([0, T]; D^3(\mathbb{R}^2)) \).
2. **Sketch of Proofs.** As in [38], we introduce the following variables. First denote the effective viscous flux by

\[ F = (2\mu + \lambda(\rho))\text{div} u - P(\rho), \]

and the vorticity by

\[ \omega = \partial_{x_1} u_2 - \partial_{x_2} u_1. \]

Also, we define that

\[ H = \frac{1}{\rho}(\mu \omega x_1 + F x_2), \quad L = \frac{1}{\rho}(-\mu \omega x_2 + F x_1). \]

**Sketch of Proof of Theorem 1.1.** We construct a sequence of approximate solutions by making use of the theory of Vaigant-Kazhikhov [38] and derive some uniform a-priori estimates which are necessary to prove Theorem 1.1.

**Step 1.** Approximation of initial data

The initial density and pressure can be approximated as

\[ \rho_0^\delta = \rho_0 + \delta, \quad P_0^\delta = P(\rho_0) + \delta, \]

for any small positive constant \( \delta > 0 \). To approximate the initial velocity, we define \( u_0^\delta \) to be the unique solution to the following elliptic problem

\[ L_{\rho_0^\delta} u_0^\delta = \nabla P_0^\delta + \sqrt{\rho_0} g \]

with the periodic boundary conditions on \( \mathbb{T}^2 \) and \( \int_{\mathbb{T}^2} u_0^\delta dx = \int_{\mathbb{T}^2} u_0 dx := \bar{u}_0 \). It should be noted that \( u_0^\delta \) is uniquely determined due to the compatibility condition (11). Then one has

\[ \| u_0^\delta \|_{H^2(\mathbb{T}^2)} \leq C, \]

and

\[ \| u_0^\delta - u_0 \|_{H^2(\mathbb{T}^2)} \leq C\delta \to 0, \quad \text{as} \ \delta \to 0. \]

For the initial data \((\rho_0^\delta, P_0^\delta, u_0^\delta)\) constructed above for each fixed \( \delta > 0 \), it is proved in [38] that the compressible Navier-Stokes equations (1) with \( \beta > 3 \) has a unique global strong solution \((\rho^\delta, u^\delta)\) such that \( c_\delta \leq \rho^\delta \leq C_\delta \) for some positive constants \( c_\delta, C_\delta \) depending on \( \delta \). We will derive the uniform bound to \((\rho^\delta, u^\delta)\) with respect to \( \delta \) and then pass the limit \( \delta \to 0 \) to get the classical solution which may contain vacuum states in an open set of \( \mathbb{T}^2 \).

For simplicity of notations, we will omit the superscript \( \delta \) of \((\rho^\delta, u^\delta)\) in the following in the case of no confusions.

**Step 2.** Elementary energy estimates

**Lemma 2.1.** There exists a positive constant \( C \) depending on \((\rho_0, u_0)\), such that

\[ \sup_{t \in [0,T]} (\| \sqrt{\rho} u \|_2^2 + \| \rho \|_2^2) + \int_0^T (\| \nabla u \|_2^2 + \| \omega \|_2^2 + \| (2\mu + \lambda(\rho))^{\frac{1}{2}} \text{div} \ u \|_2^2) dt \leq C. \]

**Step 3.** Density estimates

Just as in [38], we can prove

**Lemma 2.2.** For any \( k \geq 1 \), it holds that

\[ \sup_{t \in [0,T]} \| \rho(\cdot,t) \|_k \leq C k^\frac{2}{\beta - 2}. \]

**Step 4: First-order derivative estimates of the velocity**
Lemma 2.3. There exists a positive constant $C$, such that
\[
\sup_{t \in [0,T]} \int (\mu \omega^2 + \frac{F^2}{2\mu + \lambda(\rho)})dx + \int_0^T \int \rho(H^2 + L^2)dxdt \leq C. \tag{21}
\]

Step 5: Second order derivative estimates for the velocity:

Lemma 2.4. There exists a positive constant $C$ independent of $\delta$, such that
\[
\sup_{t \in [0,T]} \int \mu(H_{x_1} - L_{x_2})^2 + (2\mu + \lambda(\rho))(H_{x_2} + L_{x_1})^2 dxdt \leq C. \tag{22}
\]

Step 6. Upper bound of the density: We are now ready to derive the upper bound for the density in the super-norm independent of $\delta$, which is crucial for the proof of Theorem 1.1.

Lemma 2.5. It holds that
\[
\int_0^T \|(F, \omega)\|_\infty^3 dt \leq C. \tag{23}
\]

With Lemma 2.5 in hand, we can obtain the uniform upper bound for the density.

Lemma 2.6. It holds that
\[
\rho(t, x) \leq C, \quad \forall (t, x) \in [0, T] \times T^2. \tag{24}
\]

With the approximate solutions and basic estimates at hand, we can derive some uniform estimates on their higher order derivatives. The details are referred to [24] and they are omitted here. The proof of Theorem 1.1 is finished.

The proof of Theorem 1.2 is referred to [24]. To prove Theorems 1.3-1.4, we need the following weighted estimates.

The following Lemma is about the Caffarelli-Kohn-Nirenberg inequalities.

Lemma 2.7. (1) $\forall h \in C_0^\infty(\mathbb{R}^2)$, it holds that
\[
\|\left|\frac{|x|^\alpha}{r} h\right|\leq C \|\left|\frac{|x|^\alpha}{r} \nabla h\right|\|_p \|\left|\frac{|x|^\beta}{r} h\right|\|_q^{1-\theta}
\]
where $1 \leq p, q < \infty, 0 < r < \infty, 0 \leq \theta \leq 1, \frac{1}{p} + \frac{\alpha}{2} > 0, \frac{1}{q} + \frac{\beta}{2} > 0, \frac{1}{r} + \frac{\kappa}{2} > 0$ and satisfying
\[
\frac{1}{r} + \frac{\kappa}{2} = \theta\left(\frac{1}{p} + \frac{\alpha-1}{2}\right) + (1-\theta)\left(\frac{1}{q} + \frac{\beta}{2}\right),
\]
and
\[
\kappa = \theta \sigma + (1-\theta)\beta,
\]
with $0 \leq \alpha - \sigma$ if $\theta > 0$ and $0 \leq \alpha - \sigma \leq 1$ if $\theta > 0$ and $\frac{1}{p} + \frac{\alpha-1}{2} = \frac{1}{r} + \frac{\kappa}{2}$. (Best constant for Caffarelli-Kohn-Nirenberg inequality)

(2) $\forall h \in C_0^\infty(\mathbb{R}^2)$, it holds that
\[
\|\left|\frac{|x|^b h}_p\| \leq C_{a,b} \|\left|\frac{|x|^\alpha}{r} \nabla h\right|_2
\]
where $a > 0, a-1 \leq b \leq a$ and $p = \frac{2}{a-b}$. If $b = a-1$, then $p = 2$ and the best constant in the inequality (27) is
\[
C_{a,b} = C_{a,a-1} = a.
\]

The following weighted energy estimates is fundamental and crucial.
Lemma 2.8. For $\alpha > 0$ and $\alpha^2 < 4(\sqrt{2} - 1)$, it holds that for sufficiently large $m > 1$ and $\forall t \in [0, T]$,

$$
\int_{\mathbb{R}^2} |x|^\alpha (\rho |u|^2 + \rho'')(t, x) \, dx + \int_0^t \left[ \| |x|^{\frac{\alpha}{2}} \nabla u \|^2_2(s) + \| |x|^{\frac{\alpha}{2}} \text{div} u \|^2_2(s) + \| |x|^{\frac{\alpha}{2}} \sqrt{\lambda(\rho)} \text{div} u \|^2_2(s) \right] \, ds \\
\leq C_\alpha \left[ 1 + \int_0^t \left( \| \rho \|^2_{2m+1}(s) + 1 \right) (\| \nabla u \|^2_2(s) + 1) \, ds \right],
$$

(28)

where the positive constant $C_\alpha$ may depend on $\alpha$ but is independent of $m$.

For more details, it is referred to [25].

REFERENCES


E-mail address: jiuqs@mail.cnu.edu.cn
E-mail address: wangyi@amss.ac.cn
E-mail address: zpxin@ims.cuhk.edu.hk
OSCILLATING WAVES AND OPTIMAL SMOOTHING EFFECT FOR ONE-DIMENSIONAL NONLINEAR SCALAR CONSERVATION LAWS

Pierre Castelli
Lycée Jacques Audiberti
63 boulevard Wilson
06600, Antibes, France

Stéphane Junca
Labo. J. A. Dieudonné, UMR CNRS 7351
Université de Nice Sophia-Antipolis
Parc Valrose, 06108 Nice cedex 02, France
and
Team Coffee
INRIA Sophia Antipolis Méditerranée
2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex, France

Abstract. Lions, Perthame, Tadmor conjectured in 1994 an optimal smoothing effect for entropy solutions of nonlinear scalar conservation laws ([19]). In this short paper we will restrict our attention to the simpler one-dimensional case. First, supercritical geometric optics lead to sequences of \( C^\infty \) solutions uniformly bounded in the Sobolev space conjectured. Second we give continuous solutions which belong exactly to the suitable Sobolev space. In order to do so we give two new definitions of a nonlinear flux and we introduce fractional BV spaces.

1. Introduction and nonlinear flux definitions. We focus on oscillating smooth solutions for one-dimensional scalar conservations laws:

\[
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad u(0,x) = u_0(x), \quad t > 0, \quad x \in \mathbb{R}. \tag{1}
\]

The aim of this paper is to build solutions related to the maximal regularity or the uniform Sobolev bounds conjectured in [19] for entropy solutions. In the one-dimensional case, piecewise smooth solutions with the maximal regularity are obtained in [12] for power-law fluxes. We seek supercritical geometric optics expansions and some special oscillating solutions. Our constructions are valid for each \( C^\infty \) flux and show that one cannot expect a better smoothing effect.

The more complex multidimensional case is dealt with in [17, 4]. For recent other approaches we refer the reader to [9, 7, 11, 8, 15, 14]. Recall that the first famous BV smoothing effect for uniformly convex flux was given by the Oleinik one-sided Lipschitz condition in the 1950s (see for instance the books [10, 18]). For solutions

2000 Mathematics Subject Classification. Primary: 35L65, 35B65; Secondary: 35B10, 35B40, 35C20.

Key words and phrases. Conservation laws, nonlinear flux, geometric optics, oscillations, Sobolev spaces, smoothing effect, fractional BV spaces.
with bounded entropy production, the smoothing effect is weaker than for entropy solutions ([12, 14]).

Let us give various definitions of nonlinear flux from [19, 17, 2]. Throughout the paper, $K$ denotes a compact real interval.

**Definition 1.** [Lions-Perthame-Tadmor nonlinear flux, [19]] $f \in C^1(K, \mathbb{R})$ is said to be a nonlinear flux on $K$ with degeneracy $\alpha$ if there exists a constant $C > 0$ such that for all $\delta > 0$,

$$\sup_{\tau^2 + \xi^2 = 1} \text{(measure}\{v \in K, |\tau + \xi f'(v)| < \delta\}) \leq C\delta^\alpha. \quad (2)$$

In [19], the authors proved a smoothing effect for entropy solutions in some Sobolev space. They obtained uniform Sobolev bounds with respect to $L^\infty$ bounds of initial data. Moreover, they conjectured a better smoothing effect:

$$u_0 \in L^\infty(\mathbb{R}) \Rightarrow u(t, .) \in W^{s, 1}_{loc}(\mathbb{R}^n), \text{ for all } s < \alpha$$

where the parameter $\alpha$ is defined in Definition 1. They proved a weaker smoothing effect which was improved in [21]. The conjecture (3) is still an open problem.

In [17] was given another definition related to the derivatives of the flux. It generalizes a notion of nonlinear flux arising in geometric optics ([5]). The next one-dimensional definition of smooth nonlinear flux is simpler than in the multidimensional case ([1, 17]).

**Definition 2.** [Smooth nonlinear flux, [17]] $f \in C^\infty(K, \mathbb{R})$ is said to be a nonlinear flux on $K$ with degeneracy $d$ if

$$d = \max_{u \in K} \left( \min \left\{ \begin{array}{l} k \geq 1, \frac{d^{1+k}f}{du^{1+k}}(u) \neq 0 \end{array} \right\} \right) < +\infty. \quad (4)$$

For the Burgers’ equation or for uniformly convex fluxes, the degeneracy is $d = 1$. That is the minimal possible value. For the cubic flux $f(u) = u^3$ on $K = [-1, 1]$, the degeneracy is $d = 2$. The cubic flux is “less” nonlinear than the quadratic flux. Notice that, with this definition, a linear flux is not nonlinear: $d = +\infty$ with the natural convention $\min(\emptyset) = +\infty$.

This definition is equivalent to Definition 1 for $C^\infty$ flux with $\alpha = \frac{1}{d}$ ([1, 17]). Therefore the Lions-Perthame-Tadmor parameter $\alpha$ is for smooth flux the inverse of an integer.

The conjectured smoothing effect (3) is proved for the first time in fractional $BV$ spaces for the class of nonlinear (degenerate) convex fluxes ([2]).

**Definition 3.** [Nonlinear degenerate convex flux, [3, 2]] Let $f$ belong to $C^1(I, \mathbb{R})$ where $I$ is an interval of $\mathbb{R}$. We say that the degeneracy of $f$ on $I$ is at least $p$ if the continuous derivative $a(u) = f'(u)$ satisfies:

$$0 < \inf_{I \times I} \frac{|a(u) - a(v)|}{|u - v|^{p}} \quad (5)$$

The lowest real number $p$, if there exists, is called the degeneracy of $f$ on $I$. If there is no $p$ such that (5) is satisfied, we set $p = +\infty$.

Let $f \in C^2(I)$. We say that a real number $y \in I$ is a degeneracy point of $f$ on $I$ if $f''(y) = 0$ (i.e. $y$ is a critical point of the function $a(u)$).

For instance, if $f$ is the power-law flux on $[-1, 1]$: $f(u) = |u|^{1+\alpha}$ where $\alpha > 0$, then the degeneracy is $p = \max(1, \alpha)$, ([3, 2]).
Remark 1. Definition 3 implies the convexity (or the concavity) of the flux $f$.

Indeed, by definition there exists $C > 0$ such that $|f'(u) - f'(v)| \geq C|u - v|^p$. Hence the difference $f'(u) - f'(v)$ never vanishes for $u \neq v$. Since the flux is continuous, $a(u) - a(v)$ has got a constant sign for $u > v$, which implies the monotonicity of $f'$ and then the convexity (or the concavity) of the flux.

Remark 2. Definition 3 is less general than Definition 1. Nevertheless, if $f$ satisfies (5) then it also satisfies (2) with $\alpha = \frac{1}{p}$, and also (4) with $d = p$ when $f$ is smooth.

The paper is organized as follows. The sequence given in Section 2 is exactly uniformly bounded in the Sobolev space conjectured in [19]. Furthermore, this sequence is unbounded in each smoother Sobolev space. In Section 3, we build solutions with the suitable regularity (3).

2. Supercritical geometric optics. We give a sequence of high frequency waves with small amplitude exactly uniformly bounded in the Sobolev space conjectured in [19]. The construction uses a WKB expansion ([5, 20]).

Theorem 4. Let $f \in C^\infty(K, \mathbb{R})$ be a nonlinear flux with degeneracy $d$ defined by (4). There exists a constant state $u \in K$ such that for any smooth periodic function $U_0$ satisfying for all $0 < \varepsilon \leq 1$, for all $x \in \mathbb{R}$, $u_0(x) = u + \varepsilon U_0(t, x) \in K$, the following properties hold:

1. there exists a positive time $T$ such that the entropy solution $u^\varepsilon$ of equation (1) with $u_0 = u_0^\varepsilon$ is smooth on $[0, T] \times \mathbb{R}$ for all $0 < \varepsilon \leq 1$,

2. the sequence $(u^\varepsilon)$ is uniformly bounded in $W^{s,1}_{loc}([0, T] \times \mathbb{R})$ for $s = \alpha = \frac{1}{d}$ and unbounded for $s > \alpha$ when $U_0' \neq 0$ a.e.

The key point is to construct a sequence of very high frequency waves near the state $u$ where the maximum in (4) is reached. Next we compute the optimal Sobolev bounds uniformly with respect to $\varepsilon$ on the WKB expansion:

$$u^\varepsilon(t, x) = u + \varepsilon U \left( t, \frac{\varphi(t, x)}{\varepsilon^d} \right) + \varepsilon r_\varepsilon(t, x).$$

To estimate the remainder in Sobolev norms, we build a smooth sequence of solutions. It is quite surprising to have such smooth sequence on uniform time strip $[0, T]$. Indeed, it is a sequence of solutions with no entropy production, without shock. But for any higher frequency, the life span $T_\varepsilon$ of $u_\varepsilon$ as a continuous solution goes towards 0 and oscillations are canceled ([17]). Thus the construction is optimal.

Remark 3. The uniform life span of the smooth sequence $(u^\varepsilon)$ is at least

$$T \sim \frac{1}{\sup_\theta \left| \frac{dU_0}{d\theta} \right|},$$

as one can see in [17]. So we can build such smooth sequence for any large time $T$ and any non constant initial periodic profile $U_0$ small enough in $C^1$. But we cannot take $T = +\infty$ since shocks always occur when $U_0$ is not constant.
**Remark 4.** For $C^\infty$ flux, the parameter $\alpha$ in Definition 1 is always the inverse of an integer. To get supercritical geometric optics expansions for all $\alpha \in [0,1]$ and not only $\alpha \in \{1/n, n \in \mathbb{N}^*\}$, we shall consider power-law flux $f(u) = |u|^{1+p}$, where $p = 1/\alpha \in [1, +\infty]$, as in [12]. In this case, $u = 0$ and the sequence is simply $u^\varepsilon(t,x) = \varepsilon U(t,\varphi(t,x)/\varepsilon^d)$, the exact entropy solution of (1) and $U(t,\theta) = U_0(\theta)$.

**Proof.** We give a sketch of the proof (see [17] for more details).

- Existence of $u$: the map $u \mapsto \min\{k \geq 1, f^{(1+k)}(u) \neq 0\}$ is upper semi-continuous, so it achieves its maximum on the compact $K$.
- WKB expansion ([13, 16, 5, 17]): we plug the ansatz $u^\varepsilon(t,x) = u + \varepsilon U_\varepsilon(t,\varphi(t,x)/\varepsilon^d)$ into (1). Notice that the exact profile $U_\varepsilon$ depends on $\varepsilon$.

Set $\lambda = f'(u)$ and $b = f^{(1+d)}(u)/(1+d)! \neq 0$. After simplification, the Taylor expansion of the flux $f(u + \varepsilon U_\varepsilon) = f(u) + \varepsilon \lambda U_\varepsilon + \varepsilon^{1+d} b U_\varepsilon^{1+d} - \varepsilon^{2+d} R_\varepsilon(U_\varepsilon)$ gives an equation for the exact profile $U_\varepsilon$ and the phase $\varphi$:

$$\frac{\partial U_\varepsilon}{\partial t} + b \frac{\partial U_\varepsilon^{1+d}}{\partial \theta} = \varepsilon \frac{\partial R_\varepsilon(U_\varepsilon)}{\partial \theta}, \quad U_\varepsilon(0,\theta) = U_0(\theta), \quad \varphi(t,x) = x - \lambda t. \quad (6)$$

The profile, which does not depend on $\varepsilon$, is

$$\frac{\partial U}{\partial t} + b \frac{\partial U^{1+d}}{\partial \theta} = 0, \quad U(0,\theta) = U_0(\theta). \quad (7)$$

- Existence of smooth solutions for a time $T > 0$ independent of $\varepsilon$: it is a consequence of the method of characteristics. Indeed, the characteristics of equation (6) are a small perturbation of characteristics of equation (7).
- Approximation in $C^1([0,T] \times \mathbb{R})$: it comes again from the method of characteristics since $\varepsilon R_\varepsilon \to 0$.

Notice that the expansion is valid in $L^1_{loc}$ after shock waves ([5]). But it is not enough to estimate the Sobolev norms.

- Sobolev estimates: roughly speaking, the order of growth of the $s$ fractional derivative $d^s/dx^s U_0(x)$ is $\varepsilon^{-sd}$. This estimate for the profile $U$ is propagated along the characteristics on $[0,T]$. We have the same estimate for $U_\varepsilon$ since $U_\varepsilon$ is near $U$ in $C^1$. Then we get the Sobolev bounds for $u_\varepsilon$.

\[ \square \]

**3. Oscillating solutions.** In this section we give exact continuous solutions with the Sobolev regularity conjectured in [19]. Indeed, we choose a suitable initial data such that the regularity is not spoiled by the nonlinearity of the flux for a positive time $T$. Furthermore, the conjectured smoothing effect is proved for the first time in fractional $BV$ spaces ([2]) for the degenerate convex class of nonlinear flux given by Definition 3. The next theorem shows the optimality of this smoothing effect. The optimality was also given in [12] in Besov spaces framework. Let us introduce the $BV^s$ spaces.
Definition 5 (Fractional BV spaces). Let \( I \) be a non empty interval of \( \mathbb{R} \). A partition \( \sigma \) of the interval \( I \) is a finite ordered subset: \( \sigma = \{x_0, x_1, \cdots, x_n\} \subset I \), \( x_0 < x_1 < \cdots < x_n \). We denote by \( S(I) \) the set of all partitions of \( I \). Let \( s \) belong to \([0, 1]\) and \( p = \frac{1}{s} \geq 1 \). The s-total variation of a real function \( u \) on \( I \) is

\[
TV^s u(I) = \sup_{\sigma \in S(I)} \sum_{k=1}^{n} |u(x_k) - u(x_{k-1})|^p.
\]

\( BV^s(I) \) is the space of real functions \( u \) such that \( TV^s u(I) < +\infty \).

\( BV^s \) spaces are introduced in [2] for applications to conservation laws. These spaces measure the regularity of regulated functions: \( BV = BV^1 \subset BV^s \subset L^\infty \). Indeed, \( BV^s(K) \) is very close to the Sobolev space \( W^{s,1/R}(K) \) ([2]):

- \( BV^s(K) \subset W^{s-\eta,1/R}(K) \) for all \( 0 < \eta < s \).
- \( BV^s(K) \neq W^{s,1/R}(K) \)

We now give continuous functions which have the \( BV^s \) regularity.

Proposition 1 (A continuous \( BV^s \) function [3]). Let \( 0 < s < 1 \), \( 0 < \eta < 1-s \) and let \( g = g_{s, \eta} \) be the real function defined on \([0, 1]\) by \( g(0) = 0 \) and for all \( x \in [0, 1] \):

\[
g(x) = x^b \cos \left(\frac{\pi}{x^c}\right), \quad \text{where} \quad b = s + \frac{s^2}{\eta} \quad \text{and} \quad c = \frac{s}{\eta}.
\]

The function \( g \) belongs to \( BV^s([0, 1]) \cap C^0([0, 1]) \) but not to \( BV^{s+\eta}([0, 1]) \).

Notice that such example do not provide a function which belongs to \( BV^s \) but not to \( \bigcup_{\epsilon > 0} BV^{s+\epsilon} \).

Proof. The extrema of \( g \) are achieved on \( x_k = k^{-1/c} \). Let \( p = \frac{1}{s} > 1 \), \( q \leq p \) and

\[
V_q = \sum_{k=1}^{+\infty} |g(x_{k+1}) - g(x_k)|^q.
\]

Since \( qb/c = q(s + \eta) \), the asymptotic behavior \(|g(x_{k+1}) - g(x_k)|^q \sim 2^q q^{-\eta} \) when \( k \to +\infty \) yields \( V_q = +\infty \) when \( q = 1/(s + \eta) \) and \( V_p < +\infty \). First this implies \( g \notin BV^{s+\eta} \). Second, for such oscillating function with diminishing amplitudes, we choose the optimal infinite partition to compute the s-total variation (see Proposition 2.3, p. 6 in [2]). Then \( g \) belongs to \( BV^s \).

We are now able to find oscillating initial data with the critical Sobolev exponent propagated by the nonlinear conservation law (1).

Theorem 6. Assume \( f \in C^\infty(K, \mathbb{R}) \) be nonlinear in the sense of Definition 2. We denote by \( d \) its degeneracy and \( s = \frac{1}{d} \). For any \( \eta > 0 \) and any time \( T > 0 \) there exists a solution \( u \in C^0([0, T] \times \mathbb{R}, \mathbb{R}) \) such that for all \( t \in [0, T] \)

\[
u(t, \cdot) \in BV^s(\mathbb{R}, \mathbb{R}) \quad \text{and} \quad u(t, \cdot) \notin BV^{s+\eta}(\mathbb{R}, \mathbb{R})
\]

The idea follows the K-S Cheng construction ([6]) with the function \( g \) given in Proposition 1.
Proof. Let \( u \in K \) a point where the maximum of degeneracy of \( f \) is achieved. We also suppose that \( u \in \hat{K} \) (the proof of Theorem \( 6 \) is quite similar if \( u \in \partial K \)).

We define the initial condition \( u_0 \) by:

\[
\begin{cases}
  u_0(x) = u & \text{if } x < 0 \\
  u_0(x) = u + \delta g(x) & \text{if } 0 \leq x \leq 1 \\
  u_0(x) = u - \delta & \text{if } 1 < x
\end{cases}
\]

where \( \delta > 0 \) is chosen such that for all \( x \in [0,1] \), \( u + \delta g(x) \in K \). Notice that for all \( x \in [0,1] \), \(-1 \leq g(x) \leq 1 \) and \( g(1) = -1 \). Moreover, \( u_0 \in BV^\infty([0,1]) \) and \( u_0 \notin BV^{\infty+\eta}([0,1]) \).

Then, following the method of characteristics, we define the function \( u(t,x) \) by:

\[
\begin{cases}
  u(t,x) = 0 & \text{if } x < 0 \\
  u(t,x) = u + \delta g(y) & \text{if } x = y + ta(u + \delta g(y)), \quad 0 \leq y \leq 1 \\
  u(t,x) = u - \delta & \text{if } 1 + ta(u - \delta) < x
\end{cases}
\]

Let \( t > 0 \) and for all \( y \),

\[
\theta_t(y) = y + ta(u + \delta g(y)).
\]

Considering the change of variable \( y = x - a(u)t \), we can assume without loss of generality that \( f'(u) = a(u) = 0 \). Since \( f \in C^\infty(K,\mathbb{R}) \), we derive from a Taylor expansion that

\[
a(u) = \frac{1}{d!} \left( a^{(d)}(u)(u - u)^d + \int_u^u (u - s)^d a^{(1+d)}(s) ds \right).
\]

Defining

\[
I_n(y) = \frac{1}{d!} \int_0^1 (1 - r)^d a^{(1+d)}(u + r \delta g(y)) dr,
\]

\[
J_n(y) = \frac{1}{d!} \int_0^1 r(1 - r)^d a^{(2+d)}(u + r \delta g(y)) dr,
\]

we get then:

\[
\theta_t(y) = y + t \delta^d g(y)^d \left( \frac{1}{d!} a^{(d)}(u) + \delta g(y) I_n(y) \right).
\]

Note that \( g, I_n, J_n \) are bounded on \([0,1] \).

For \( y \neq 0 \), since \( b \delta d = 1+c \), we have \( \frac{|g(y)|^d}{y} = O(y^c) \) at 0. Thus \( \theta_t \) is differentiable at 0 and \( \frac{d\theta_t}{dy}(0) = 1 \). For \( y \neq 0 \), we have

\[
\frac{d\theta_t}{dy}(y) = 1 + t \delta^d h_n(y),
\]

where

\[
h_n(y) = g(y)^{d-1} g'(y) \left( \frac{1}{(d-1)!} a^{(d)}(u) + (d + 1)\delta g(y) I_n(y) + \delta^2 g(y)^2 J_n(y) \right).
\]
For $y \neq 0$, since $bd = 1 + c$, we have
\[ g(y)^{d-1}g'(y) = \left( y^b \cos \left( \frac{\pi}{y^c} \right) \right)^{d-1} \left( b y^{b-1} \cos \left( \frac{\pi}{y^c} \right) + \pi cy^{b-c-1} \sin \left( \frac{\pi}{y^c} \right) \right), \]
\[ |g(y)^{d-1}g'(y)| \leq \left| \cos \left( \frac{\pi}{y^c} \right) \right|^{d-1} \left( b |y|^c \right) \left| \cos \left( \frac{\pi}{y^c} \right) \right| + \pi c \left| \sin \left( \frac{\pi}{y^c} \right) \right|. \]

Thus $g(y)^{d-1}g'(y)$ is bounded on $[0, 1]$.

Since $h_n$ is bounded on $[0, 1]$, there exists $T_3 > 0$ such that for all $y \in [0, 1]$ and for all $t \in [0, T]$, $\frac{d\theta_t}{dy}(y) > 0$. Notice that $\lim_{\delta \to 0} T_3 = +\infty$. We can take $\delta > 0$ small enough such that $T_3 > T$.

Thus for all $t \in [0, T]$, $\theta_t$ is an homeomorphism between $[0, 1]$ and $[0, 1 + ta(y - \delta)]$. Then $u(t, x)$ is a continuous solution of equation (1) on $[0, T] \times \mathbb{R}$. Furthermore, since $u_0 \in BV^s(I)$ and $u_0 \notin BV^{s+\eta}(I)$, where $I = [0, 1]$, we deduce that for all $t \in [0, T]$, $u(t, \cdot) \in BV^s(J)$ and $u(t, \cdot) \notin BV^{s+\eta}(J)$, where $J = \theta_t(I) = [0, 1 + ta(y - \delta)]$.

Finally, since $u(t, \cdot)$ is constant outside $J$, we have proved that $u(t, \cdot) \in BV^s(\mathbb{R})$ and $u(t, \cdot) \notin BV^{s+\eta}(\mathbb{R})$.

**Remark 5.** As in Remark 4, Theorem 6 is restricted for critical exponent $s$ such that $\frac{1}{s} \in \mathbb{N}$. To obtain all exponents $s \in [0, 1]$, following [12], we can consider a power-law flux with $p = \frac{1}{s}$: $f(u) = |u|^{1+p}$. Our construction is quite similar as in the proof of Theorem 6 with $u = 0$ and $\delta > 0$ small enough.

**REFERENCES**


E-mail address: Pierre.Castelli@ac-nice.fr
E-mail address: junca@unice.fr
A WENO-TVD FINITE VOLUME SCHEME FOR THE APPROXIMATION OF ATMOSPHERIC PHENOMENA

DANTE KALISE
Johann Radon Institute for Computational and Applied Mathematics
Austrian Academy of Sciences
Altenberger Straße 69, 4040 Linz, Austria

Abstract. We present a numerical scheme for the approximation of atmospheric processes in two dimensions. The scheme belongs to the class of methods proposed in [13], where a modification of the classical WENO scheme [6] was developed by replacing first-order fluxes with a high-order TVD approximation constructed via a flux-centered limiter approach. The accuracy and performance of the method is assessed by standard tests involving advective and convective motion in two dimensions.

1. Introduction. Advection and convection, understood as the class of phenomena related to fluid flow motion or transport, are concepts covering different relevant situations in atmospheric modeling. In this work, we shall be concerned with the development of an accurate scheme able to handle both types of motion in a computationally efficient manner. Models which describe such behavior are, in a first step, linear scalar advection models, including space-dependent velocity fields and, in a more elaborated formulation, the set of Euler equations for gas dynamics, parameterized in scales that are typical of atmospheric phenomena.

Achieving an accurate and physically meaningful numerical approximation of such models is undoubtedly, a challenging task. During the last decades, a method which has gained popularity among the atmospheric modeling community is the finite volume (FV) framework [14]. In a computationally efficient way, it preserves many aspects of the underlying physics (as conservation for instance) with a high level of accuracy, which is fundamental for applications such as numerical weather prediction.

In the FV context there is a considerable amount of available methods oriented to the solution of hyperbolic system of conservation laws. We are concerned with a particular class of those methods, the so-called WENO (weighted essentially non-oscillatory) schemes [6], which combine high-order, non-oscillatory reconstructions of the data with the use of monotone, low-order fluxes together with a suitable total variation diminishing (TVD) integration in time in order to generate a scheme of global high-order of accuracy preserving the non-oscillatory feature. In a recent version of the method [13], the choice of its first-order monotone flux is replaced by a high-order, TVD flux; in particular such construction is performed via a flux-limiter centered (FLIC) approach, which has the computational advantage of not requiring

2000 Mathematics Subject Classification. Primary: 65M99; Secondary: 65Z05.
Key words and phrases. Finite volume methods, High-order methods, WENO reconstruction, TVD fluxes, Atmospheric modeling.

The author acknowledges support from the Marie Curie ITN SADCO (Grant No. 264735).
any wave propagation information. The novelty of the present article is to apply this latter idea in two-dimensional problems.

An important part of this work is devoted to the numerical validation of the proposed scheme. We first address a 2D model of space-dependent advection, to then study convective phenomena based upon the Euler equations, with relevant cases for atmospheric modeling. The performance of the scheme is assessed in terms of accuracy, its ability to preserve monotonicity in the presence of sharp solutions, its robustness with respect to flux parameters, correct front locations and energy conservation.

2. A numerical scheme for the system of balance laws. In this section we present a finite volume scheme for a two-dimensional system of balance laws of the form

\[ \partial_t Q + \partial_x F(Q) + \partial_z H(Q) = 0, \quad (1) \]

where \( Q \) is a vector of conserved variables and \( F, H \) are physical fluxes; if a source term is present, such as gravity and Coriolis force in the 2D Euler equations, we opt for a splitting approach as suggested in [11] for the specific numerical flux of this article. In order to approximate eq. (1), we begin by meshing the spatial domain \( \Omega_{x,z} \) into uniform control volumes \( \Omega_{i,j} = \left[ x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}} \right] \times \left[ z_{j-\frac{1}{2}}, z_{j+\frac{1}{2}} \right] \) of size \( \Delta x \Delta z \); inside every control volume we average with respect to \( x \) and \( z \) leading to the semi-discrete scheme

\[ \frac{dQ_{i,j}(t)}{dt} = -\frac{1}{\Delta x} (F_{i+\frac{1}{2},j} - F_{i-\frac{1}{2},j}) - \frac{1}{\Delta z} (H_{i,j+\frac{1}{2}} - H_{i,j-\frac{1}{2}}) \equiv L_{i,j}(Q), \quad (2) \]

where

\[ Q_{i,j} = \frac{1}{\Delta x \Delta z} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{z_{j-\frac{1}{2}}}^{z_{j+\frac{1}{2}}} Q(x,z,t) \, dz \, dx, \quad (3) \]

\[ F_{i+\frac{1}{2},j} = \frac{1}{\Delta z} \int_{z_{j-\frac{1}{2}}}^{z_{j+\frac{1}{2}}} F(Q(x_{i+\frac{1}{2}},z,t)) \, dz, \quad (4) \]

\[ H_{i,j+\frac{1}{2}} = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} H(Q(x,z_{j+\frac{1}{2}},t)) \, dx. \quad (5) \]

The computation of expressions (4)-(5) is performed by first replacing the integrals by suitable Gaussian quadrature rules, and then establishing a numerical flux approximation at given Gauss points.

The first step towards the construction of the high-order numerical scheme consists in increasing the spatial accuracy by performing a reconstruction procedure upon the set of averaged cell values. In our case we follow the 2D WENO reconstruction proposed in [3], which yields one polynomial of prescribed degree per cell. After the reconstruction procedure is completed, any computation requiring values of the vector of conserved variables is performed by taking extrapolated boundary values; note that such approach also allows high-order of accuracy for the treatment of source terms, which will often require the evaluation at Gauss points inside the cell.

The next step in our numerical scheme deals with the calculation of the numerical fluxes (4)-(5). The classical WENO scheme replaces the original physical flux at the cell interfaces by a consistent, first-order numerical flux, which is evaluated at extrapolated boundary values arising from spatial reconstruction. Instead, we proceed as in [13], where the low-order flux is replaced by a second-order, TVD
flux. We make use of a 2D extension of the flux-limiter-centered scheme (FLIC) approach presented in [4], which is second-order, centered and non-oscillatory. In our case, it consists of a flux-limited version of a generalized Lax Wendroff flux, using as a low-order block the FORCE (first order centered) flux [8], which can be interpreted as an average of Lax-Friedrichs and Lax-Wendroff types of fluxes:

\[
F_{i+\frac{1}{2},j}^{\text{FLIC}} = F_{i+\frac{1}{2},j}^{\text{FORCE}} + \psi_{i+\frac{1}{2},j} \left( F_{i+\frac{3}{2},j}^{\text{LW}} - F_{i+\frac{1}{2},j}^{\text{FORCE}} \right),
\]

(6)

where

\[
F_{i+\frac{1}{2},j}^{\text{FORCE}} = \frac{1}{2} \left( F_{i+\frac{1}{2},j}^{L} Q_{i+\frac{3}{2},j}^{L} + F_{i+\frac{1}{2},j}^{R} Q_{i+\frac{3}{2},j}^{R} \right),
\]

(7)

\[
F_{i+\frac{1}{2},j}^{\text{LW}} = \frac{1}{2} \left( F_{i+\frac{1}{2},j}^{L} \right) + \frac{1}{2} \left( F_{i+\frac{1}{2},j}^{R} \right) - \frac{\Delta t}{4\Delta x} \left( Q_{i+\frac{3}{2},j}^{R} - Q_{i+\frac{1}{2},j}^{L} \right),
\]

(8)

Analogous formulas can be derived for the \( H_{i+\frac{1}{2},j} \); also note that even though the formulas are written along the boundary \( i+\frac{1}{2} \), \( j \), the use of the Gaussian quadrature formula will replace the axes \( j \) by Gauss points and therefore this subscript must be understood in that sense.

Note that the fluxes include the parameter \( \Delta t \), despite that so far we are dealing uniquely with space discretization: this pseudotime arises from the averaging operators originating the numerical fluxes. The \( \Delta t \) parameter which generates an updated state will only appear after the time discretization of (2).

The function \( \psi_{i+\frac{1}{2},j} = \psi_{i+\frac{1}{2},j}(r_{i+\frac{1}{2},j}^{L}, r_{i+\frac{1}{2},j}^{R}) \) is a flux limiter, which is a standard component of a TVD flux; for the numerical tests presented in this work, we use the SUPERBEE limiter, which on its centered version reads:

\[
\psi(r) = \begin{cases} 
0, & \text{if } r \leq 0, \\
2r, & \text{if } 0 \leq r \leq \frac{1}{2}, \\
1, & \text{if } \frac{1}{2} \leq r \leq 1, \quad \phi_g = \frac{1 - |c|}{1 + |c|}, \\
\min \{2, \phi_g + (1 - \phi_g)r \} r \geq 1
\end{cases}
\]

(11)

where \( c \) corresponds to the Courant number. The limiter depends on the flow parameter \( r \), which will be defined upon a physical quantity \( e \) of the system. Once \( e \) has been obtained from the discretized variables, left and right flow parameters are given by

\[
r_{i+\frac{1}{2},j}^{L} = \frac{e_{i+\frac{1}{2},j}^{R} - e_{i+\frac{1}{2},j}^{L}}{e_{i+\frac{3}{2},j}^{R} - e_{i+\frac{3}{2},j}^{L}}, \quad r_{i+\frac{1}{2},j}^{R} = \frac{e_{i+\frac{1}{2},j}^{R} - e_{i+\frac{3}{2},j}^{L}}{e_{i+\frac{3}{2},j}^{R} - e_{i+\frac{3}{2},j}^{L}},
\]

(12)

and finally,

\[
\psi_{i+\frac{1}{2},j} = \min(\psi(r_{i+\frac{1}{2},j}^{L}), \psi(r_{i+\frac{1}{2},j}^{R})).
\]

(13)

Note that the presented building block for the approximation of physical fluxes is centered, in the sense that it can be fully computed from the set of extrapolated left and right values at the cell interface, without requiring computations related to wave propagation information, which makes it computationally cheaper than its upwind counterpart.

The above described procedure starts with a set of averaged values and ends with a numerical approximation of the space operators involved in eq. (2). The resulting
scheme is still continuous in time, and we conclude this section by discretizing this operator in a way that is consistent with the choices that we have made in the generation of the space discretization operator. In order to preserve high-order of accuracy and non-oscillatory properties in time, we consider the well-known family of explicit TVD Runge-Kutta schemes [9], in particular its third order version

\[
Q_{i,j}^{n+1} = \frac{1}{3}Q_{i,j}^{n} + \frac{2}{3}Q_{i,j}^{n+\frac{2}{3}} + \frac{2}{3}\Delta t L_{i,j}(Q_{i,j}^{n+\frac{2}{3}}). 
\]

(16)

3. Numerical tests for advection and convection. We present numerical tests dealing with advective and convective motion. We test our scheme on its third-order version, meaning that we perform a quadratic WENO reconstruction in space, and time discretization is done with the previously mentioned third order RK-TVD.

3.1. Doswell frontogenesis. We address a kinematic frontogenesis problem, originally presented in [7]. It is a standard test in atmospheric modeling, and allows us to assess the performance of the scheme in the treatment of sharp fronts; numerical experiments with this test case in the context of this article can be found in [12].

For this test, we set the domain \( \Omega = [-5, 5]^2 \), and the flow is governed by

\[
\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x} (a(x,z)Q) + \frac{\partial}{\partial z} (b(x,z)Q) = 0, \tag{17}
\]

\[
a = -zf(r), \quad b = xf(r), \quad f(r) = \frac{1}{r}v(r), \quad v(r) = \bar{v} \text{sech}^2(r) \tanh(r), \quad r = \sqrt{x^2 + z^2}, \quad \bar{v} = 2.59807. \tag{18}
\]

The initial condition for this test is given by

\[
Q(x,z,0) = \tanh\left(\frac{z}{\delta}\right), \tag{20}
\]

generating the following exact solution

\[
Q(x,z,t) = \tanh\left(\frac{z \cos(\nu t) - x \sin(\nu t)}{\delta}\right). \tag{21}
\]

For this test, time stepping is selected according to

\[
\Delta t = CFL \min \left(\frac{\Delta x}{\max_{\Omega} |a|}, \frac{\Delta z}{\max_{\Omega} |b|}\right), \tag{22}
\]

with \( CFL \) the classical Courant number which is set to \( CFL = 0.45 \).

The parameter \( \delta \) is related with the thickness of the front zone. We first present numerical experiments with a value of \( \delta = 1 \), in order to generate a smooth solution to verify the convergence rates; table 1 shows that consistent convergence rates are obtained with the proposed scheme in both discrete \( L_\infty \) and \( L_1 \) norms. A second study is performed with sharp fronts, taking \( \delta = 10^{-6} \). Figure 1 illustrates the capacity of the scheme in tracking sharp fronts; it can be observed that no spurious oscillations are generated, despite the sharpness of the solution. Moreover, the scheme proves to be very robust with respect to the choice of the limiter, which is an interesting feature as it often happens that the choice of the limiter is highly dependent on the problem.
Table 1. Convergence rates for the Doswell frontogenesis problem at $t = 4 \ [s]$, with $\delta = 1$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$L_\infty$ error</th>
<th>$L_\infty$ order</th>
<th>$L_1$ error</th>
<th>$L_1$ order</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>3.4786e-001</td>
<td>1.2719e-002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.1140e-001</td>
<td>1.6</td>
<td>3.5136e-003</td>
<td>1.9</td>
</tr>
<tr>
<td>200</td>
<td>3.3302e-002</td>
<td>1.8</td>
<td>7.3045e-004</td>
<td>2.3</td>
</tr>
<tr>
<td>400</td>
<td>5.4778e-003</td>
<td>2.6</td>
<td>1.0541e-004</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Figure 1. Doswell frontogenesis problem. Results at $t = 4[s]$ for sharp initial condition with $\delta = 10^{-6}$ and $200 \times 200$ elements. Left: contour plot of the front. Right: cut at $x = 0$ with different limiters.

3.2. Euler equations: Convective tests. Throughout this section we study the behavior of our scheme when faced to convective phenomena. Our starting point corresponds to the set of equations describing the evolution in time of a 2D dry air atmosphere. Imposing conservation of mass, momentum and energy, and considering effects of gravity, together with neglecting friction and rotation effects, leads to a set of 2D inviscid primitive equations for the atmosphere written in conservative form

$$\partial_t Q + \partial_x \mathcal{F} + \partial_z \mathcal{H} = \mathcal{S},$$

where

$$Q = \begin{bmatrix} \rho \\ \rho u \\ \rho w \\ \rho \theta \end{bmatrix}, \quad \mathcal{F} = \begin{bmatrix} \rho u \\ \rho u^2 + \mathcal{P} \\ \rho u w \\ \rho u \theta \end{bmatrix}, \quad \mathcal{H} = \begin{bmatrix} \rho w \\ \rho w u \\ \rho w^2 + \mathcal{P} \\ \rho w \theta \end{bmatrix}, \quad \mathcal{S} = \begin{bmatrix} 0 \\ 0 \\ -\rho g \\ 0 \end{bmatrix}.$$
Table 2. Extreme values for the density current test case with $\Delta x = \Delta z = 50[m]$ at $t = 900[s]$.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\theta'_\text{min}$</th>
<th>$\theta'_\text{max}$</th>
<th>$u_{\text{min}}$</th>
<th>$u_{\text{max}}$</th>
<th>$w_{\text{min}}$</th>
<th>$w_{\text{max}}$</th>
<th>Front loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inviscid</td>
<td>-11.7296</td>
<td>0.8950</td>
<td>-18.6235</td>
<td>32.2655</td>
<td>-18.9703</td>
<td>22.6452</td>
<td>1.498×10^4</td>
</tr>
</tbody>
</table>

via

$$\theta = T \left( \frac{P}{P_0} \right)^{-R_d/c_p}.$$  \hfill (25)

The system is closed by the equation of state for an ideal gas

$$P = C_0 (\rho \theta)^\gamma, \quad C_0 = \frac{R_d^\gamma}{P_0 R_d/c_v}.$$ \hfill (26)

Model parameters are: the gravitational acceleration $g = 9.81[m s^{-2}]$, the atmospheric pressure at sea level $P_0 = 10^5[Pa]$, the gas constant for dry air $R_d = 287[JK^{-1}kg^{-1}]$, the specific heat of dry air at constant pressure and volume $C_p = 1004[JK^{-1}kg^{-1}]$, the specific heat of dry air at constant volume $C_v = 717[JK^{-1}kg^{-1}]$ and its ratio $\gamma = C_p/C_v = 1.4$. The expression for the total energy of the system (internal+kinetic+potential) is given by

$$e = E_{\text{Int}} + E_{\text{Kin}} + E_{\text{Pot}} = c_v T + \frac{1}{2}(u^2 + w^2) + gz,$$ \hfill (27)

which will be used as flow parameter in the limiter computation. Time stepping is selected according to

$$\Delta t = CFL \min \left( \frac{\Delta x}{(\max_{\Omega \bar{\theta}_x})}, \frac{\Delta z}{(\max_{\Omega \bar{\theta}_z})} \right),$$ \hfill (28)

with $CFL = 0.4$ and where $s_x$ and $s_z$ are maximum characteristic speeds in the $x$ and $z$ direction respectively. The gravity source term is treated via a splitting approach as in [11].

With solid boundary conditions and a warmer background state, the cold bubble drops experiencing deformation until it hits the boundary, generating horizontal shear displacement, and eventually developing Kevin-Helmholtz rotors. Final simulation time is $t = 900[s]$.

In figure 2, we study the effect of resolution at final time. There is a clear convergent behavior of the solution, and increasing resolution provides a better insight into the development of Kevin-Helmholtz rotors; at the highest tested resolution, with $\Delta x = \Delta z = 50[m]$, it can be clearly appreciated the generation of three rotors, which is similar to the results obtained in the aforementioned references for this
Figure 2. Density current test case. From top to bottom: potential temperature colormap with $\Delta x = \Delta z = 200$, $100$ and $50$[m] at $t = 900$[s].

Extremal values for the finest mesh are presented in table 2, and are in accordance with the range of values previously reported in the literature; in particular, the front location, which is a relevant quantity in this test case, coincides the results published in [2, 1]. The energy plot shown in figure 3 is qualitatively comparable to the one presented in [5] (even though is not exactly the same density current test case), exhibiting a sustained increment in the kinetic energy of the system, while both internal and potential energy decrease throughout the simulation.

4. Concluding remarks. We have presented a high-order WENO-TVD scheme with particular focus on the adequate resolution of advective and convective motion at atmospheric scales. The numerical evidence suggests that the scheme is able to reproduce accurate and physically meaningful results in an efficient manner. In the future, we expect to improve the time integration of the scheme by the use of semi-implicit techniques allowing larger time steps.

REFERENCES

**Figure 3.** Normalized energy (with respect to the initial total value) of the system for the density current test case.

**References:***


*E-mail address: dante.kalise@oeaw.ac.at*
AN ENHANCEMENT TO THE AUFS FLUX SPLITTING SCHEME BY SUN AND TAKAYAMA

FRIEDEMANN KEMM
Institute for Applied Mathematics and Scientific Computing
Brandenburg University of Technology
Platz der Deutschen Einheit 1
03046 Cottbus, Germany

Abstract. The AUFS-scheme by Sun and Takayama is a flux splitting scheme without breakdown of discrete shock profiles, usually called carbuncle, but still with a fine resolution of entropy waves. Unfortunately, in numerical tests, the viscosity on entropy waves turns out to be too small and the viscosity on shear waves to be too high. In this paper, we provide fixes to overcome these deficiencies.

1. Introduction. In their original paper, Steger and Warming [8] present not only a splitting of the Euler flux into a left- and a right-going part but also a more general framework which allows to split the flux function itself into different parts based on the splitting of the wave speeds. For the standard Steger-Warming scheme, they are split into a non-positive and a non-negative part $\lambda^-$ and $\lambda^+$. They also present a splitting of the flux which results from splitting the wave speeds into an advective part and an acoustic part. The wave speeds for the advective part are all set to the flow velocity $u$, the acoustic wave speeds to $-c$, 0 and $c$ with the speed of sound $c$. The resulting splitting of the 1d-Euler flux reads as

$$f(q) = uq + P(q) = u \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix} + \begin{pmatrix} 0 \\ p \\ pu \end{pmatrix}.$$

It is to be noted that the wave speeds used in the construction of the splitting are not the actual eigenvalues of the according flux-Jacobians. The Jacobian for the advective part is not even diagonalizable, and the eigenvalues for the central part are zero and $\pm \sqrt{\frac{\gamma - 1}{\gamma}} c$. If for the advective part, $u$ is replaced by some averaged velocity $\bar{u}$ and for the acoustic or pressure part $P$ the wave speeds are again split into a non-positive and a non-negative part, this still leads to a useful scheme. For the 2d-case, the numerical viscosity of the central part then reads as

$$\delta Q_{SW} = \frac{1}{2c_l} \begin{pmatrix} p_l \\ u_l p_l \\ v_l p_l \\ H_l p_l \end{pmatrix} - \frac{1}{2c_r} \begin{pmatrix} p_r \\ u_r p_r \\ v_r p_r \\ H_r p_r \end{pmatrix},$$

2000 Mathematics Subject Classification. Primary: 76N15, 76M12, 76L05; Secondary: 35L65, 65M06.

Key words and phrases. Flux Vector Splitting, Positivity of Pressure, contact and shear waves.
where the subscripts refer to the left and right state. Thus, the numerical flux can be written as
\[ g(q_l, q_r) = \bar{u} q_{up} + \frac{1}{2} (P_l + P_r) + \delta Q_{SW}. \]
The subscript up denotes the upwind value, e.g. for \( \bar{u} > 0 \) the left state.

Sun and Takayama [9] modify this scheme in two ways. First, they introduce some upwinding to the central part by
\[ g(q_l, q_r) = \bar{u} q_{up} + M P_{up} + (1 - M) \left( \frac{1}{2} (P_l + P_r) + \delta Q_{SW} \right), \]
where \( M \) is an estimate for the local Mach-number. Second, they replace the left and right speed of sound in the denominators of (1) by an averaged speed of sound \( \bar{c} \) to get
\[ \delta Q_{ST} = \frac{1}{2\bar{c}^2} \begin{pmatrix} p_l - p_r \\ p_l u_l - p_r u_r \\ p_l v_l - p_r v_r \\ \bar{c}^2 (p_l - p_r) + \frac{1}{2} (p_l u_l^2 - p_r u_r^2) \end{pmatrix}. \] (2)

As a justification for that, they argue that for an isentropic gas, the left and right speed of sound are the same, and both numerical viscosities would coincide with that for a Rusanov-type scheme
\[ \delta Q_{Rus} = \frac{\bar{c}}{2} \begin{pmatrix} \rho_l - \rho_r \\ \rho_l u_l - \rho_r u_r \\ \rho_l v_l - \rho_r v_r \\ E_l - E_r \end{pmatrix}. \] (3)

For an ideal (or even a real) gas, however, their choice for the viscosity \( \delta Q_{ST} \) differs from the others in one important point: in the absence of pressure and velocity jumps, it completely vanishes. Thus, the resulting method exactly resolves any pure entropy wave. At a first glance this is an appealing property, which, as is shown in their paper [9], leads to high resolution in many test cases. But this is not without risk. As Gressier, Villedieu, and Moschetta [3] prove, for a large class of FVS-schemes, there is an inherent incompatibility between the exact resolution of entropy waves and positivity. The consequences can be seen in Figure 1. There, a

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{scatter.png}
\caption{Scatter type plot of pressure in the supersonic region of the steady shock test at \( t = 16 \) when computed with original AUFS.}
\end{figure}

scatter plot of the supersonic part of the steady shock test, introduced by Dumbser.
et al. [1], is shown. The test was introduced to investigate the tendency of numerical schemes to evolve a carbuncle. It starts with a strong steady shock and small randomized numerical noise added to the primitive variables. After a short time, the original AUFS-scheme produces negative pressure. As is proved in [5], AUFS is a member of the above mentioned class of FVS-schemes for which there is an inherent incompatibility between the exact resolution of entropy waves and positivity.

Another issue of the AUFS-scheme is the third component of the viscosity $\delta Q$. In a shear wave, for all variants, i.e. (1), (2), and (3), it only vanishes if the shear wave is trivial, i.e. if there is no jump in the tangential velocity $v$. On the one hand, this seems to be the main reason why AUFS does not show the carbuncle. On the other hand, it leads to a poor resolution of boundary layers. As can be seen from Section 3.2, the resolution for the original AUFS is even poorer than with $\delta Q_{SW}$.

In this paper, we investigate the loss in the resolution of entropy waves that results from using $\delta Q_{SW}$ or $\delta Q_{Rus}$ instead of $\delta Q_{ST}$ and show how we can lower the viscosity on shear waves without sacrificing the robustness of the scheme.

2. Adjustment of the numerical viscosity. To investigate and adjust the numerical viscosity of the central pressure flux, it is important to keep in mind that for any contact wave, shear or entropy wave, there is no jump in the pressure nor in the normal velocity $u$. As a consequence, in that case, in (2) all but the third component of $\delta Q_{ST}$ vanish. The third component only vanishes if there is no shear flow, i.e. no jump in the tangential velocity $v$. This means that the original AUFS-scheme [9] exactly resolves any pure entropy wave but no pure nontrivial shear wave. As already mentioned in the introduction, for a flux vector splitting scheme, it would be desirable to have it the other way: exact resolution of pure shear waves but not of pure entropy waves. In this way one could achieve both, nice resolution of boundary layers and positivity of the scheme at the same time.

The way to retain the positivity of the scheme is quite simple: replace the original viscosity $\delta Q_{ST}$ (2) of the AUFS-scheme by the Steger-Warming viscosity $\delta Q_{SW}$ (1) or the Rusanov viscosity $\delta Q_{Rus}$ (3). As can be seen in Section 3, the price for the positivity is not too high. The loss in the overall accuracy is acceptable. If we use the Steger-Warming viscosity $\delta Q_{SW}$, we even gain a much better resolution of shear waves. This can be seen in the results of Section 3.2.

To reduce the viscosity on shear waves, a first idea would be to replace the viscosity for the pressure flux by a HLLC-type viscosity as it is derived in [5]. Unfortunately, this would not only reduce the viscosity on the shear waves but also on entropy waves. They would be resolved exactly, again leading to lack of positivity and an unstable scheme. The solution of our problem is to use a weighted average of the Rusanov- or Steger-Warming-viscosity with the HLLC-viscosity, e.g.

$$\delta Q = \theta \delta Q_{Rus} + (1 - \theta) \delta Q_{HLLC},$$

where $Q_{HLLC}$ is

$$\delta Q_{HLLC_{sym}} = \frac{\bar{c}}{2} \begin{pmatrix} 0 & \rho_l (u^* - u_l) - \rho_r (u^* - u_r) \\ \rho_l (u^* - u_l) - \rho_r (u^* - u_r) & 0 \\ (E_l - E_r) - \frac{1}{2} u^2 (\rho_l - \rho_r) \end{pmatrix}.$$ 

One might be tempted to use the residual in the Rankine-Hugoniot condition for waves travelling with speed $u$ as an indicator for shear waves as we did it in our carbuncle corrected HLLEM-solver HLLEMCC [4]. But that indicator is not able
to distinguish between shear end entropy waves. Furthermore, in the context of a FVS-method, one would want to resort to an indicator that can be obtained with less computational effort.

For both, entropy and shear waves, there is no jump in the pressure nor in the normal velocity $u$. In addition, for pure entropy waves there is no jump in the transverse velocity $v$, and for pure shear waves, there is no jump in the density. Thus, we choose as an indicator

$$\theta = \left( \frac{1}{3} \left( \frac{|p_l - p_r|}{p_l + p_r} + \frac{|u_l - u_r|}{|u_l| + |u_r|} + \frac{|\rho_l - \rho_r|}{\rho_l + \rho_r} \right) \right)^\alpha$$

with $\alpha \in [0, 1)$.

In principle, it would be possible to use any other monotone increasing function $h$ that maps $[0, 1]$ unto $[0, 1]$ instead of $(\cdot)^\alpha$. But numerical tests with piecewise linear functions indicate that we need $h(\xi) \to \infty$ as $\xi \to 0$ to get reasonable results.

### 3. Numerical results

As numerical tests, we employ only two-dimensional examples. To test the resolution of shear-waves, we employ a Kelvin-Helmholtz instability. To test the solvers for the carbuncle, we choose Dumbser’s aforementioned steady shock test [1], the colliding flow problem proposed by LeVeque [6, Section 7.7], and the famous Quirk-test [7]. As a last example, we show results for the Elling test [2] which tests the scheme for its ability to produce carbuncle type structures where they occur in physics.

As point of comparison, we give results computed with our HLLEMCC-scheme [4], a much more elaborate and, thus, more expensive scheme, which, at the one hand, prevents the breakdown of discrete shock profiles, but, on the other hand, still gives the exact flux for all linear waves, i.e. for both, shear and entropy waves. Wherever possible, we also show results of the original AUFS-scheme. For three tests, the steady shock, the colliding flow problem, and the Elling test, this is not possible, since the scheme quite soon fails due to the lack of positivity.

#### 3.1. Parameters for the tested schemes

To perform numerical tests, we first have to fix the parameters in the schemes: $\bar{u}$, $\bar{c}$, $M$, and $\alpha$. For the averaged velocity $\bar{u}$ and the estimate for the Mach-number $M$ we stick to the original settings of Sun and Takayama. For $\bar{u}$, we just take the arithmetic mean of $u_l$ and $u_r$. The Mach-number estimate is given by

$$M = \begin{cases} \bar{u} - \min\{0, u_l - c_l, \hat{u} - \hat{c}\}, & \text{if } \bar{u} > 0, \\ \bar{u} - \max\{0, u_r + c_r, \hat{u} + \hat{c}\}, & \text{if } \bar{u} \leq 0, \end{cases}$$

where the values with the hat are the isentropic approximations

$$\hat{u} = \frac{1}{2} (u_l + u_r) + \frac{c_l - c_r}{\gamma - 1}, \quad \hat{c} = \frac{1}{2} (c_l + c_r) + \frac{1}{4} (\gamma - 1)(u_l - u_r)$$

with the gas constant $\gamma$. For the Steger-Warming viscosity $\delta Q_{SW}$, we take $\hat{c}$ as the arithmetic mean of $c_l$ and $c_r$, for the Rusanov viscosity $\delta Q_{Rus}$ and the HLLC-viscosity $\delta Q_{HLLC}$, we take it to be their maximum. For $\alpha$, we test different settings: In the Steger-Warming case $\alpha = 0.05$ and $\alpha = 0.1$, in the Rusanov case $\alpha = 0.1$ and $\alpha = 0.2$. 
3.2. Kelvin-Helmholtz instability. For the Kelvin-Helmholtz instability, which tests the schemes ability to resolve shear waves, we start with a flow which consists of three parts. In a region close to the middle line, it is uniformly directed to the left. Atop and below, it is directed in the opposite direction. Pressure and density are constant everywhere. To trigger the instability, we slightly disturb the $y$-component of the flow-velocity. Since for a first order scheme on our given $100 \times 100$-grid we cannot expect the instability to evolve, we resort to a second order scheme with direction-wise minmod on primitive variables. As the results in Figure 2 show, this test reveals a severe deficiency of the original AUFS-scheme: The viscosity on shear waves is even higher than with the Steger-Warming viscosity $\delta Q_{SW}$. For $\alpha = 0.1$ in the Steger-Warming case and $\alpha = 0.2$ in the Rusanov case, the results of the modified schemes are even in the range of those for the reference scheme.

3.3. Steady shock test. The steady shock problem was introduced by Dumbser et al. [1] as test for the carbuncle. We consider the worst case: The shock is located directly on a cell face. According to Dumbser et al. [1] this situation is most likely to evolve a carbuncle-like structure. We add artificial numerical noise of amplitude $10^{-6}$ to the primitive variables in the initial state. The results in Figure 3 show that, except for the original AUFS, which fails due to the lack of positivity, all variants nicely reproduce the steady shock, even at time $t = 1000$. The oscillations alongside the shock are even smaller than in the reference scheme.

3.4. Colliding flow. LeVeque [6, Section 7.7] outlined a test problem for the carbuncle instability. It consists of a pair of slowly moving shocks, initialised by a strong colliding flow. To trigger the carbuncle, LeVeque disturbs the initial state in one grid point. In this study we superimpose artificial numerical noise of amplitude $10^{-6}$ onto the initial state instead of disturbing it in just one point. The noise is generated randomly. This allows us to make sure that the resulting structure of the solution is independent of the initial perturbation. For this test, we find in Figure 4 that lowering the viscosity leads to increasing overshoots near the shocks and, for
3.5. Quirk test. To test a scheme for its tendency to evolve the carbuncle, Quirk [7] introduced a test problem which is known as Quirk-test. Contrary to the preceding example, it is not a one-dimensional Riemann problem, but consists of a shock running down a duct. The shock is caused by Dirichlet-type boundary conditions on the left boundary. Originally, a disturbance of the middle grid line was used to trigger the instability [7]. Because the computations are done with a Cartesian code, we instead use numerical noise in the same manner as for the steady shock and the colliding flow problem. The only difference lies in the amplitude of the perturbation, here $10^{-3}$. Although one would expect the Quirk test to be the hardest one, Figure 5 clearly shows that for this type of schemes, there is no difficulty in reproducing the running shock perfectly, even with the original AUFS-flux.

3.6. Elling test. The Elling test is a variant of the steady shock test. It was introduced by Elling in [2] as an example where a carbuncle type phenomenon should occur due to physical reasons. It is a symmetrized version of the interaction of a shock with a boundary layer. Instead of imposing a boundary layer, in the middle slice of the computational grid, the flow speed on the upwind side of the
Figure 5. Scatter plot of density for Quirk test for different numerical viscosities; HLLEMCC as reference scheme.

Figure 6. Contour plot of density for Elling test with 2nd order scheme at time $t = 100$ for different numerical viscosities; HLLEMCC as reference scheme.

shock is set to zero. The question answered by this test is if in viscous computations the numerical viscosity of the inviscid scheme would outweigh the physical viscosity. If the numerical viscosity is too high, the scheme can not correctly reproduce the interaction of a shock with a boundary layer. For this reason, like in the case of the Kelvin-Helmholtz instability 3.2, we resort to a second order scheme with direction-wise minmod on primitive variables. The results in Figure 6 show that, except for the original AUFS, which fails due to the lack of positivity, all variants allow for a carbuncle like structure. The variants with the full Steger-Warming or Rusanov type viscosities $\delta Q_{SW}$ (1) and $\delta Q_{Rus}$ (3) respectively, try to retain at least part of the shock although it should break down due to physical reasons. The best results are obtained with the Rusanov viscosity $\delta Q_{Rus}$ (3) and $\alpha = 0.2$. They are close to the results for the reference scheme.

4. Conclusions and outlook. We showed that, although the original AUFS scheme by Sun and Takayama [9] lacks positivity and good resolution of shear layers, by some easy modifications, it is possible make it into a reliable scheme without breakdown of discrete shock profiles—usually called carbuncle—but with a
good resolution of shear waves. This was done by replacing their tricky numerical viscosity on the central part by standard viscosities and weighting these with the viscosity of a HLLC-type solver for the pressure flux. The indicator used for the weights is based on the relative jumps in pressure, normal velocity and density.

As can be seen from our numerical results, it is possible to convert the original AUFS method by Sun and Takayama into a numerical scheme that is reliable, of low cost, and suits to many practical applications, both, in engineering and in computational physics.

Acknowledgments. The author was partially supported by the DFG (Deutsche Forschungsgemeinschaft).

REFERENCES


E-mail address: kemm@math.tu-cottbus.de
ON POLYMER FLOODING PROBLEM IN OIL RESERVOIR SIMULATION

K. SUDARSHAN KUMAR, G. D. VEERAPPA GOWDA AND C. PRAVEEN

Tata Institute of Fundamental Research
Centre For Applicable Mathematics
Sharadanagar
Bangalore 560065, India

Abstract. We propose a high order finite volume scheme by using the idea of numerical flux developed in [2] to study the Buckley-Leverett model in two dimensional oil reservoir simulation with polymer flooding. We have considered the effect of gravity and allowed the permeability to be discontinuous in space variable. In the presence of gravity the flux functions need not be monotone and hence the exact Riemann problem is complicated and computationally expensive. To overcome this difficulty a Godunov type numerical scheme using the idea of discontinuous flux are used. Using slope limiter in the space variable and strong stability preserving Runge-Kutta scheme in the time variable, high order accuracy is achieved. Numerical simulations are performed in various test cases by considering the effect of gravity and heterogeneity of the medium.

1. Introduction. In this work we propose a high resolution finite volume scheme for the numerical simulation of Buckley-Leverett model with polymer flooding by using the idea of numerical flux developed in [2, 3]. For simplicity we let $\Omega = [0,1] \times [0,1]$ denote the two dimensional oil reservoir. Let $s \in [0,1]$ denote the saturation of aqueous phase and $c \in [0, c_0]$ denote the concentration of the polymer added to the aqueous phase, where $c_0$ is a non negative finite number. Then in the absence of capillary pressure the governing equations form a system of hyperbolic conservation laws [12] given by

$$
\begin{align*}
    s_t + \nabla \cdot F(s,c,x) &= 0 \\
    (sc)_t + \nabla \cdot (cF(s,c,x)) &= 0
\end{align*}
$$

(1)

where $(x,t) \in \Omega \times (0,\infty)$ and the flux $F : [0,1] \times [0,c_0] \times \Omega \rightarrow \mathbb{R}^2$ is given by $F = (F_1, F_2)$,

$$
F_1(s,c,x) = v_1(x)f(s,c), \quad f(s,c) = \frac{\lambda_w(s,c)}{\lambda_w(s,c) + \lambda_o(s)}
$$

(2)

$$
F_2(s,c,x) = [v_2(x) - (\rho_w - \rho_o)g\lambda_o(s,c)K(x)]f(s,c).
$$

(3)

Here $\rho_w, \rho_o$ are the densities of water and oil, $g$ is the acceleration due to gravity. The quantities $\lambda_w$ and $\lambda_o$ are the mobilities of the water and oil phase respectively.
and \( v = (v_1, v_2) \in \mathbb{R}^2 \) is the total velocity given by Darcy law \([8]\).

\[
v = -\left( (\lambda_w + \lambda_o)K(x) \frac{\partial p}{\partial x_1}, (\lambda_w + \lambda_o)K(x) \frac{\partial p}{\partial x_2} + (\lambda_w \rho_w + \lambda_o \rho_o)gK(x) \right)
\]  \( (4) \)

where \( K : \Omega \to [0, \infty) \) is the permeability of rock and allowed to be discontinuous in \( x \) and \( p : \Omega \to \mathbb{R} \) is the pressure. If we assume incompressibility of the flow and if there are no sources, then the velocity is governed by

\[
\nabla \cdot v = 0 \quad \text{in} \quad \Omega
\]  \( (5) \)

with some suitable boundary conditions for pressure on \( \partial \Omega \). For instance in the inlet part of the boundary, water is pumped in at a high pressure \( p = p_I \) while a lower pressure \( p = p_O \) is maintained on outlet part of the boundary, see Fig.1. On the remaining part of the boundary, the normal velocity is set to zero, which gives a Neumann condition on pressure. Equations \((1) \) and \((5) \) form a system of coupled algebraic-differential equations, there is no time evolution equation for the pressure (or velocity). A commonly used model for the mobilities are

\[
\lambda_w(s, c) = \frac{s^2}{\mu_w(c)}, \quad \lambda_o(s) = \frac{(1-s)^2}{\mu_o}
\]  \( (6) \)

where \( \mu_w, \mu_o \) are the viscosities of water and oil and \( \mu_w = \mu_w(c) \) is increasing in \( c \).

In the absence of polymer flooding, equivalently if the flux function is independent of \( c \) then this problem \((1) \) reduces to a two dimensional scalar equation. It is well known that in the heterogeneous media, that is when the permeability \( K(x) \) is discontinuous, fingering instability \([5]\) will develop and which results in an inefficient oil recovery. For example see Fig.3(a). As the concentration \( c \) increases, thickness of water, equivalently viscosity of water increases and the fingering effects reduces which leads to an efficient oil recovery see Fig.3(b). In the presence of the concentration \( c \) the system \((1) \) becomes coupled and non-strictly hyperbolic. When the concentration \( c \) is smooth, existence and uniqueness theory in one dimension is established in \([16]\) but we deal here with the case when \( c \) need not be smooth. For this system, developing a Godunov type upwind schemes are difficult as it needs a solution of Riemann problems. Most often numerical methods requires the calculation of eigenvalues and eigenvectors of the Jacobian matrix of the system. Here by using the idea of discontinuous flux we reduce the system to two scalar and uncoupled equations with discontinuous coefficients. Next study each scalar equation by using the idea of discontinuous flux. This approach does not require detailed information about the eigen structure of the full system. Scalar conservation laws with discontinuous flux have been studied by many authors \([1, 4, 6, 7, 9, 11, 13]\). In particular, in \([3]\) a Godunov type finite volume scheme is proposed and convergence to a proper entropy solution is proved, with certain conditions on the flux functions. In one dimensional case this problem was studied in \([2]\) and there proposed a finite volume scheme and named numerical flux as DFLU. This DFLU flux works even in cases where the upstream mobility fails \([15]\). Here we are extending DFLU to a multi dimensional case with high order accuracy. The difficulties of developing an upwind type numerical schemes in a highly heterogeneous media in the presence of gravity attracts the importance of the proposed work.

1.1. Discretization of the domain \( \Omega = [0, 1] \times [0, 1] \). Consider the Cartesian grid obtained by taking the cross product of the one-dimensional partitions \( \{x_i, i = 1, \ldots, n_x\} \) and \( \{y_j, j = 1, \ldots, n_y\} \) with \( x_1 = y_1 = 0 \) and \( x_{n_x} = y_{n_y} = 1 \). We
also introduce one layer of grid points on all four sides of $\Omega$ which will be referred to as ghost points. Thus the grid point indices range over $0 \leq i \leq n_x + 1$ and $0 \leq j \leq n_y + 1$. The grid defines the cells $Q_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, for $0 \leq i \leq n_x$ and $0 \leq j \leq n_y$. The number of real cells where the solution is supposed to be computed in the domain $\Omega$ are $n_c = (n_x - 1) \times (n_y - 1)$ (which does not count the ghost cells).

1.2. Solution of pressure equation. Define $\mu := (\lambda_w + \lambda_o) K$ and $\theta := (\lambda_w \rho_w + \lambda_o \rho_o) g K$. Integrating equation (5) over cell $Q_{i,j}$ and using the finite volume theorem, we obtain the finite volume approximation

$$(v_{i+\frac{1}{2},j} - v_{i-\frac{1}{2},j}) \Delta y + (v_{i,j+\frac{1}{2}} - v_{i,j-\frac{1}{2}}) \Delta x = 0$$

where the velocity at the cell face is given by

$$v_{i+\frac{1}{2},j} = -\mu \frac{\partial p}{\partial x} \bigg|_{i+\frac{1}{2},j}, \quad v_{i,j+\frac{1}{2}} = -(\mu \frac{\partial p}{\partial y} + \theta) \bigg|_{i,j+\frac{1}{2}}$$

After a suitable approximation of these values the system of equations for the pressure can be put in the form

$$A p = b$$

where $A \in \mathbb{R}^{n_c \times n_c}$ and $b \in \mathbb{R}^{n_c}$. This matrix equation is solved using conjugate gradient method.

1.3. Solution of saturation and concentration. Integrating equations of (1) over cell $Q_{i,j}$, we obtain the following finite volume approximations

$s_{i,j}^{n+1} = s_{i,j}^n - \left( \frac{\Delta t}{\Delta x} [F^n_{i+\frac{1}{2},j} - F^n_{i-\frac{1}{2},j}] + \frac{\Delta t}{\Delta y} [F^n_{i,j+\frac{1}{2}} - F^n_{i,j-\frac{1}{2}}] \right)$

$$s_{i,j}^{n+1} c_{i,j}^{n+1} = s_{i,j}^n c_{i,j}^n - \left( \frac{\Delta t}{\Delta x} [(cF)^n_{i+\frac{1}{2},j} - (cF)_i^{n-\frac{1}{2},j}] + \frac{\Delta t}{\Delta y} [(cF)^n_{i,j+\frac{1}{2}} - (cF)_i^{n-\frac{1}{2},j}] \right)$$

In one dimension, construction of numerical fluxes is explained in [3] in detail. Extending the idea of [3], we define the DFLU numerical fluxes as

$$F^n_{i+\frac{1}{2},j} = F_{DFLU} (s_{i+\frac{1}{2},j}^n, s_{i-\frac{1}{2},j}^n, c_{i+\frac{1}{2},j}^n, c_{i-\frac{1}{2},j}^n, v_{i+\frac{1}{2},j}, K_{i,j}, K_{i+1,j})$$

$$= \max \left\{ F_1(\max(s_{i,j}^n, \theta_{F_1}^n), c_{i,j}^n, K_{i,j}), F_1(\min(s_{i+1,j}^n, \theta_{F_1}^n), c_{i+1,j}^n, K_{i+1,j}) \right\}$$

where $(\theta_{F_1})_{i,j} = \text{argmin} \ F_1(...)$. Similarly we define $F^n_{i+\frac{1}{2},j}$ and $(cF)^n_{i,j+\frac{1}{2}}$ in the y direction.

High order reconstruction: We implement the minmod reconstruction in the space variable to achieve second order accuracy in space. After the reconstruction the updated value of saturation and concentration is given by (9) with the high order accurate flux

$$F^n_{i+\frac{1}{2},j} = F_{DFLU} (s_{i+\frac{1}{2},j}^{nL}, s_{i+\frac{1}{2},j}^{nR}, c_{i+\frac{1}{2},j}^{nL}, c_{i+\frac{1}{2},j}^{nR}, v_{i+\frac{1}{2},j}, K_{i,j}, K_{i+1,j})$$
\[ G^n_{i+\frac{1}{2},j} = \begin{cases} c^n_{L,i+\frac{1}{2},j} & \text{if } F^n_{i+\frac{1}{2},j} > 0, \\ c^n_{R,i+\frac{1}{2},j} & \text{if } F^n_{i+\frac{1}{2},j} \leq 0, \end{cases} \]

where \( s^n_{L,i+\frac{1}{2},j} \) and \( s^n_{R,i+\frac{1}{2},j} \) are the reconstructed values of saturation and concentration respectively. For the time stepping we apply a third order accurate, strong stability preserving Runge-Kutta scheme, due to Shu-Osher. If the explicit scheme (9) is stable in a norm then the Runge-Kutta scheme is also stable in the same norm under the same time step restriction (see [10]).

2. Stability results of the numerical scheme. Saturation and concentration satisfies a sort of maximum principle, which is stated as following lemmas (we omit proof here). Let

\[ M = \sup_s \{ \frac{\partial F_1}{\partial s}, \frac{\partial F_2}{\partial s}, \frac{F_1}{s}, \frac{F_2}{s} \} \quad \text{and} \quad \lambda^x = \frac{\Delta t}{\Delta x}, \quad \lambda^y = \frac{\Delta t}{\Delta y}, \]

**Lemma 2.1.** Let \( s_0 \in [0,1] \) be the initial data and let \( \{ s^n_{i,j} \} \) be the corresponding solution calculated by the finite volume scheme (9) using DFLU flux along with minmod limiter. If CFL condition \( \max\{\lambda^x M, \lambda^y M\} \leq \frac{1}{4} \) is satisfied, then

\[ 0 \leq s^n_{i,j} \leq 1 \quad \forall \ i, j \in \mathbb{Z} \quad \text{and} \quad n \in \mathbb{Z}^+. \]  

**Lemma 2.2.** Let \( \{ c^n_{i,j} \} \) be the solution calculated by the finite volume scheme (9) by using DFLU flux with minmod limiter. Under the CFL condition \( \max\{\lambda^x M, \lambda^y M\} \leq \frac{1}{4} \), concentration \( c \) satisfies the maximum principle

\[ \min\{c^n_{i,j}, c^n_{i+1,j}, c^n_{i,j+1}\} \leq c^{n+1}_{i,j} \leq \max\{c^n_{i,j}, c^n_{i+1,j}, c^n_{i,j+1}\}, \quad \forall \ n \in \mathbb{Z}^+, \ i \in \mathbb{Z}. \]

3. Numerical experiments in two dimensions. Here we have chosen an example of the quarter five-spot problem in the domain \([0,1] \times [0,1]\). To study the effect of gravity, numerical experiments are performed in the presence of gravity as well as in the absence of gravity. Also to study the polymer flooding effect, numerical experiments are performed for various concentration of the polymer. The flux function \( F = (F_1, F_2) \) takes the same form as in equation (2) and (3) with

\[ \lambda_w = \frac{s^2}{0.5 + c}, \quad \lambda_o = (1 - s)^2, \quad \rho_w g = 2 \quad \text{and} \quad \rho_o g = 1 \]

and velocity \( v \) across the grid point is calculated by using (7).

3.1. Initial and boundary conditions. The simulations are performed in a computational domain \( \Omega = [0,1] \times [0,1] \) for \( t \in [0,1] \). The initial condition on \( s \) is \( s(x,0) \equiv 0 \), i.e. initially the reservoir domain is completely filled with oil (indicated in red, see Fig.1). Through the inlet part of the boundary we pump water with a high pressure \( p = p_I \) and we keep the outlet part of the boundary with a low pressure \( p = p_O \), on remaining part of the boundary normal velocity is set zero (see Fig.1).
3.2. **Permeability of the porous media.** We consider a heterogeneous porous medium with an absolute permeability $K(x)$. In order to illustrate the robustness of the proposed numerical scheme we consider the two model porous media. The first test case corresponds to a heterogeneous medium with a continuous random permeability, given by

$$K(x) = \min\{\max\{ \sum_{i=0}^{N} \Phi_i(x), 0.5 \}, 1.5 \},$$ (12)

and

$$\Phi_i(x) = \exp\left(-\frac{|x - x_i|}{0.05}\right)^2,$$

where $x_i$ are $N$ randomly chosen locations inside the domain. Here we have taken $N = 100$. The second test case corresponds to a heavily heterogeneous medium with hard rocks and the permeability is given by choosing $N$ random locations $x_i$ and

$$K(x) = \begin{cases} 0.01 & \text{if } x \in B(x_i, 0.0015) \text{ for some } i \in \{1, 2, \ldots, N\}, \\ 1 & \text{elsewhere}. \end{cases}$$ (13)

The permeability fields for these two test cases are shown in Fig. 2. We analyze and study the simulations performed using DFLU scheme in the following experiments.

**Experiment 1:** Simulations in this experiment were performed using the spatial permeability distribution given in (12), shown in Fig. 2(a). We inject water through the inlet boundary with an inlet pressure $p_I = 8.0$ and inlet concentration $c = 0.0$. It produces fingering effects, consequently when the water front touches the outlet boundary a large amount of oil is stuck in the remaining portion of the domain, which reduces the efficiency of oil recovery, this is shown in Fig. 3(a). To avoid this fingering instability, water is injected along with some polymer with concentration $c = 7.0$. As a result the fingering instability almost disappears and the amount of oil produced at the recovery well (outlet boundary) is increased. This is shown in Fig. 3(b).

**Experiment 2:** The permeability fields chosen as in expression (13), shown in Fig. 2(b), which corresponds to a heavily heterogeneous medium with hard rocks. The result obtained in Fig. 4(a) corresponds to the saturation profile with an inlet concentration $c = 0$ and Fig. 4(b) with an inlet concentration $c = 8$. A consistent
The behavior of the saturation profile shows that our proposed scheme works well with varying spatial discontinuity in the media.

**Experiment 3:** This experiment is mainly to study the effect of gravity in saturation profile. The simulations analyzed in this experiment were performed using the spatial permeability distributions given in (13), shown in Fig.2(b) and inlet concentration $c = 8$. The expression involving the gravity term is considered along $y$ direction. The resulting figures are shown in Fig.5(a) with the absence of gravity term and in Fig.5(b) with the presence of gravity term. Observe that presence of gravity significantly affects the saturation profile.
Conclusion. A high resolution finite volume method is presented to perform the simulations of two-phase flow in porous media by using the idea of discontinuous flux. The effect of gravity in two-phase flow is analyzed. Also the efficiency of the numerical method was demonstrated by performing numerical simulations corresponding to two-phase flow in heterogeneous media.

Acknowledgments. Authors would like to thank Prof. Adimurthi for his useful suggestions and discussions.
REFERENCES


E-mail address: sudarshan@math.tifrbng.res.in
E-mail address: gowda@math.tifrbng.res.in
E-mail address: praveen@math.tifrbng.res.in
AN ERROR ESTIMATE FOR VISCOS APPROXIMATE SOLUTIONS TO DEGENERATE ANISOTROPIC CONVECTION-DIFFUSION EQUATIONS

CHRISTIAN KLINGENBERG AND UJJWAL KOLEY
Institut für Mathematik
Julius-Maximilians-Universität Würzburg
Campus Hubland Nord, Emil-Fischer-Straße 30
97074, Würzburg, Germany

Abstract. We consider a viscous approximation for a nonlinear degenerate convection-diffusion equations in two space dimensions, and prove an $L^1$ error estimate. Precisely, we show that the $L^1_{loc}$ difference between the approximate solution and the unique entropy solution converges at a rate $O(\varepsilon^{1/2})$, where $\varepsilon$ is the viscous parameter.

1. Introduction. In this paper, we are interested in certain “viscous” approximations of entropy solutions of the following Cauchy problem

\begin{equation}
\begin{cases}
u_t + f(u) y = u_{xx}, & (x, y, t) \in \Pi_T, \\
u(x, y, 0) = u_0(x, y), & (x, y) \in \mathbb{R}^2,
\end{cases}
\end{equation}

where $\Pi_T = \mathbb{R} \times \mathbb{R} \times (0, T)$ with $T > 0$ fixed, $u : \Pi_T \rightarrow \mathbb{R}$ is the unknown function and $f : \mathbb{R} \rightarrow \mathbb{R}$ is the convective flux function. The main characteristics of this type of equations is that it has mixed parabolic-hyperbolic type, due to the directional separation of the diffusion and convection effects: while matter is convected along the $y$ axis, it is simultaneously diffused along orthogonal direction.

The existence of solutions of equation (1) can be obtained by the classical method of adding a vanishing viscosity, in other terms a diffusion, in the missing direction (along the $y$ axis). Since the most characteristics of Equation (1) is that it has mixed parabolic-hyperbolic type, or in other words, it has strong degeneracy due to the lack of diffusion in the $y$-direction, it is difficult to establish the uniqueness of solutions of (1). In [1], authors have proved the existence as well as uniqueness of such problems.

In this paper we are interested in certain approximate solutions of (1) coming from solving the uniformly parabolic problem

\begin{equation}
u^\varepsilon_t + f(u^\varepsilon) y = u_{xx}^\varepsilon + \varepsilon u_{yy}^\varepsilon.
\end{equation}

We refer to $u^\varepsilon$ as a “viscous” approximate solution of (1). Since the convergence of $u^\varepsilon$ to the unique entropy solution $u$ of (1) as $\varepsilon \downarrow 0$ is well known, so our interest here is to give an explicit rate of convergence for $u^\varepsilon$ as $\varepsilon \downarrow 0$, i.e., an $L^1$ error

2000 Mathematics Subject Classification. Primary: 58F15, 58F17; Secondary: 53C35.

Key words and phrases. Degenerate convection-diffusion equations, entropy solutions, viscous approximation, doubling of variables, rate of convergence.

The second author is supported by Alexander von Humboldt Foundation.

741
estimate for viscous approximate solutions. There are several ways to prove such an error estimate. One way is to view it as a consequence of a continuous dependence estimate. Combining the ideas of [1] with a variant of Kruzkov’s “doubling of variables” device for (1), we prove that \( \|u^\varepsilon - u\|_{L^1(\Pi_T)} = O(\varepsilon^{1/2}) \). Although our proof is of independent interest, it may also shed some light on how to obtain an error estimates for numerical methods.

2. Preliminaries. Independently of the smoothness of the initial data, due to the lack of diffusion in the \( x \)-direction, jumps may form in the solution \( u \). Therefore we consider solutions in the weak sense, i.e.,

**Definition 2.1.** Set \( \Pi_T = \mathbb{R} \times \mathbb{R} \times (0,T) \). A function 

\[ u(t,x) \in C([0,T];L^1(\mathbb{R}^2)) \cap L^\infty(\Pi_T) \]

is a weak solution of the initial value problem (1) if it satisfies:

(a) For all test functions \( \varphi \in D(\mathbb{R}^2 \times [0,T]) \)

\[
\iint_{\Pi_T} (u\varphi_t + f(u)\varphi_y + u\varphi_{xx}) \, dx \, dy \, dt + \iint_{\mathbb{R}^2} u_0(x,y)\varphi(x,y,0) \, dx \, dy = 0.
\] (3)

(b) \( u = u(x,y,t) \) is continuous at \( t = 0 \) as a function: \([0,T) \rightarrow L^1(\mathbb{R}^2)\).

Since weak solutions are not uniquely determined by their initial data, one must impose an additional entropy condition to single out the physically relevant solution.

**Definition 2.2.** A weak solution \( u \) of the initial value problem (1) is called an entropy solution, if the following entropy inequality holds for all test functions \( 0 \leq \varphi \in D(\mathbb{R}^2 \times (0,T)) \):

\[
\iint_{\Pi_T} |u - \psi(x)| \varphi_t + \text{sign}(u - \psi(x))(f(u) - f(\psi(x)))\varphi_y \, dx \, dy \, dt
\]

\[
\geq \iint_{\Pi_T} -|u - \psi(x)| \varphi_{xx} - \text{sign}(u - \psi(x))\psi_{xx}\varphi \, dx \, dy \, dt.
\] (4)

**Remark 1.** Note that the above entropy condition is inspired by Kružkov’s entropy condition for scalar conservation laws. However, due to the presence of diffusion term \( u_{xx} \) in (1), Kružkov’s entropy condition has to be modified in this case. This is done by introducing as entropy test functions all functions of the form \( |u - \psi(x)| \) and \( \psi \) smooth, while Kružkov’s definition asks for \( \psi \) to be a constant.

Regarding the regularized problem (2), it is well known [1] that for all \( \varepsilon > 0 \) this problem has a unique solution

\[ u^\varepsilon \in C([0,T];L^1(\mathbb{R}^2)) \cap L^\infty(\Pi_T). \]

In fact,

\[ u^\varepsilon \in C([0,T];W^{2,p}(\mathbb{R}^2)) \cap C^1([0,T];L^p(\mathbb{R}^2)), \]

for all \( 1 < p < \infty \). In addition, \( u^\varepsilon \) satisfies the following properties

(a) \( \int_{\mathbb{R}^2} u^\varepsilon(x,y,t) \, dx \, dy = \int_{\mathbb{R}^2} u_0(x,y) \, dx \, dy \),

(b) \( TV(u^\varepsilon(\cdot,\cdot,t)) \leq TV(u_0) \),

(c) \( \|\partial_t u^\varepsilon(t)\|_1 \leq C[TV(u_0) + TV((u_0)_x)] \).

Now we are in a position to state our main result, which is the following
Main Theorem. Let \( u \) be the unique entropy solution to (1) and \( u^\varepsilon \) be as defined by (2). Assume that \( u_0 \in BV \), \( f \) is Lipschitz continuous. Choose a constant
\[
M > \max_{|u|<\|u_0\|_{L^\infty(\mathbb{R}^2)}} |f'(u)|,
\]
and another constant \( L > MT \), where \( T > 0 \). Then there exists a constant \( C \), independent of \( \varepsilon \), but depending on \( f, L, T \) and \( u_0 \), such that
\[
\int_{\mathbb{R}} \int_{-L-Mt}^{L-Mt} |u(t,x,y) - u^\varepsilon(t,x,y)| \, dx dy \leq C\varepsilon^{1/2} \text{ for } t \leq T.
\]

3. Proof of the main theorem. The theorem will be proved by a “doubling of the variables” argument, which was introduced by Kruzhkov [3, 4] in the context of hyperbolic conservation laws.

First, observe that from (4) we have for any test function \( \varphi \) with compact support in \( \mathbb{R} \times \mathbb{R} \times (0,T) \) and \( u = u(x,z,s) \),
\[
\int \int \int_{\Omega_T} |u - \psi(z)| \varphi_x + \text{sign}(u - \psi(z))(f(u) - f(\psi(z))) \varphi_z \, dx dz ds
\]
\[
\geq \int \int \int_{\Omega_T} -|u - \psi(z)| \varphi_{xx} - \text{sign}(u - \psi(z))\psi_{xx} \varphi \, dx dz ds,
\]
(5)

For the regularized equation (2), we start not with the entropy condition, but in the argument leading up to this condition. To do that, first define the regularized counterpart of the signum function as
\[
\text{sign}_\eta(\sigma) = \begin{cases} 
\text{sign}(\sigma) & |\sigma| > \eta, \\
\sin\left(\frac{\pi\sigma}{2\eta}\right) & \text{otherwise},
\end{cases}
\]
where \( \eta > 0 \) and the signum function is defined as
\[
\text{sign}(\sigma) = \begin{cases} 
-1 & \sigma < 0, \\
0 & \sigma = 0, \\
1 & \sigma > 0,
\end{cases}
\]
Set
\[
\psi_\eta(u, \psi) = \int_u^\psi \text{sign}_\eta(z - \psi) \, dz.
\]

This is a convex entropy for all \( \psi \). Set \( u^\varepsilon = u^\varepsilon(x,y,t) \) and rewrite (2) as
\[
u^\varepsilon_t + (f(u^\varepsilon) - f(\psi))_y = u^\varepsilon_{xx} + \varepsilon(u^\varepsilon - \psi(x))_{yy},
\]
and multiply this with \( \psi_\eta(u, \psi) \varphi \) where \( \varphi \) is a test function with compact support in \( \mathbb{R} \times \mathbb{R} \times (0,T) \). Observe that the solution \( u^\varepsilon \) of (2) is smooth. Hence after a partial integration, we arrive at
\[
\int \int \int_{\Omega_T} \psi_\eta(u^\varepsilon, \psi) \varphi_t + Q_\eta(u^\varepsilon, \psi) \varphi_y \, dx dy dt
\]
\[
= \int \int \int_{\Omega_T} -\text{sign}_\eta(u^\varepsilon - \psi) u^\varepsilon_{xx} \varphi - \varepsilon \text{sign}_\eta(u^\varepsilon - \psi)(u^\varepsilon - \psi(x))_{yy} \varphi \, dx dy dt,
\]
where we have used $Q'_\eta(u, \psi) = \psi'_\eta(u, \psi)f'(u)$. Next, taking limit as $\eta \to 0$, we end up with the parabolic equality

\[
\int \int \int_{\Pi_T} |u^\varepsilon - \psi(x)| \varphi_t + \text{sign}(u^\varepsilon - \psi(x))(f(u^\varepsilon) - f(\psi(x)))\varphi_y \, dx \, dy \, dt
\]

\[
= \int \int \int_{\Pi_T} -\text{sign}(u^\varepsilon - \psi(x))u^\varepsilon_{x} \varphi - \varepsilon \text{sign}(u^\varepsilon - \psi(x))(u^\varepsilon - \psi(x))_{yy} \varphi \, dx \, dy \, dt.
\]

At this point we are ready to use “doubling of the variables” device. First, using the entropy inequality (5) for the solution $u = u(x, z, s)$ with $\psi(x) = u^\varepsilon(x, y, t)$, we get for $(y, t) \in \mathbb{R} \times (0, T)$

\[
\int \int \int_{\Pi_T} |u - u^\varepsilon| \varphi_s + \text{sign}(u - u^\varepsilon)(f(u) - f(u^\varepsilon))\varphi_z \, dx \, dz \, ds
\]

\[
\geq \int \int \int_{\Pi_T} -|u - u^\varepsilon| \varphi_{xx} - \text{sign}(u - u^\varepsilon)u^\varepsilon_{xx} \varphi \, dx \, dz \, ds.
\]

Similarly, from the parabolic equality (6) for the solution $u^\varepsilon = u^\varepsilon(x, y, t)$ with $\psi(x) = u(x, z, s)$, we get for $(z, s) \in \mathbb{R} \times (0, T)$

\[
\int \int \int_{\Pi_T} |u^\varepsilon - u| \varphi_t + \text{sign}(u^\varepsilon - u)(f(u^\varepsilon) - f(u))\varphi_y \, dx \, dy \, dt
\]

\[
= \int \int \int_{\Pi_T} -\text{sign}(u^\varepsilon - u)u^\varepsilon_{x} \varphi - \varepsilon |u^\varepsilon - u|_{yy} \varphi \, dx \, dy \, dt.
\]

We now integrate (7) over $(y, t) \in \mathbb{R} \times (0, T)$ and (8) over $(z, s) \in \mathbb{R} \times (0, T)$. Addition of those two results yields

\[
\left( \int \int \int_{Q_T} |u^\varepsilon - u| (\varphi_t + \varphi_s) + \text{sign}(u^\varepsilon - u)(f(u^\varepsilon) - f(u))(\varphi_y + \varphi_z) \, dX \right)
\]

\[
\geq \left( \int \int \int_{Q_T} -|u^\varepsilon - u| \varphi_{xx} - \varepsilon |u^\varepsilon - u|_{yy} \varphi \, dX \right),
\]

\[
:= \int \int \int_{Q_T} Q_1 + Q_2 \, dX,
\]

where $dX = dx \, dy \, dz \, dt \, ds$ and $Q_T = \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times (0, T) \times (0, T)$.

Following Kruzkov and Kuznetsov [3, 4] we now specify a nonnegative test function $\varphi = \varphi(x, y, t, z, s)$ defined in $Q_T$. To this end, let $\omega \in C_0^\infty(\mathbb{R})$ be a function satisfying

\[
\text{supp}(\omega) \subset [-1, 1], \quad \omega(\sigma) \geq 0, \quad \int \omega(\sigma) \, d\sigma = 1,
\]

and define $\omega_r(x) = \omega(x/r)/r$. Next, let us choose $\phi \in C_0^\infty(\mathbb{R})$ such that

\[
\phi = \begin{cases} 
1, & |x| < 1, \\
0, & |x| \geq 2,
\end{cases}
\]
and $0 \leq \phi \leq 1$ when $1 \leq |x| \leq 2$. Then we define $K_\beta(x) = \phi(x/\beta)$. We will let $\beta \to \infty$ later. Furthermore, let $h(z)$ be defined as

$$h(z) = \begin{cases} 0, & z < -1, \\ z + 1 & z \in [-1, 0], \\ 1 & z > 0. \end{cases}$$

and set $h_\alpha(z) = h(\alpha z)$. Let $\nu < \tau$ be two numbers in $(0, T)$, for any $\alpha > 0$ define

$$H_\alpha(t) = \int_{-\infty}^{t} \omega_\alpha(\xi) d\xi,$$

$$\Psi(y, t) = (H_\alpha(t - \nu) - H_\alpha(t - \tau)) \left( h_\alpha(y - L_l(t)) - h_\alpha(y - L_r(t) - \frac{1}{\alpha}) \right)$$

$$=: \chi^\alpha_{(\nu, \tau)}(t) \chi_{(L_l, L_r)}(y, t)$$

where the lines $L_l, r$ are given by

$$L_l(t) = -L + Mt, \quad L_r(t) = L - Mt$$

where $M$ and $L$ are positive numbers, $M$ will be specified below. With $0 < r < \min \{\nu, T - \tau\}$ and $\alpha_0 \in (0, \min \{\nu - r, T - \tau - r\})$ we set

$$\varphi(x, y, t, z, s) = K_\beta(x) \Psi(y, t) \omega_r(y - z) \omega_{\tau_0}(t - s). \quad (10)$$

We note that $\varphi$ has compact support and also that we have,

$$\varphi_t + \varphi_s = K_\beta(x) \Psi_t(y, t) \omega_r(y - z) \omega_{\tau_0}(t - s),$$

$$\varphi_y + \varphi_z = K_\beta(x) \Psi_y(y, t) \omega_r(y - z) \omega_{\tau_0}(t - s).$$

For the record, we note that

$$\Psi_t(y, t) = -\chi^{\alpha_0}_{(\nu, \tau)}(t) M \left( h'_\alpha(y - L_l(t)) + h'_\alpha(y - L_r(t) - \frac{1}{\alpha}) \right)$$

$$+ (\omega_\alpha(t - \nu) - \omega_\alpha(t - \tau)) \chi_{(L_l, L_r)}(y, t), \quad (11)$$

$$\Psi_y(y, t) = \chi^{\alpha_0}_{(\nu, \tau)}(t) \left( h'_\alpha(y - L_l(t)) - h'_\alpha(y - L_r(t) - \frac{1}{\alpha}) \right).$$

We shall let all the “small parameters” $\alpha, \alpha_0, r, \tau_0, \varepsilon$ and $\Delta x$ be sufficiently small and the “large parameter” $\beta$ be sufficiently big, but fixed.

Starting the first term on the left of (9), we write

$$\int_{Q_T} |u^\varepsilon - u| (\varphi_s + \varphi_t) \, dX \leq \int_{\Pi_T} |u^\varepsilon(x, y, t) - u(x, y, t)| K_\beta(x) \Psi_t dxdydt$$

$$+ \int_{\Pi_T} \int_{\mathbb{R}} |u(x, y, t) - u(x, y, s)| K_\beta(x) |\Psi_t(x, t)| \omega_{\tau_0}(t - s) \, dxdydsdt$$

$$+ \int_{Q_T} |u(x, y, s) - u(x, z, s)| K_\beta(x) |\Psi_t(x, t)| \omega_{\tau_0}(t - s) \omega_r(x - y) \, dX. \quad \gamma$$

Following [2], it is easy to find that

$$\beta + \gamma \leq C (r_0 + r). \quad (12)$$
To continue the estimate with the first term on the left of (9), we split $\delta$ as follows

$$
\delta = - \int \int \int_{\Pi_T} \chi_{(\nu, \tau)}^0(t) M \left( h'_{\alpha}(y - L_l(t)) + h'_{\alpha}(y - L_r(t) - \frac{1}{\alpha}) \right) K_{\beta}(x) |u^\varepsilon(x, y, t) - u(x, y, t)| \, dx \, dy \, dt
+ \int \int \int_{\Pi_T} \chi_{(L_l, L_r)}(y, t) |u^\varepsilon(x, y, t) - u(x, y, t)| K_{\beta}(x) (\omega_{\alpha_0}(t - \nu) - \omega_{\alpha_0}(t - \tau)) \, dx \, dy \, dt,
$$

$$
:= \delta_1 + \delta_2.
$$

The term $\delta_1$ will be balanced against the first order derivative term on the left hand side of (9). To estimate $\delta_2$ we set $e(x, y, t) = |u^\varepsilon(x, y, t) - u(x, y, t)|$ and following [2], we find

$$
\delta_2 \leq \int \chi_{(L_l, L_r)}^0(y, \nu) K_{\beta}(x) |u^\varepsilon(x, y, \nu) - u(x, y, \nu)| \, dx \, dy
- \int \chi_{(L_l, L_r)}^0(y, \tau) K_{\beta}(x) |u^\varepsilon(x, y, \tau) - u(x, y, \tau)| \, dx \, dy + C_{\alpha_0}.
$$

(13)

Now we rewrite the “first derivative term” on the left hand side of (9). Doing this, we get

$$
\int_{Q_T} K_{\beta}(x) \text{sign}(u^\varepsilon - u) (f(u^\varepsilon) - f(u)) (\varphi_y + \varphi_z) \, dX
= \int_{Q_T} K_{\beta}(x) \sgn(x, y, z, t, s) \left( f(u^\varepsilon(x, y, t)) - f(u(x, y, t)) \right)
\Psi_y(y, t) \omega_r(y - z) \omega_{\gamma_0}(t - s) \, dX
+ \int_{Q_T} K_{\beta}(x) \sgn(x, y, z, t, s) \left( f(u(x, y, t)) - f(u(x, z, s)) \right)
\Psi_y(y, t) \omega_r(y - z) \omega_{\gamma_0}(t - s) \, dX
=: \delta_3 + \delta_4,
$$

where we have set $\sgn(x, y, z, t, s) = \text{sign}(u^\varepsilon(x, y, t) - u(x, z, s))$. We proceed as follows

$$
|\delta_4| \leq \int_{Q_T} |f(u(x, y, t)) - f(u(x, z, s))| \chi_{(\nu, \tau)}^0(t) K_{\beta}(x)
\omega_{\gamma_0}(t - s) \omega_r(y - z) \left( h'_{\alpha}(y - L_l(t)) + h'_{\alpha}(y - L_r(t) - \frac{1}{\alpha}) \right) \, dX.
$$

(14)

We follow [2] to estimate each of these two terms to conclude

$$
|\delta_4| \leq C (r_0 + r).
$$

Again, choosing $M$ larger than the Lipschitz norm of $f$ implies that

$$
\delta_1 + \delta_3 \leq 0.
$$

(15)
Collecting all the terms we see that
\[
\begin{align*}
\int_{Q_T} K_\beta(x) \chi^{\alpha}_{(L_t, L_r)}(y, \tau) |u^\varepsilon(x, y, \tau) &- u(x, y, \tau)| \, dxdy \\
&\leq \int_{Q_T} K_\beta(x) \chi^{\alpha}_{(L_t, L_r)}(y, \nu) |u^\varepsilon(x, y, \nu) - u(x, y, \nu)| \, dxdy \\
&\quad + C (r_0 + r + \alpha_0 + \alpha) + \left| \int_{Q_T} Q_1 + Q_2 \, dX \right|.
\end{align*}
\] (16)

In order to estimate \( Q_1 \) and \( Q_2 \), we proceed as follows
\[
\left| \int_{Q_T} Q_1 \right| \, dX = \int_{Q_T} |u^\varepsilon - u| \, \varphi_{xx} \, dX
\]
\[
= \int_{Q_T} |u^\varepsilon - u| K_\beta(x) \Psi(y, t) \omega_r(y - z) \omega_{r_0}(t - s) \, dX
\]
\[
\leq ||u^\varepsilon + u||_{L^\infty} \int_{Q_T} \frac{1}{\beta^2} \phi''(\frac{x}{\beta}) \Psi(y, t) \omega_r(y - z) \omega_{r_0}(t - s) \, dX
\]
\[
\leq \frac{C}{\beta^2} \int_{Q_T} \int_{Q_T} \phi''(\frac{x}{\beta}) \chi^{\alpha}_{(\nu, \tau)}(t) \chi^{\alpha}_{(L_t, L_r)}(y, t) \, dxdydt
\]
\[
\leq \frac{CK}{\beta^2}.
\]

where \( K := \int_{R} |\phi''(y)| \, dy \).

Next,
\[
\left| \int_{Q_T} Q_2 \right| \, dX = \varepsilon \int_{Q_T} |u^\varepsilon - u|_y \varphi_y \, dX
\]
\[
= \varepsilon \int_{Q_T} \text{sign}(u^\varepsilon - u)(u^\varepsilon)_y K_\beta(x) \Psi(y, t) \omega_r(y - z) \omega_{r_0}(t - s)
\]
\[
+ \varepsilon \int_{Q_T} \text{sign}(u^\varepsilon - u)(u^\varepsilon)_y K_\beta(x) \Psi(y, t) \omega'_r(y - z) \omega_{r_0}(t - s) \, dX
\]
\[
:= Q_{2,1} + Q_{2,2}.
\]

Each of the above terms can be approximated as follows
\[
\left| \int_{Q_T} Q_{2,1} \right| \, dX = \varepsilon \int_{Q_T} \text{sign}(u^\varepsilon - u)(u^\varepsilon)_y K_\beta(x) \Psi(y, t) \omega_r(y - z) \omega_{r_0}(t - s) \, dX
\]
\[
\leq \varepsilon K \int_{\Pi_T} (u^\varepsilon)_y(x, y, t) \chi^{\alpha}_{(\nu, \tau)}(t) \left( h'_\alpha(y - L_1(t)) - h'_\alpha(y - L_r(t) - \frac{1}{\alpha}) \right) \, dxdydt
\]
\[
\leq \varepsilon K \varepsilon \alpha \int_{\Pi_T} |u^\varepsilon| \, dxdy
\]
\[
\leq CK \varepsilon \alpha,
\]

where \( K = \|K_\beta\|_{\infty} \) and \( C = \text{TV}(u^\varepsilon) \). Similarly for the other term
\[
\left| \int_{Q_T} Q_{2,2} \right| \, dX = \varepsilon \int_{Q_T} \text{sign}(u^\varepsilon - u)(u^\varepsilon)_y K_\beta(x) \Psi(y, t) \omega'_r(y - z) \omega_{r_0}(t - s) \, dX
\]
\[
\leq \frac{K\varepsilon}{r} \int_{\Pi_T} (u^\varepsilon)_y(x, y, t) \chi^{\alpha}_{(\nu, \tau)}(t) \chi^{\alpha}_{(L_t, L_r)}(y, t) \, dxdydt
\]
\[
\leq CK \varepsilon \alpha,
\]
where again $K = \|K\|_{\infty}$ and $C = TV(u^\varepsilon)$.

Therefore
\[
\left| \int_{\Pi_T^2} Q_1 + Q_2 \, dX \right| \leq C\varepsilon + \frac{C}{\beta\alpha} + \frac{C\varepsilon}{r}.
\]

(17)

where $C$ depends on (among other things) $L$ and $T$, but not on the parameters $\alpha_0$, $\alpha$, $r_0$, $r$, $\beta$ or $\varepsilon$.

Now we have proved the following Lemma:

**Lemma 3.1.** Assume that $u$ and $u^\varepsilon$ take values in the interval $[-K,K]$ for some positive $K$. Let $M > \max_{v \in [-K,K]} |f'(v)|$. Then if $T \geq \tau > \nu \geq 0$ and $L - Mt > 0$, we have
\[
\int_{\mathbb{R}} \int_{-L+Mt}^{L-Mt} |u^\varepsilon(x,y,\tau) - u(x,y,\tau)| \, dx \, dy \\
\leq \int_{\mathbb{R}^2} |u^\varepsilon(x,y,\nu) - u(x,y,\nu)| \, dx \, dy \\
+ C \left[ r_0 + r + \alpha + \varepsilon + \frac{1}{\alpha\beta} + \frac{\varepsilon}{r} \right].
\]

(18)

This follows from (16) and (17), observing that we can send $\alpha_0$ to zero. Now we let $u(x,y,t)$ be the unique entropy solution of (1). Also note that since we let $\beta$ tends to $\infty$ so $\gamma = \frac{1}{\beta}$ is small. We set $\alpha = r = r_0$ and $\varepsilon = \gamma$, and assume that $\alpha$ is sufficiently small, then
\[
\int_{\mathbb{R}} \int_{-L+Mt}^{L-Mt} |u_{\Delta x}(x,t) - v(x,t)| \, dx \, dy \leq C \left( \alpha + \frac{\varepsilon}{\alpha} \right),
\]

(19)

for some constant $C$ which is independent of the small parameters $\alpha, \varepsilon$. This follows from (18). Then setting $\varepsilon = \alpha^2$ proves the main theorem.

**Acknowledgments.** The second author acknowledges support from the Alexander von Humboldt Foundation, through a Humboldt Research Fellowship for postdoctoral researchers.

**REFERENCES**


E-mail address: klingenberg@mathematik.uni-wuerzburg.de
E-mail address: toujjwal@gmail.com
1-D CONSERVATIVE SYSTEMS: A GEOMETRIC APPROACH

MICHAEL BENFIELD
Department of Mathematics, North Carolina State University
Raleigh, NC 27695, USA

HELGE KRISTIAN JENSSEN AND IRINA A. KOGAN
Department of Mathematics, Penn State University, University Park
State College, PA 16802, USA
and
Department of Mathematics, North Carolina State University
Raleigh, NC 27695, USA

Abstract. We consider general 1-d systems of hyperbolic conservation laws with the aim of better understanding of the structure ("geometry") of such systems. For a given frame $\mathcal{R}$ of vector fields, we derive two new overdetermined PDE systems: the $F(\mathcal{R})$- and $\eta(\mathcal{R})$-systems. The former gives all conservative systems with eigenframe $\mathcal{R}$, while the latter provides extensions for these conservative systems. We also describe a recent result for non-rich, strictly hyperbolic, $3 \times 3$ conservative systems, namely: two such systems with the same eigen curves must necessarily also share Hugoniot loci.

1. Introduction. We consider 1-d, conservative $n \times n$-systems
\[ u_t + f(u)_x = 0, \quad t \geq 0, \; x \in \mathbb{R}, \] (1)
where the unknown $u = u(t, x) = (u^1, \ldots, u^n)^T \in \mathbb{R}^n$ is the vector of conserved quantities. The smooth flux $f : \Omega^{\text{open}} \subset \mathbb{R}^n \to \mathbb{R}^n$ is assumed to be hyperbolic: its Jacobian $Df(u)$ is $\mathbb{R}$-diagonalizable at each state $u \in \Omega$. In this case the eigenvectors $r_i|_u$ of $Df(u)$ form a basis for each $u \in \Omega$, and we refer to $\mathcal{R} := \{r_1, \ldots, r_n\}$ as an eigenframe of (1). The system (1) is strictly hyperbolic provided the corresponding eigenvalues $\lambda^1(u), \ldots, \lambda^n(u)$ are distinct at each $u \in \Omega$.

While there is a well-developed theory for near-equilibrium solutions to strictly hyperbolic systems (1) (see [2]), far less is known about large variation solutions. In fact, there are systems admitting blowup solutions. While blowup of gradients is well-known, the blowup in [1, 3, 6] is of the solution itself in $L^\infty$ and/or BV. Although similar phenomena had been observed earlier, the examples in [1, 3, 6] were new in that the solutions remain uniformly strictly hyperbolic as they explode. Also, there is nothing pathological about the flux $f(u)$ in these systems: it may be polynomial. This type of behavior is still poorly understood. In particular, it is not known if physical systems (e.g. compressible Euler) can exhibit similar behavior. The examples in [1, 3, 6] are quite special: starting from a 1-parameter family of $2 \times 2$ systems for $(u^1, u^2)$, say, and then adding a third, decoupled conservation law...
for the parameter $u^2$, one obtains a $3 \times 3$-system (1). The same type of systems also admit space-time periodic solutions [8]. However, it now appears that the analytic structure of these systems is too special to be of further interest. Still, these systems do exhibit an interesting geometric structure in $u$-space. This motivates an investigation of more general geometric issues for (1). E.g., can we prescribe curves in $u$-space and find systems (1) with these as rarefaction curves (integral curves of eigenfields)? Do the resulting systems possess extensions? Can we prescribe other types of geometric quantities for a system of conservation laws: shock curves, characteristic values, interaction coefficients, or combinations of such?

Before describing some recent results, we review the setup and findings in [4, 5]. In [4], we addressed the problem of determining the class of conservative systems (1) such that $Df(u)$ has a prescribed eigenframe $\mathfrak{R} = \{r_1, \ldots, r_n\}$. This was formulated as an algebraic-differential system, the “$\lambda(\mathfrak{R})$-system,” for the corresponding eigenvalues $\lambda^1, \ldots, \lambda^n$. This $\lambda$-system was then analyzed by appealing to various integrability theorems (Frobenius, Darboux, Cartan-Kähler), and a complete breakdown of the case $n = 3$ was given. In general, once the $\lambda$’s are found, the flux $f$ can be determined by successive integration of $n$ first order linear ODEs. It is a non-obvious fact that there are frames $\mathfrak{R}$ (see Example 2) that only admit trivial fluxes $f(u) = \lambda u + c$ ($\lambda \in \mathbb{R}, c \in \mathbb{R}^n$). To discuss the issue of extensions of (1) we recall the following definition and proposition [2]:

Definition 1.1. A smooth function $\eta : \Omega \to \mathbb{R}$ is an extension for (1) provided the map $u \mapsto \nabla \eta(u)Df(u)$ is the $u$-gradient of a scalar function $q : \Omega \to \mathbb{R}$. An extension $\eta$ is an entropy for (1) provided it is a convex function of the conserved quantities, i.e. the Hessian $D^2\eta(u)$ is positive semi-definite on $\Omega$.

Proposition 1. Let $\mathfrak{R} = \{r_i\}_{i=1}^n$ be the eigenframe of (1). Then $\eta : \Omega \to \mathbb{R}$ is an extension for (1) if and only if

$$\text{for each pair } 1 \leq i \neq j \leq n, \text{ either } \lambda^j = \lambda^i \text{ or } (D^2\eta)(r_i, r_j) = 0. \quad (2)$$

In [5], we analyzed the following question: Given a frame $\mathfrak{R}$, how large is the class of functions $\eta$ which satisfy the orthogonality condition:

$$\text{(D}^2\eta)(r_i, r_j) = 0. \quad (3)$$

Such functions provide all extensions for strictly (and a subset of all extensions for non-strictly) hyperbolic systems with eigenframe $\mathfrak{R}$. In [5], we reformulated the orthogonality condition (3) as an over-determined algebraic-differential system, the “$\beta(\mathfrak{R})$-system.” The unknowns in this $\beta$-system are the lengths of the frame-vectors $r_i$ as measured with respect to the inner-product $D^2\eta$. The Hessian $D^2\eta$ is determined by these lengths and the given frame $\mathfrak{R}$. In turn, the extensions $\eta$ can be determined, in principle, by integration of $n(n + 1)$ first order linear ODEs.

The results in [4, 5] provide information about how many conservative systems there are with a given eigenframe, and how many extensions these systems are equipped with. This information is given in terms of the number of free parameters or functions that determine a general solution of the $\lambda$- and $\beta$-systems. E.g., for the $3 \times 3$-case there are only two possibilities if the frame admits strictly hyperbolic systems: either the resulting systems are all rich$^1$ or they form a 1-parameter family (up to addition of a trivial flux). The latter class is more interesting one: it

$^1$We refer to [2] for this and other standard terminology.
covers systems with eigenframe that of the full Euler system, as well as the blowup examples described above.

While the analysis in [4, 5] gives the sizes of the solutions sets for the λ- and β-systems, it does not explicitly provide the fluxes and their extensions. As noted above this requires additional, and potentially challenging, ODE integrations.

In this paper, we address this practical drawback by deriving two overdetermined, but purely differential, systems \( f(\mathfrak{R}) \) and \( \eta(\mathfrak{R}) \), whose solutions provide directly the fluxes \( f(u) \) with eigenframe \( \mathfrak{R} \), and their associated extensions \( \eta(u) \). For the former we use a coordinate-free notion of Jacobian maps\(^2\) and for the latter we use a coordinate-free notion of Hessian metrics\(^3\). By solving the \( f(\mathfrak{R}) \)- and \( \eta(\mathfrak{R}) \)-systems we avoid the additional integrations required to reconstruct fluxes and extensions from the solutions of the λ- and β-systems. Not surprisingly, the degrees of freedom in the general solutions to the \( f(\mathfrak{R}) \)- and \( \eta(\mathfrak{R}) \)-systems are determined by the degrees of freedom in the general solution the \( \lambda(\mathfrak{R}) \)- and \( \beta(\mathfrak{R}) \)-systems, respectively. In fact, the latter systems provide exactly the compatibility conditions for the former overdetermined systems.

After reviewing some background on frames and connections in Section 2, we derive the \( f(\mathfrak{R}) \)- and \( \eta(\mathfrak{R}) \)-systems and establish their relationships with the \( \lambda(\mathfrak{R}) \)- and \( \beta(\mathfrak{R}) \)-systems in Sections 3 and 4. Section 5 provides some examples. Finally, in Section 6, we describe a recent result relating Hugoniot loci and eigenframes.

2. Frames and connections.

2.1. Frames and coframes. Let \( \Omega \subset \mathbb{R}^n \) be an open set. Denote by \( \mathcal{X}(\Omega) \) and \( \mathcal{X}^*(\Omega) \) the sets of smooth vector fields and smooth one-forms on \( \Omega \subset \mathbb{R}^n \), respectively. A frame \( \{r_1, \ldots, r_n\} \) is a set of smooth vector fields which are linearly independent at each point \( u \in \Omega \). A coframe \( \{\ell^1, \ldots, \ell^n\} \) is a set of smooth one-forms which are linearly independent at each point \( u \in \Omega \). If \( \ell^i(r_j) = \delta^i_j \) (Kronecker delta), then the coframe and frame are dual. Given coordinates \( (u^1, \ldots, u^n) \) on \( \Omega \), the associated coordinate frame is \( \{\partial/\partial u^1, \ldots, \partial/\partial u^n\} \), with dual coframe \( \{du^1, \ldots, du^n\} \).

For any frame \( \mathfrak{R} = \{r_i\}_{i=1}^n \), its structure coefficients \( c^k_{ij} \) are defined by
\[
[r_i, r_j] = \sum_{k=1}^n c^k_{ij} r_k,
\]
(5)

where the coefficients \( c^k_{ij} \) are the connection components relative to \( \mathfrak{R} \). The connection has a unique extension to an \( \mathbb{R} \)-bilinear map \( \mathcal{X}(\Omega) \times T(\Omega) \to T(\Omega) \), where

\[\nabla_{r_i} Y = \nabla_Y r_i, \quad \nabla_Y (h Y) = \nabla_Y h \cdot Y + h \cdot \nabla_Y Y.
\]

2.2. Connections. An affine connection \( \nabla \) on \( \Omega \) is an \( \mathbb{R} \)-bilinear map
\[
\mathcal{X}(\Omega) \times \mathcal{X}(\Omega) \to \mathcal{X}(\Omega), \quad (X, Y) \mapsto \nabla_X Y
\]
such that for any smooth function \( h \) on \( \Omega \)
\[
\nabla_{hX} Y = h \nabla_X Y, \quad \nabla_X (h Y) = (X h) Y + h \nabla_X Y.
\]
(4)

A connection is uniquely defined by prescribing it on a frame:
\[
\nabla_{r_i} r_j = \sum_{k=1}^n \Gamma^k_{ij} r_k,
\]
(5)

where the coefficients \( \Gamma^k_{ij} \) are the connection components relative to \( \mathfrak{R} \). The connection has a unique extension to an \( \mathbb{R} \)-bilinear map \( \mathcal{X}(\Omega) \times T(\Omega) \to T(\Omega) \), where

\(^2\)The notion of a Jacobian map was introduced in Remark 2.14 of [5] to provide an alternative, coordinate-free derivation of the λ-system. It is different from the differential of a map, which in coordinates, is also given by a Jacobian matrix.

\(^3\)The notion of a Hessian metric has been extensively studied [7]. In Remark 2.15 of [5], it was used to provide an alternative, coordinate-free derivation of the β-system.
\[ T(\Omega) \] is a set of all tensor products of \( X(\Omega) \) and \( \mathcal{X}(\Omega) \) (see [7]). We record the following properties of this extension:

\[ \nabla_X h = X(h), \quad \nabla_X (h\omega) = (Xh)\omega + f \nabla_X \omega, \quad \nabla_X(\omega(Y)) = [\nabla_X \omega](Y) + \omega(\nabla_X Y), \quad (6) \]

where \( h \) is a function, \( \omega \) is a one-form, and \( X, Y \) are vector-fields on \( \Omega \). Note that \( (5)-(6) \) imply

\[ \nabla_{\tilde{r}_i} \ell_j = -\sum_{k=1}^n \Gamma_{ij}^k \ell_k. \quad (7) \]

We now fix the connection \( \nabla \) on \( \Omega \) by requiring the state variables \( \{u_i\}_{i=1}^n \) of (1) to be affine coordinates on \( \Omega \), i.e. we impose

\[ \nabla \frac{\partial}{\partial u_i} = 0 \quad \text{for all } i, j = 1, \ldots, n. \quad (8) \]

This connection has the following two important properties:

\[ c'_{jk} = \Gamma_{jk}^i - \Gamma_{ij}^k \quad \text{(Symmetry)}, \]

\[ \nabla_X \circ \nabla_Y - \nabla_Y \circ \nabla_X = \nabla_{[X,Y]} \quad \text{(Flatness)}, \]

where \( X, Y \in \mathcal{X}(\Omega) \). Finally, we observe that if

\[ r_i = \sum_{k=1}^n R_k \frac{\partial}{\partial u_i} \quad \text{and} \quad \ell_i = \sum_{k=1}^n L_k \frac{\partial}{\partial u_i} \quad (i = 1, \ldots, n) \quad (11) \]

then (4) and (8) imply \( \Gamma_{ij}^k (u) = L_k (D_u R_j) R_i \).

3. The \( f(\mathcal{R}) \)-system. We fix a frame \( \mathcal{R} = \{r_i\}_{i=1}^n \) on \( \Omega \) and ask for the class of hyperbolic systems (1) whose flux-Jacobians \( Df \) has eigenframe \( \mathcal{R} \). This was formulated in [4] as an overdetermined algebraic-differential system for the eigenvalues \( \lambda^i(u) \). To derive a purely differential system, which directly yields the components of the flux \( f \), we apply:

**Proposition 2.** Let \( \{u_i\}_{i=1}^n \) be affine coordinates on a simply connected open subset \( \Omega \subset \mathbb{R}^n \), and \( f = (f^1, \ldots, f^n) : \Omega \to \mathbb{R}^n \) be a smooth map. Then \( r_i |u \) is an eigenvector of the Jacobian matrix \( Df(u) \) if and only if there is a function \( \lambda^i : \Omega \to \mathbb{R} \) with

\[ \nabla r_i, \quad \lambda^i r_i, \quad (12) \]

where the vector field \( F \) is defined by \( F := f^1 \frac{\partial}{\partial u^1} + \cdots + f^n \frac{\partial}{\partial u^n} \). If (12) is satisfied, then \( \lambda^i(u) \) is the corresponding eigenvalue of \( Df(u) \).

Hence, we seek a vector field \( F \in \mathcal{X}(\Omega) \) such that (12) is satisfied for \( i = 1, \ldots, n \). We write \( F = \sum_{i=1}^n \tilde{f}^i r_i \). Then, using (4), we find that (12) is equivalent to

\[ \sum_{j=1}^n \left( r_i(\tilde{f}^j) + \sum_{k=1}^n \tilde{f}^k \Gamma_{ik}^j \right) r_j = \lambda^i r_i. \quad (13) \]

Collecting the coefficients, we obtain the systems

\[ r_i(\tilde{f}^j) + \sum_{k=1}^n \tilde{f}^k \Gamma_{ik}^j = 0 \quad \text{for } i \neq j, \quad \text{and} \quad \lambda^i = r_i(\tilde{f}^i) + \sum_{k=1}^n \tilde{f}^k \Gamma_{ik}^i. \quad (14) \]

The \( f(\mathcal{R}) \)-system (14) consists of \( n^2 - n \) linear 1st order PDEs in \( n \) unknowns \( \{\tilde{f}^i\}_{i=1}^n \). For a given solution of (14), the flux \( f \) is determined by \( f = R(\tilde{f}^i), \ldots, \)
Substituting (20) into (19) and collecting coefficients of where

\[
\ell
\]

...directly

\[
\Omega \subset \mathbb{R}^n
\]

Let

Proposition 3.

\[
\beta
\]

...termined algebraic-differential system for the lengths \( \beta^i(u) \). To derive a purely differential system, which directly yields the extensions of (1), we apply:

**Proposition 3.** Let \( \{u_i\}_{i=1}^n \) be affine coordinates on a simply connected open subset \( \Omega \subset \mathbb{R}^n \), and \( \eta : \Omega \to \mathbb{R} \). Then \( \{r_i|u\}_{i=1}^n \) are orthogonal with respect to the inner product defined by the Hessian \( D^2_r \eta(u) \) if and only if there exist functions \( \beta^i : \Omega \to \mathbb{R} \), \( i = 1, \ldots, n \), such that

\[
\nabla_{r_i} \, d\eta = \beta^i \ell^i, \quad \text{for all } i = 1, \ldots, n,
\]

where \( \{\ell^1, \ldots, \ell^n\} \) is the dual coframe to \( \mathcal{R} \). If (19) is satisfied, then \( \beta^i \) is the length of the vector \( r_i|u \) relative to the Hessian inner product \( D^2_r \eta \).

Hence, we seek \( \eta : \Omega \to \mathbb{R}^n \) such that (19) is satisfied for some functions \( \beta^i \). Since \( d\eta = \sum_j r_j(\eta) \ell^j \), and using (6)-(7), we obtain

\[
\nabla_{r_i} \, d\eta = \sum_j \left( r_i(r_j(\eta)) - \sum_k r_k(\eta) \Gamma^k_{ij} \right) \ell^j.
\]

Substituting (20) into (19) and collecting coefficients of \( \ell^i \), we obtain the systems

\[
r_i(r_j(\eta)) - \sum_k \Gamma^k_{ij} r_k(\eta) = 0 \quad \text{for } i \neq j, \quad \text{and} \quad \beta^i = r_i(\eta) - \sum_k \Gamma^k_{ii} r_k(\eta).
\]

---

\(^4\)See [4] for Frobenius type systems and Frobenius' Theorem adapted to the present context.
The $\eta(\mathfrak{R})$-system (21)$_1$ consists of $n(n - 1)$ 2nd order PDEs for the single unknown $\eta$. For a given solution of (21)$_1$, the lengths $\beta^i$ of the vectors $r_i|_a$ relative to the inner product $D^2_a \eta$ are given by (21)$_2$.

We next derive necessary and sufficient conditions for the existence of solutions to the overdetermined $\eta(\mathfrak{R})$-system. Again, the flatness condition (10) implies

$$\nabla_{r_i} \nabla_{r_j} d\eta - \nabla_{r_j} \nabla_{r_j} d\eta = \nabla_{[r_i,r_j]} d\eta = \sum_k c^k_{ij} \nabla_{r_i} d\eta. \quad (i \neq j) \quad (22)$$

Substitution of (19) into (22) gives

$$\nabla_{r_i}(\beta_j \ell^j) - \nabla_{r_j}(\beta_i \ell^i) = \sum_k c^k_{ij} \beta_k \ell^k. \quad (23)$$

By expanding (23) using (7), and then collecting the coefficients of $\ell^i$, we obtain the following algebraic-differential system for the $\beta^i$:

$$r_i(\beta_j) = \beta_j(\Gamma^i_{ij} + c^i_{ij}) + \beta^j \Gamma^i_{jj}, \quad \text{for } i \neq j, \quad (24)$$

$$\beta_k c_{ij} + \beta_j \Gamma^i_{ik} - \beta_i \Gamma^j_{jk} = 0 \quad \text{for } i < j, i, j \neq k. \quad (25)$$

This is exactly the $\beta(\mathfrak{R})$-system derived in [5]. Thus, a necessary condition for the existence of a solution to the $\eta(\mathfrak{R})$-system is the existence of a solution to the $\beta(\mathfrak{R})$-system. Vice versa, given a solution $\{\beta^i\}^n_{i=1}$ of the $\beta(\mathfrak{R})$-system, we introduce $\eta_i := r_i(\eta)$ and rewrite (21)$_1, 2$ as a first order Frobenius type system in the $n$ unknowns $\eta_i$:

$$r_i(\eta_j) - \sum_k \ell^k c_{ij} \eta_k = 0 \quad \text{for } i \neq j, \quad \text{and } \quad r_i(\eta_j) - \sum_k \Gamma^k_{ij} \eta_k = \beta^i. \quad (26)$$

A direct computation shows that this is a compatible system. By Frobenius’ Theorem, we conclude that any solution $\{\beta^i\}^n_{i=1}$ of the $\beta(\mathfrak{R})$-system, together with a choice of values $\eta_1(\bar{u}), \ldots, \eta_n(\bar{u})$ at some point $\bar{u} \in \Omega$, provide a unique solution of the system (26) satisfying (26)$_2$. Finally, if $\{\eta_i\}^n_{i=1}$ solve (26)$_1, 2$, then $(\eta_1, \ldots, \eta_n) := (\eta_1, \ldots, \eta_n) R^{-1} (R$ given by (11)), provide a solution $\eta$ of the combined system (21)$_1, 2$ via $\eta_i = \frac{\partial \eta}{\partial \sigma_i}$.

5. Examples. We include some representative examples where we use Maple code developed by the authors$^5$ to calculate explicit fluxes and their extensions from a given eigenframe. Cf. Examples 6.5, 6.6, 6.8 and 6.13 in [5].

Example 1. Consider a rich orthogonal frame: $r_1 = (u^1, u^2, 0)^T$, $r_2 = (-u^2, u^1, 0)^T$, $r_3 = (0, 0, 1)^T$. Solving the $f(\mathfrak{R})$-system, we obtain the family of fluxes depending on three arbitrary functions of one-variable:

$$f(u) = \begin{pmatrix}
-u^1 \int_{a}^{u^1} F_2 \left(\frac{\sqrt{(u^1)^2 + (u^2)^2}}{a} - \frac{u^2}{a}\right) du + u^1 F_1 \left(\frac{(u^1)^2}{a^2} + \frac{(u^2)^2}{a^2}\right) - F_1 \left(\frac{u^2}{a}\right) u^2 \\
-u^2 \int_{a}^{u^2} F_3 \left(\frac{\sqrt{(u^1)^2 + (u^2)^2}}{a} - \frac{u^1}{a}\right) du + u^2 F_1 \left(\frac{(u^1)^2}{a^2} + \frac{(u^2)^2}{a^2}\right) - F_1 \left(\frac{u^2}{a}\right) u^2
\end{pmatrix}$$

Solving the $\eta(\mathfrak{R})$-system, we obtain the family of extensions also depending on three arbitrary functions of one-variable:

$$\eta(u) = \int_{a}^{u^1} G_2 \left(\frac{\sqrt{(u^1)^2 + (u^2)^2}}{a} - \frac{u^2}{a}\right) du + G_1 \left(\frac{(u^1)^2}{a^2} + \frac{(u^2)^2}{a^2}\right) + G_3(u^2). \quad (27)$$

$^5$http://www.math.ncsu.edu/~iakogan/symbolic/geometry_of_conservation_laws.html
There are strictly hyperbolic systems in this class, and any extension of such a system is given by (27). For this frame, it is much harder (and beyond the capability of MAPLE) to obtain fluxes and extensions if we first solve the $\lambda(\mathcal{R})$ and $\beta(\mathcal{R})$-systems, and then try to find $f$ and $\eta$ by solving ODE’s. It is not difficult, however, to find $\lambda$’s and $\beta$’s, after we found $f$ and $\eta$, using, for instance, (14) and (21).2

**Example 2.** For the rich frame $r_1 = (u^1, u^2, 0)^T$, $r_2 = (-u^2, u^1, 0)^T$, $r_3 = (-u^2, u^1, 1)^T$, we solve the $f(\mathcal{R})$-system and obtain only trivial fluxes:

$$f(u) = c_1 (u^1, u^2, u^3)^T + (c_2, c_3, c_4)^T,$$

with $\lambda_1 = \lambda_2 = \lambda_3 = c_1 \in \mathbb{R}$. The $\eta(\mathcal{R})$-system provides a family of extensions that satisfy the orthogonality condition (3): $\eta(u) = a_1 u^1 + a_2 u^2 + G(u^3)$. Obviously, any function $\eta(u)$ satisfies the full condition (2) and, hence, is an extension.

**Example 3.** For the non-rich frame $r_1 = (-1, 0, 1 - u^2)^T$, $r_2 = \left(\frac{u^3}{(u^2)^2 - 1}, -1, u^1\right)^T$, $r_3 = (1, 0, 1 - u^2)^T$, we solve the $f(\mathcal{R})$-system and obtain a family of fluxes:

$$f(u) = c_1 \left(\begin{array}{c}
\frac{u^3 - u^1 u^2 - u^1}{2} \\
\frac{1}{2} (u^2)^2 - u^2 \\
(1 + u^2) (u^1 - u^1 u^2 - u^3)
\end{array}\right) + c_2 \left(\begin{array}{c}
u^1 \\
u^2 \\
u^3
\end{array}\right) + \left(\begin{array}{c}
c_3 \\
c_4 \\
c_5
\end{array}\right),$$

which is, up to a trivial flux, depends on one constant. Using (14), we find that $\lambda_1 = c_2 - 2 c_1$, $\lambda_2 = c_2 + c_1 (u^2 - 1)$ and $\lambda_3 = c_2$. Observe that there are strictly hyperbolic system with this frame. Solving the $\eta(\mathcal{R})$-system, we obtain a family of extensions

$$\eta(u) = a_1 u^1 + a_3 u^3 + G(u^2),$$

depending on one arbitrary function. These are all the possible extensions of a strictly hyperbolic system with the given frame. From (21), we find that $\beta_1 = \beta_3 \equiv 0$ and $\beta_2 = G''(u^2)$.

**Example 4.** For the non-rich frame $r_1 = (1, u^2, 0)^T$, $r_2 = (u^3, 1, 0)^T$, $r_3 = (0, 0, 1)^T$, we solve the $f(\mathcal{R})$-system and obtain a family of fluxes:

$$f(u) = c_1 \left(u^1, u^2, F(u^3)\right)^T + (c_2, c_3, c_4)^T,$$

which is, up to a trivial flux, depends on one arbitrary function of one variable. Using (14), we find that $\lambda_1 = \lambda_2 = c_1$ and $\lambda_3 = F'(\bar{u}_3)$. Therefore, there are no strictly hyperbolic system with this frame. Solving the $\eta(\mathcal{R})$-system, we obtain a family of extensions:

$$\eta(u) = a_1 u^1 + a_2 u^2 + G(u^3),$$

which satisfy the orthogonality condition (3). Note, however, that there may be other extensions that satisfy condition (2) due to equality of eigenvalues $\lambda_1 = \lambda_2$.

6. **On Hugoniot loci and eigencurves for $3 \times 3$-systems.** Together with [4, 5], the analysis above provides detailed information about systems (1) with a given eigenframe: the $f(\mathcal{R})$- and $\eta(\mathcal{R})$-systems provide the fluxes and entropies, while the $\lambda(\mathcal{R})$- and $\beta(\mathcal{R})$-systems tell us how many such there are. Note that prescribing $\mathcal{R}$ amounts to prescribing rarefaction curves, and these make up one part of the wave curves. The other part is the Hugoniot locus $\mathcal{H}$. For a given flux $f$ and state $\bar{u} \in \Omega$,

$$\mathcal{H}_{f, \bar{u}} := \{ u \in \Omega \mid \exists s \in \mathcal{R} \text{ such that } s(u - \bar{u}) = f(u) - f(\bar{u})\}.$$

To complement our analysis above, it would seem natural to ask: “Given a family of curves in $\Omega$, what is the set if systems (1) with these as Hugoniot loci?” However,
this is not a well-formulated problem: due to the symmetry relation “\( \bar{v} \in \mathcal{H}_{f,\bar{u}} \Leftrightarrow \bar{u} \in \mathcal{H}_{f,v} \)” it is hard to prescribe a family of curves that could be Hugoniot loci. Unless, of course, one starts with the Hugoniot loci of a given system \((1)\). Thus, the natural complement to our earlier analysis is to ask: “Given a hyperbolic system \((1)\), what other systems have the same Hugoniot loci?”

We shall describe a partial result for \(3 \times 3\)-systems in this direction. First, there are simple examples showing that conservation laws with the same eigenframe does not, in general, share the same Hugoniot loci. This is not surprising since \(\mathcal{H}_{f,\bar{u}}\) depends on all of \(f\), and not only its eigenframe. However, we have the following:

**Proposition 4.** Consider a strictly hyperbolic, non-rich \(3 \times 3\)-system \((1)\). Then all systems with the same eigenframe also have the same Hugoniot loci.

**Proof.** Let \(\mathcal{R}\) be the eigenframe and \(g\) the flux corresponding to our system. The case of a strictly hyperbolic non-rich system only occurs in the IIa subcase from [4]. The proof of Theorem 3.2 in that paper invokes the Frobenius Integrability Theorem to show that for any choice of the initial conditions \(\lambda^2(\bar{u})\) and \(\lambda^3(\bar{u})\) for two of the eigenvalues, there is a unique Jacobian with the eigenframe \(\mathcal{R}\). Let \(A\) be the Jacobian with eigenvalues satisfying initial condition \(\lambda^2(\bar{u}) = 1\) and \(\lambda^3(\bar{u}) = 0\). Let \(a_2\) and \(a_3\) be the eigenvalues of \(Dg(\bar{u})\) corresponding to \(r_2\) and \(r_3\). A direct computation show that \(B = (a_2 - a_3)A + a_3 I\) is a Jacobian with with eigenframe \(\mathcal{R}\) and eigenvalues satisfying initial conditions \(\lambda^2(\bar{u}) = a_2\) and \(\lambda^3(\bar{u}) = a_2\). By the proof of Theorem 3.2 in [4] it is the unique such Jacobian.

Thus the flux \(g\) must be of the form \(g(u) := (a_2 - a_3) f(u) + a_3 \bar{u} + \bar{v}\), where \(f(u)\) is a flux with the Jacobian \(A\), and \(\bar{v} \in \mathbb{R}^n\) is a fixed vector. Note that if \(f(u) - f(v) = s(u - v)\), then \(g(u) - g(v) = [s(a_2 - a_3)](u - v)\) and thus the Hugoniot loci for \(f\) are Hugoniot loci for \(g\). Since \(a_2 \neq a_3\), we can also write \(f(u) = b_1 g(u) + b_2 \bar{u} + \bar{v}'\) for some \(b_1, b_2,\) and \(\bar{v}'\), and determine by the same argument that also the Hugoniot loci for \(g\) are Hugoniot loci for \(f\). Thus all systems with this eigenframe have the same Hugoniot loci.

To appreciate the relevance of this result, we recall (see Introduction) that given frame in \(\mathbb{R}^3\), there are only two possibilities for it to admit strictly hyperbolic systems: either these are all rich, or they form a 1-parameter family (up to a trivial flux). Thus, Proposition 4 applies to all systems in the latter 1-parameter class. In particular, it applies to any system with eigenframe that of the full Euler system, or that of the blowup examples in [1, 3, 6].

*E-mail address*: mike.benfield@gmail.com  
*E-mail address*: hkjenssen@gmail.com  
*E-mail address*: iakogan@math.ncsu.edu
CONVERGENCE OF A FINITE DIFFERENCE SCHEME FOR 2 × 2 KEYFITZ-KRANZER SYSTEM

Ujjwal Koley
Institut für Mathematik
Julius-Maximilians-Universität Würzburg
Campus Hubland Nord, Emil-Fischer-Strae 30
97074, Würzburg, Germany

Nils Henrik Risebro
Centre of Mathematics for Applications (CMA)
University of Oslo, P.O. Box 1053
Blindern, N-0316, Oslo, Norway

Abstract. We are concerned with the convergence of a numerical scheme for the initial value problem associated to the $2 \times 2$ Keyfitz-Kranzer system of equations. In this paper we prove the convergence of a finite difference scheme to a weak solution.

1. Introduction. In this paper, we are concerned with a symmetrically hyperbolic system of two equations

\[
\begin{align*}
  u_t + (u\phi(r))_x &= 0, \\
  v_t + (v\phi(r))_x &= 0,
\end{align*}
\]

(1)

with initial data

\[
(u(x,0), v(x,0)) = (u_0(x), v_0(x)), \quad x \in \mathbb{R},
\]

(2)

where $r(x,t) = \sqrt{u^2(x,t) + v^2(x,t)}$, $\Pi_T = \mathbb{R} \times (0, T)$ with $T > 0$ fixed, and $u, v : \Pi_T \to \mathbb{R}$ are the unknown functions. Regarding the function $\phi$, the basic assumption is that $\phi : \mathbb{R} \to \mathbb{R}$ is a given (sufficiently smooth) scalar function (see Section 2 for precise assumptions). Systems of this type was first considered in [2] and later on by several other authors, as a prototypical example of a non-strict hyperbolic system. Note that (1) is a non-strict hyperbolic system with first characteristic field is always linearly degenerate and the second characteristic field is either genuinely nonlinear or linearly degenerate, depending on the behavior of $\phi$.

Due to the nonlinearity, discontinuities in the solution may appear independently of the smoothness of the initial data and weak solutions must be sought. A weak solution is defined as follows:

Definition 1.1. We call the pair $(u,v)$ a weak solution of the Cauchy problem (1)–(2) if

(a) $u$ and $v$ are in $L^\infty(\Pi_T)$. 

2000 Mathematics Subject Classification. Primary: 35D30, 35L40; Secondary: 35Q35.

Key words and phrases. Keyfitz-Kranzer system, finite difference scheme, convergence, existence, compensated compactness.

759
(b) $u$ and $v$ satisfy (1) in the sense of distributions on $\Pi_T$, i.e., the following identities,
\begin{align}
\int_{\Pi_T} u \psi_t + u \phi(r) \psi_x \, dx dt + \int_{\mathbb{R}} u_0 \phi(x,0) \, dx &= 0, \\
\int_{\Pi_T} v \psi_t + v \phi(r) \psi_x \, dx dt + \int_{\mathbb{R}} v_0 \phi(x,0) \, dx &= 0,
\end{align}
hold for each smooth test function $\psi$ with compact support in $\Pi_T$.

In this paper, we propose a upwind semi discrete finite difference scheme and prove the convergence of the approximate solution to the weak solution of (1). In what follows, we first prove the strong convergence of approximate solution $r_{\Delta x} = \sqrt{u_{\Delta x}^2 + v_{\Delta x}^2}$ using compensated compactness argument [1, 4]. Next, we prove a BV estimate of $\varphi_{\Delta x} := \tan^{-1}(\frac{u_{\Delta x}}{v_{\Delta x}})$. Then Helly’s theorem combined with the strong convergence of $r_{\Delta x}$ gives the required strong convergence of $u_{\Delta x}$ and $v_{\Delta x}$.

2. Mathematical Framework. In this section we present some mathematical tools that we shall use in the analysis. To start with the basic assumptions on the initial data and the function $\phi(r)$, we assume that $\phi$ is a twice differentiable function $\phi : [0, \infty) \to [0, \infty)$ so that
(a) $\phi(r) > 0$ and $\phi'(r) \geq 0$ for all relevant $r$;
(b) $\phi(r), \phi'(r)$ and $\phi''(r)$ are bounded for all relevant $r$;
(c) $\sqrt{u_0^2 + v_0^2} \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$.

Note that we shall assume above assumptions throughout the paper. We also use the following compensated compactness result.

**Theorem 2.1.** Let $\Omega \subset \mathbb{R} \times \mathbb{R}^+$ be a bounded open set, and assume that $\{u^\varepsilon\}$ is a sequence of uniformly bounded functions such that $|u^\varepsilon| \leq M$ for all $\varepsilon$. Also assume that $f : [-M, M] \to \mathbb{R}$ is a twice differentiable function. Let $u^\varepsilon \rightharpoonup u$ and $f(u^\varepsilon) \rightharpoonup v$, and set
\begin{align}
(\eta_1(s), q_1(s)) &= (s - k, f(s) - f(k)), \\
(\eta_2(s), q_2(s)) &= \left(f(s) - f(k), \int_{s-k}^s (f'(\theta))^2 \, d\theta \right),
\end{align}
where $k$ is an arbitrary constant. If
\[ \eta_1(u^\varepsilon)_t + q_1(u^\varepsilon)_x \]
is in a compact set of $H^{-1}_{\text{loc}}(\Omega)$ for $i = 1, 2$,
then
\begin{enumerate}
\item[(a)] $v = f(u)$, a.e. $(x,t)$,
\item[(b)] $u^\varepsilon \to u$, a.e. $(x,t)$ if $\text{meas} \{u \mid f''(u) = 0 \} = 0$.
\end{enumerate}

For a proof of this theorem, see the monograph of Lu [4]. The following compactness interpolation result (known as Murat’s lemma [3]) is useful in obtaining the $H^{-1}_{\text{loc}}$ compactness needed in Theorem 2.1.

**Lemma 2.2.** Let $\Omega$ be a bounded open subset of $\mathbb{R}^2$. Suppose that the sequence \{L$^\varepsilon$\}$_{\varepsilon > 0}$ of distributions is bounded in $W^{-1,\infty}(\Omega)$. Suppose also that
\[ L_\varepsilon = L_{1,\varepsilon} + L_{2,\varepsilon}, \]
where $\{L_{1,\varepsilon}\}$_{\varepsilon > 0} is in a compact subset of $H^{-1}(\Omega)$ and $\{L_{2,\varepsilon}\}$_{\varepsilon > 0} is in a bounded subset of $M_{\text{loc}}(\Omega)$. Then $\{L_\varepsilon\}$_{\varepsilon > 0} is in a compact subset of $H^{-1}_{\text{loc}}(\Omega)$. 

3. Semi Discrete Finite Difference Scheme. We start by introducing the necessary notations. Given \( \Delta x > 0 \), we set \( x_j = j \Delta x \) and \( x_{j+1/2} = x_j + \Delta x/2 \) for \( j \in \mathbb{Z} \) and for any function \( u = u(x) \), we define \( u_j = u(x_j) \). Let \( D_\pm \) denote the discrete forward and backward differences, i.e.,

\[
D_\pm u_j = \mp \frac{u_j - u_{j \pm 1}}{\Delta x}.
\]

To a sequence \( \{w_j\}_{j \in \mathbb{Z}} \) we associate the function \( w_{\Delta x} \) defined by

\[
w_{\Delta x}(x) = \sum_{j \in \mathbb{Z}} w_j \mathbb{I}_I(x),
\]

where \( \mathbb{I}_A \) denotes the characteristic function of the set \( A \). We will use following standard notations:

\[
\|u_{\Delta x}\|_{L^\infty(\mathbb{R})} = \sup_{j \in \mathbb{Z}} |u_j|, \quad \|u_{\Delta x}\|_{L_1(\mathbb{R})} = \Delta x \sum_{j \in \mathbb{Z}} |u_j|,
\]

\[
\|u_{\Delta x}\|_{L^2(\mathbb{R})} = \left( \Delta x \sum_{j \in \mathbb{Z}} |u_j|^2 \right)^{1/2}, \quad |u_{\Delta x}|_{BV(\mathbb{R})} = \sum_{j \in \mathbb{Z}} |u_j - u_{j-1}|.
\]

Now let \( \{u_j(t)\}_{j \in \mathbb{Z}} \) and \( \{v_j(t)\}_{j \in \mathbb{Z}} \) satisfy the (infinite) system of ordinary differential equations,

\[
\begin{aligned}
    u'_j + D_- (\phi(r_j)u_j) &= 0, \\
    v'_j + D_- (\phi(r_j)v_j) &= 0,
\end{aligned}
\]

with initial values

\[
u_j(0) = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} u_0(x) \, dx \quad \text{and} \quad v_j(0) = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} v_0(x) \, dx.
\]

Here \( r_j = \sqrt{u_j^2 + v_j^2} \). It is natural to view (5) as an ordinary differential equation in \( L^2(\mathbb{R}) \times L^2(\mathbb{R}) \). It is easy to show that the right hand side of (5) is Lipschitz continuous in \( L^2(\mathbb{R}) \times L^2(\mathbb{R}) \), which essentially gives the local (in time) existence and uniqueness of differentiable solutions. The next lemma shows that the \( L^2 \) norm remains bounded if it is bounded initially, so the solution can be defined up to any time.

Lemma 3.1. Let \( \{u_j(t)\}, \{v_j(t)\} \) be defined by (5), and let \( r_j = \sqrt{u_j^2 + v_j^2} \). Then

\[
\begin{aligned}
    \|r_{\Delta x}(t)\|_{L^1(\mathbb{R})} &\leq \|r_{\Delta x}(0)\|_{L^1(\mathbb{R})}, \\
    \|r_{\Delta x}(t)\|_{L^2(\mathbb{R})} &\leq \|r_{\Delta x}(0)\|_{L^2(\mathbb{R})}.
\end{aligned}
\]

Furthermore, there is a constant \( C \), independent of \( \Delta x \) and \( T \), such that

\[
\int_0^T \left( \sum_j \int_{r_j-\Delta x}^{r_j} (r_j^2 - s^2) \phi'(s) \, ds + \Delta x \sum_j \phi_{j-1} \Delta x \left( (D_+ u_j)^2 + (D_+ v_j)^2 \right) \right) \, dt \leq C.
\]

Proof. Set \( U = (u, v) \) and observe that \( r_j = |U_j| \). We can rewrite the system (5) as

\[
(U_j)_t + D_- (U_j \phi(r_j)) = 0.
\]
Let \( \eta = \eta(U) \) be a differentiable function \( \eta : \mathbb{R}^2 \to \mathbb{R} \), take the inner product of (7) with \( \nabla \eta(U_j) \) to get

\[
\frac{d}{dt} \eta(U_j) + D_- (\phi_j \eta(U_j)) + [(\nabla \eta(U_j), U_j) - \eta(U_j)] D_- \phi_j + \phi_j - \frac{\Delta x}{2} d^2 \eta_{j-1/2} (D_- U_j, D_- U_j) = 0. \tag{8}
\]

Here \( \phi_j = \phi(r_j) \), and \( d^2 \eta \) denotes the Hessian matrix of \( \eta \), so that

\[
d^2 \eta_{j-1/2} = d^2 \eta(U_j-1/2)
\]

for some \( U_j-1/2 \) between \( U_j \) and \( U_{j-1} \). By a limiting argument, the function \( \eta(U) = |U| \) can be used. This function is convex, i.e., \( d^2 \eta(U) \geq 0 \). This means that

\[
\frac{d}{dt} r_j + D_- (r_j \phi_j) \leq 0. \tag{9}
\]

Multiplying by \( \Delta x \) and summing over \( j \) we get

\[
\| r_{\Delta x}(t) \|_{L^1(\mathbb{R})} \leq \| U_0 \|_{L^1(\mathbb{R})}. \tag{10}
\]

Furthermore, choosing \( \eta(U) = |U|^2 \) in (8) we get

\[
\frac{d}{dt} r_j^2(t) + D_- (r_j^2 \phi_j) + r_j^2 D_- \phi_j + \phi_j - \frac{\Delta x}{2} (D_- U_j)^2 = 0.
\]

We have that

\[
D_- (r_j^2 \phi_j) + r_j^2 D_- \phi_j = \frac{2}{\Delta x} \int_{r_{j-1}}^{r_j} s \phi(s) + s^2 \phi'(s) \, ds + \frac{1}{\Delta x} \int_{r_{j-1}}^{r_j} (r_j^2 - s^2) \phi'(s) \, ds,
\]

where

\[
g(r) = 2 \int_0^r s \phi(s) + s^2 \phi'(s) \, ds. \tag{11}
\]

Using this we find that

\[
\| r_{\Delta x}(t) \|_{L^2(\mathbb{R})} \leq \| U_0 \|_{L^2(\mathbb{R})}, \tag{12}
\]

since, by the assumption that \( \phi' \geq 0 \),

\[
\int_{r_{j-1}}^{r_j} (r_j^2 - s^2) \phi'(s) \, ds \geq 0.
\]

Hence \( \|(u_{\Delta x}(t), v_{\Delta x}(t))\|_{L^2(\mathbb{R})^2} \) is bounded independently of \( \Delta x \) and \( t \). Therefore, the exists a differentiable solution \( (u_{\Delta x}(t), v_{\Delta x}(t)) \) to (5) for all \( t > 0 \). Furthermore, we have the bound

\[
\int_0^T \left( \sum_j \int_{r_{j-1}}^{r_j} (r_j^2 - s^2) \phi'(s) \, ds + \Delta x \sum_j \phi_j - 1 \Delta x |D_- U_j|^2 \right) \, dt \leq C,
\]

for some constant \( C \) which is independent of \( t \) and \( \Delta x \). \( \square \)

**Lemma 3.2.** If there is a constant \( R \) such that \( r_j(0) \leq R \) for all \( j \), then \( r_j(t) \leq R \) for all \( j \) and \( t > 0 \).

If \( 0 < u_j(0) \) and \( 0 < v_j(0) \) for all \( j \), and there is a constant \( C > 0 \) such that

\[
\frac{1}{C} \leq \frac{u_j(0)}{v_j(0)} \leq C,
\]
then
\[
\frac{1}{C} \leq \frac{u_j(t)}{v_j(t)} \leq C
\]
for all \( j \) and \( t > 0 \).

**Proof.** If \( j_0 \) is such that \( r_{j_0}(t_0) \geq r_{j_0-1}(t_0) \), then \( D_- f(r_{j_0}(t_0)) \geq 0 \) with \( f(r) = r \phi(r) \), since \( f \) is non-decreasing. Hence, from (9), we see that \( r'_{j_0}(t_0) \leq 0 \). This proves the first statement of the lemma.

To prove the second statement, we first show that if \( u_j(0) > 0 \), then \( u_j(t) \geq 0 \), and if \( u_{j_0}(t_0) = 0 \) for some \( t_0 > 0 \) and \( j_0 \), then \( u_j(t) = 0 \) for all \( j \leq j_0 \) and all \( t \geq t_0 \). A similar statement holds for \( v_j \). To see this, note that
\[
u_j' + u_j D_- \phi_j + \phi_{j-1} D_- u_j = 0.
\]
Assume that for some \( t_0 \) and \( j_0 \), \( u_{j_0}(t_0) = 0 \) and \( u_{j_0}(t) \geq 0 \) for \( t < t_0 \). Then \( u'_{j_0}(t_0) \leq 0 \). If \( u_{j_0-1}(t_0) > 0 \), this leads to a contradiction, hence \( u_{j_0-1}(t_0) = 0 \). By repeating the argument we get that that \( u_j(t_0) = 0 \) for all \( j < j_0 \). If both \( u_j \) and \( u_{j-1} \) are zero, then \( u'_j(t) = 0 \), hence if \( u_j(t_0) = 0 \), \( u_j(t) = 0 \) for all \( t > t_0 \). A similar statement holds for \( v_j \). This means that if \( r_{j_0}(t_0) = 0 \), both \( u_{j_0}(t_0) \) and \( v_{j_0}(t_0) \) are zero, hence \( r_j(t) = 0 \) for \( j \leq j_0 \) and \( t \geq t_0 \).

Let for the moment \( \varphi_j \) be defined by
\[
\varphi_j = \begin{cases} 
\tan^{-1} \left( \frac{v_j}{u_j} \right) & \text{if } r_j > 0, \\
\varphi_{j+1} & \text{if } r_j = 0.
\end{cases}
\]
(13)

By the previous observation, we know that \( 0 \leq \varphi_j \leq \pi/2 \). Now if \( r_j > 0 \),
\[
\varphi'_j(t) = \frac{1}{1 + (v_j/u_j)^2} \left( \frac{v_j}{u_j} \right)' = -\frac{u_j u_{j-1} \phi_{j-1}}{r_j^2} D_- \left( \frac{v_j}{u_j} \right).
\]
Therefore \( \varphi_j \) satisfies the equation
\[
\varphi'_j + \frac{u_j u_{j-1} \phi_{j-1}}{r_j^2} D_- (\tan(\varphi_j)) = 0.
\]
(14)

This equation holds for any \( j \) where \( r_j > 0 \), if \( r_{j_0} = 0 \) for some \( j_0 \), then we define \( \varphi_j(t) = \varphi_{j_0+1}(t) \) for all \( j \leq j_0 \).

We have that \( \tan \) is an increasing function, and \( (u_j u_{j-1} \phi_{j-1})/r_j^2 \geq 0 \) if \( r_j > 0 \). Therefore, if \( \varphi_j(t) > \varphi_{j-1}(t) \), then \( \varphi'_j(t) \leq 0 \). Similarly if \( \varphi_j(t) < \varphi_{j-1}(t) \), then \( \varphi'_j(t) \geq 0 \). The assumption on the initial data implies that
\[
0 < \inf_{j} \varphi_j(0) \leq \varphi_j(t) \leq \sup_{j} \varphi_j(0) < \pi/2.
\]
Incidentally, this shows that if \( u_j(t) = 0 \), then \( v_j(t) = 0 \) and vice versa.

Let now \( \eta_i(r) \) and \( q_i(r) \) be given by (4) for \( i = 1, 2 \). We then have that
\[
\frac{d}{dt} \eta_i(r_j) + D_- (q_i(r_j)) + e_{1,i} = 0,
\]
(15)
where
\[
f(r) = r \phi(r), \quad q_1(r) = f(r) - f(k) \quad \text{and}
\]
\[
e_{1,j} = \phi_{j-1} \Delta x (D_- U_j)^T \frac{1}{r_{j-1/2}} \left( I - \frac{U_{j-1/2} \otimes U_{j-1/2}}{r_{j-1/2}^2} \right) (D_- U_j).
\]
Lemma 3.3. We have that \( e_i \in \mathcal{M}_{\text{loc}}(\Pi_T) \) for \( i = 1, 2, 3 \).

Proof. The result follows since \( \phi(r) > 0 \) and

\[
\int_0^T \Delta x \sum_j \Delta x |D_r| \leq C,
\]

\[
\square
\]

Lemma 3.4. Let \((u_{\Delta x}, v_{\Delta x})\) be generated by the scheme (5) and let \( r_{\Delta x} \) be defined by \( r_{\Delta x} = \sqrt{u_{\Delta x}^2 + v_{\Delta x}^2} \). Then

\[
\{ \eta_i (r_{\Delta x}) t + q_i (r_{\Delta x}) \}_{\Delta x > 0} \text{ is compact in } H_{\text{loc}}^{-1}(\Pi_T),
\]

where \( \eta_i \) and \( q_i \) are given by (4).

Proof. Let \( i = 1 \) or \( i = 2 \), and \( \psi \) is a test function in \( H_{\text{loc}}^{1}(\Pi_T) \). we define

\[
\langle \mathcal{L}_i, \psi \rangle = \langle \eta_i (r_{\Delta x}) t + q_i (r_{\Delta x}) x, \psi \rangle
\]

\[
= \int_0^T \sum_j \int_{x_j-1/2}^{x_j+1/2} \frac{d}{dt} \eta_i(r_j(t)) \psi(x, t) + D_q_i(r_j(t)) \psi(x_{j-1/2}, t) dx \, dt
\]

\[
= \int_0^T \sum_j \int_{x_j-1/2}^{x_j+1/2} \psi(x_{j-1/2}, t) D_q_i(r_j(t)) dx \, dt
\]

\[
+ \int_0^T \sum_j \int_{x_j-1/2}^{x_j+1/2} (\psi(x_{j-1/2}, t) - \psi(x, t)) D_q_i(r_j) dx \, dt
\]

\[
\langle \mathcal{L}_i, \psi \rangle + \langle \mathcal{L}_2, \psi \rangle.
\]

By (15), (16) and Lemma 3.3 we know that \( \mathcal{L}_{i,1} \in \mathcal{M}_{\text{loc}}(\Omega) \). Regarding \( \mathcal{L}_{i,2} \) we have

\[
|\langle \mathcal{L}_{2, i}, \psi \rangle| = \left| \int_0^T \sum_j \int_{x_j-1/2}^{x_j+1/2} \int_{x_j-1/2}^{x_j+1/2} \psi_x(y, t) dy \, D_q_i(r_j(t)) \, dx \, dt \right|
\]

\[
\leq \int_0^T \sum_j \Delta x^{3/2} \left( \int_{x_j-1/2}^{x_j+1/2} (\psi_x(x, t))^2 \, dx \right)^{1/2} \|q_i\|_{L^\infty} \, |D_r| \, dt
\]

\[
\leq C \sqrt{\Delta x} \|\psi\|_{H^1(\Pi_T)}.
\]
Therefore the above estimate shows that $L_{2,i}$ is compact in $H^{-1}(\Pi_T)$. By Lemma 2.2, we conclude the sequence $\{\eta_i(r_{\Delta x})t + q_i(r_{\Delta x})\}_{\Delta x>0}$ is compact in $H^{-1}_{loc}(\Pi_T)$.

**Lemma 3.5.** If

$$\text{meas}\left\{ r \left| 2\phi'(r) + r\phi''(r) = 0 \right. \right\} = 0,$$

then there is a subsequence of $\{\Delta x\}$ and a function $r$ such that $r_{\Delta x} \rightarrow r$ a.e. $(x,t) \in \Pi_T$. We have that $r \in L^\infty([0,T]; L^1(\mathbb{R}))$.

**Proof.** The strong convergence of $r_{\Delta x}$ follows from the compensated compactness theorem, Theorem 2.1 and the compactness of $\{\eta_i(r_{\Delta x})t + q_i(r_{\Delta x})\}_{\Delta x>0}$ for $i = 1,2$.

**Lemma 3.6.** If $r_j(0) > 0$ and there is a positive constant $C$ such that $1/C \leq (v_j(0)/u_j(0)) \leq C$, and

$$\left| \frac{v_0}{u_0} \right| < \infty,$$

then there is a subsequence of $\{\Delta x\}$ and a function $\varphi \in C([0,T]; L^1_{loc}(\mathbb{R}))$ such that $\varphi_{\Delta x}(:, t) \rightarrow \varphi(:, t)$ in $L^1_{loc}(\mathbb{R})$ as $\Delta x \rightarrow 0$.

**Proof.** For $j$ such that $r_j > 0$, by (14)

$$\frac{d}{dt} \varphi_j + \frac{u_j u_j - 1 - \phi_j - 1}{r_j^2 \cos^2(\varphi_j - 1/2)} D_\varphi_j = 0,$$

where $\varphi_{j-1/2}$ is some intermediate value. Set

$$\mu_j = \frac{u_j u_j - 1 - \phi_j - 1}{r_j^2 \cos^2(\varphi_j - 1/2)} \text{ and } \theta_j = D_\varphi_j.$$

Note that $\mu_j \geq 0$, and that $\mu_j$ is bounded since $\varphi_j < \pi/2$. Then $\theta_j$ satisfies

$$\frac{d}{dt} \theta_j + \mu_j - D_\theta_j + \theta_j D_m \mu_j = 0. \quad (17)$$

Let $\eta_\alpha(\theta)$ be a smooth approximation to $|\theta|$ such that

$$\eta_\alpha''(\theta) \geq 0 \text{ and } \lim_{\alpha \rightarrow 0} \eta_\alpha(\theta) = \lim_{\alpha \rightarrow 0} (\theta \eta_\alpha(\theta)) = |\theta|.$$

We multiply (17) by $\eta_\alpha'(\theta_j)$ to get an equation satisfied by $\eta_\alpha(\theta_j)$. Observe that

$$\mu_j - 1 \eta_\alpha'(\theta_j) D_\theta_j + \theta_j \eta_\alpha'(\theta_j) D_\mu_j$$

$$= \mu_j - 1 \eta_\alpha'(\theta_j) + \theta_j \eta_\alpha'(\theta_j) D_\mu_j + \frac{\Delta x}{2} \mu_j - 1 \eta_\alpha''(\theta_j - 1/2) (D_\theta_j)^2$$

$$\geq D_\mu (\mu_j \eta_\alpha(\theta_j)) + (\theta_j \eta_\alpha'(\theta_j) - \eta_\alpha(\theta_j) D_\mu_j).$$

Hence

$$\frac{d}{dt} \eta_\alpha(\theta_j) + D_\mu (\mu_j \eta_\alpha(\theta_j)) \leq (\eta_\alpha(\theta_j) - \theta_j \eta_\alpha'(\theta_j)) D_\mu_j.$$

Now let $\alpha \rightarrow 0$ to obtain

$$\frac{d}{dt} \left| \theta_j \right| + D_\mu (\mu_j \left| \theta_j \right|) \leq 0. \quad (18)$$

If we multiply this with $\Delta x$, sum over $j$ and integrate in $t$, we find that

$$|\varphi_{\Delta x}(:, t)|_{B.V.} \leq |\varphi_{\Delta x}(:, 0)|_{B.V.} \leq |\varphi(:, 0)|_{B.V.} < \infty.$$
By Helly’s theorem, for each \( t \in [0, T] \), the sequence \( \{ \varphi_{\Delta x}(\cdot, t) \}_{\Delta x > 0} \) has a subsequence which converges strongly in \( L^1_{\text{loc}}(\mathbb{R}) \). By using a diagonal argument, we can get this convergence for a dense countable set \( \{ t_k \}_{k \in \mathbb{N}} \subset [0, T] \). Since \( \varphi_{\Delta x}(\cdot, t) \) has bounded variation, it is \( L^1_{\text{loc}} \) Lipschitz continuous in \( t \), that is
\[
\| \varphi_{\Delta x}(\cdot, t) - \varphi_{\Delta x}(\cdot, s) \|_{L^1_{\text{loc}}(\mathbb{R})} \leq \sup_j \mu_j |\varphi(\cdot, 0)|_{B.V.} |t - s|.
\]
This means that \( \varphi_{\Delta x}(\cdot, t) \) converges also for \( [0, T] \ni t \notin \{ t_k \}_{k \in \mathbb{N}} \). Furthermore, it also shows that \( \varphi = \lim_{\Delta x \to 0} \varphi_{\Delta x} \) is continuous in \( t \), with values in \( L^1_{\text{loc}}(\mathbb{R}) \).

Now we have the strong convergence of \( r_{\Delta x} \) and of \( \varphi_{\Delta x} \). This means that also \( u_{\Delta x} \) and \( v_{\Delta x} \) converge strongly to some functions \( u \) and \( v \) in \( L^\infty([0,T] ; L^1_{\text{loc}}(\mathbb{R})) \) since we have
\[
u_{\Delta x} = r_{\Delta x} \cos(\varphi_{\Delta x}) \quad \text{and} \quad \psi_{\Delta x} = r_{\Delta x} \sin(\varphi_{\Delta x}).
\]

**Theorem 3.7.** Let \( \phi \) be a twice differentiable function \( \phi : [0, \infty) \to [0, \infty) \) such that \( \phi'(r) > 0 \) and \( \phi''(r) \geq 0 \), and
\[
\text{meas}\{ r \mid 2\phi'(r) + r\phi''(r) = 0 \} = 0.
\]
Let \( u_{\Delta x} \) and \( v_{\Delta x} \) be defined by (5)-(6). If \( u_0 > 0, \ v_0 > 0 \) and \( u_0^2 + v_0^2 \in L^1(\mathbb{R}) \), and \( |v_0/u_0|_{B.V.} < \infty \) then there exists functions \( u \) and \( v \) in \( L^\infty([0,T] ; L^1_{\text{loc}}(\mathbb{R})) \) such that \( u_{\Delta x} \to u \) and \( v_{\Delta x} \to v \) as \( \Delta x \to 0 \). The functions \( u \) and \( v \) are weak solutions to (1).

**Proof.** We have already established convergence. It remains to show that \( u \) and \( v \) are weak solutions. To this end, observe that\(^1\)
\[
\int_0^T \int_{\mathbb{R}} D_- (u_{\Delta x} \phi(r_{\Delta x})) \psi(x, t) \, dx \, dt = -\int_0^T \int_{\mathbb{R}} u_{\Delta x} \phi(r_{\Delta x}) D_+ \psi(x, t) \, dx \, dt.
\]
As \( \Delta x \to 0 \), \( D_+ \psi \to \psi_x \) for any \( \psi \in C^0_0(\Omega) \). This means that \( u \) is a weak solution. Similarly we can show that \( v \) is a weak solution. Hence, the functions \( u \) and \( v \) are weak solutions to (1).

**Acknowledgments.** The first author acknowledges support from the Alexander von Humboldt Foundation, through a Humboldt Research Fellowship for post-doctoral researchers.

**REFERENCES**


E-mail address: toujjwal@gmail.com
E-mail address: nilshr@math.uio.no

\(^1\)Here we “extend” the definition of \( D_- \) and \( D_+ \) to arbitrary functions in the obvious manner.
A LOCAL DISCONTINUOUS GALERKIN SCHEME FOR
COMPRESSIBLE PHASE FIELD FLOW

MIRKO KRÄNKEL
Department of Applied Mathematics, University of Freiburg
Hermann-Herderstr. 10, D-79104 Freiburg, Germany

Abstract. We present a numerical scheme for a two-phase flow model going back to work of [4] and investigated in [1]-[3]. The numerical approach is based on a Local Discontinuous Galerkin discretization in space and an implicit Runge-Kutta time marching scheme.

1. Introduction. In the numerical simulation of two-phase flow, phasefield models are an attractive approach, because the tracking of the phase boundary is incorporated directly in solving the PDEs for the flow and no additional manipulation of the computational mesh is needed. Furthermore changes of the topology of the interface like merging of bubbles and droplets don’t need special treatment. These models in general consist of the Navier-Stokes equations and an additional equations for an additional phasefield variable which distinguishes the components or phases of the fluid, see [7] for a general survey on phase field models. In contrast to the major part of the literature found on this kind of models, where the two phases are assumed to be incompressible, see e.g. [5] or [11], we are interested in a model for compressible two phase flow. Such a model was derived in [4] and [1]. In this work and in [2] and [3] the behavior of the model in the sharp interface limit was investigated. In the special case of the model considered in [2] and [3] the classical Young-Laplace which relates the jump of the pressure in normal direction the interface to the surface tension of the interface and the possibility of phase transition is recovered in the sharp interface limit. Furthermore this model allows to incorporate different physical properties of the pure phases, as the free energy of the system contain an interpolation of the free energies for each phase. Another approach to compressible two phase flow in the case when the two phases are the liquid and the vapor phase of certain substance, is the Navier-Stokes-Korteweg model. In this case the density is used to distinguish the two phases. In [10] the numerical treatment of this model was investigated. Two of the main challenges in the numerical simulation using the Navier-Stokes-Korteweg model, were the severe time step restrictions imposed by the third order terms in momentum balance and the need for a very fine computational mesh due to small size of the interfacial layer. The numerical scheme proposed in this contribution is a first to developing an alternative to the Navier-Stokes-Korteweg model. The fact that the smallness parameter for the interface can be controlled independently of the physical behavior of the pure

2000 Mathematics Subject Classification. Primary: 76T99; Secondary: 76M10.
Key words and phrases. Discontinuous Galerkin, Two-Phase flow, Phase-field, compressible Navier-Stokes equations, numerics.
The author is supported by the DFG-CNRS research unit 563.

767
phases, and therefore allows for an artificial enlargement of the interface might lead to scheme which overcomes some of the difficulties in the Navier-Stokes-Korteweg model at least in some situations.

2. The Allen-Cahn-Navier-Stokes Model. The model from [1] consists of the compressible Navier-Stokes equations and an Allen-Cahn type equation for the phase field parameter $\phi$.

$$\begin{align*}
\partial_t \rho + \nabla \cdot (\rho v) &= 0, \\
\partial_t (\rho v) + \nabla \cdot (\rho v \otimes v + \Pi) &= 0, \\
\rho(\partial_t \phi + v \cdot \nabla \phi) &= S.
\end{align*} \tag{1}$$

for $x \in \Omega \subset \mathbb{R}^d, t > 0$.

The stress tensor has the form

$$\Pi := D(\nabla v) + P(\rho, \phi, \nabla \phi).$$

To specify the structure of the pressure tensor $P$ and the reaction source $S$, we introduce the free energy density $f = f(\rho, \phi, \nabla \phi)$, which has the form

$$f(\rho, \phi, \nabla \phi) := \frac{1}{\delta} \rho W(\phi) + \delta \rho \frac{|\nabla \phi|^2}{2} + \psi(\rho, \phi).$$

where $W(\phi)$ is double-well function with respect to $\phi$ and different heights of the minima and the function $\psi$ models the physics in the pure phases.

$$\psi(\rho, \phi) := \nu(\phi) f_2(\rho) + (1 - \nu(\phi)) f_1(\rho).$$

here $f_1, f_2$ are free energy densities for the two phases and $\nu$ is interpolation function with $\nu'(1) = \nu'(0) = 0$.

The last assumption on the interpolation function is used in [3] to derive the sharp interface limit when the smallness parameter $\delta$ goes to zero.

With $\frac{\delta f}{\delta \phi}$ denoting the variational derivative $f_\phi - \nabla \cdot (f_\phi \nabla \phi)$ and a positive constant $\eta$ denoting the reaction rate the source term $S$ is given by

$$S := -\frac{1}{\delta} \eta \frac{\delta f}{\delta \phi} = \frac{\eta}{\delta} \rho W(\phi) + \psi(\rho, \phi) - \delta \nabla \cdot (\rho \nabla \phi).$$

which gives a phase field equation of Allen-Cahn type. Finally the pressure and deformation tensor $P,D$ are given

$$P := P(\rho, \phi, \nabla \phi) = (-\psi + \rho \psi_{\mid_\rho}) I + \delta (\rho \nabla \phi \otimes \nabla \phi)$$

and the deformation tensor $D$

$$D := \mu \left( \frac{1}{2} (\nabla v + (\nabla v)^T) + \lambda \nabla \cdot v I \right)$$

with $\mu, \lambda > 0$ so that $D \cdot \nabla v > 0$. In general $\mu, \lambda$ are allowed to depend on $\phi$ allowing for different physical properties of each phase.

Remark 1. The $\delta$-scaling is choosen as proposed in [2] to archive in the limit $\delta \to 0$ the classical Young-Laplace Law, relating the jump in the pressure to the surface tension of the interface.
2.1. **The system in balance form.** Using the conservation of total mass we can write
\[
\begin{align*}
\partial_t \rho + \nabla \cdot \rho v &= 0 \\
\partial_t (\rho v) + \nabla \cdot (\rho v \otimes v + \Pi) &= 0 \\
\rho (\partial_t \phi + v \cdot \nabla \phi) &= -\frac{\eta}{\delta} \frac{\delta f}{\delta \phi}
\end{align*}
\]
as second-order balance law of the form
\[
\partial_t U + \nabla \cdot F(U) + \nabla \cdot T(U, \nabla U) + \nabla \cdot D(U, \nabla U) = S(U)
\]
(2)
The system has the boundary conditions:
\[
v = 0, \nabla \phi \cdot \nu = 0 \quad \text{on } \partial \Omega
\]
and initial conditions
\[
\rho(0, x) = \rho_0(x), v(0, x) = v_0(x), \phi(0, x) = \phi_0(x).
\]
**Remark 2.** In this form the system (1) fulfills an energy inequality which is equivalent to the entropy principle in the isothermal case. See [1] and [3].

\[
U := \begin{pmatrix} \rho \\ \rho v \\ \rho \phi \end{pmatrix} \quad F(U) := \begin{pmatrix} \rho v \\ \rho v \otimes v + p \\ \rho \phi v \end{pmatrix} \quad T(U, \nabla U) := \begin{pmatrix} 0 \\ \rho \nabla \phi \otimes \nabla \phi \\ 0 \end{pmatrix} \quad S(U) := \begin{pmatrix} 0 \\ 0 \\ -\frac{\eta}{\delta} \frac{\delta f}{\delta \phi} \end{pmatrix}
\]
The diffusion tensor
\[
D(U, \nabla U) := \begin{pmatrix} 0 \\ -D_v(U, \nabla U) \\ \eta \rho \nabla \phi \end{pmatrix}
\]
with
\[
D_v = \mu_1 \nabla \cdot v I + \mu_2 \left( \frac{1}{2} (\nabla v + (\nabla v)^T) - \frac{1}{d} \nabla \cdot v I \right)
\]
**Remark 3.** The viscous flux \(D(U, \nabla)\) can be expressed as
\[
D(U, \nabla U)_i = \sum_{j=1}^{d+2} A_{i,j}(U) \partial_j U
\]
In 2 dimensions the matrices \(A_{i,j}, i = 1..d+2\) are given as
\[
A_{11} := \begin{pmatrix} 0 \\ -(2\mu - \lambda) \frac{U_{x}}{U_{x}^2} \\ -\mu \frac{U_{x}}{U_{x}} \\ -\eta \frac{U_{x}}{U_{x}} \end{pmatrix} \quad A_{12} := \begin{pmatrix} 0 \\ -\lambda \frac{U_{x}}{U_{x}} \\ -\mu \frac{U_{x}}{U_{x}} \\ -\eta \frac{U_{x}}{U_{x}} \end{pmatrix}
\]
3. The numerical scheme. To discretize the system of equations we used a Local-Discontinuous-Galerkin Method, similar to the LDG-method developed in [10, Diehl] based on the work of Cockburn and Shu [8]. We write the system (1)) as a system of first order equations. On a given computational mesh, the resulting equations are tested with piecewise polynomial functions, which are in general discontinuous across the boundaries. The values on the edges are then approximated by numerical fluxes. The first order formulation allows a convenient way to treat the nonlinearity $\rho \nabla \phi \otimes \nabla \phi$ in the momentum balance as $\nabla \phi$ is approximated by an auxiliary variable allowing to use the $L^2$-projection of the nonlinearity. See e.g. [9] where this technique is used for equation with nonlinear diffusion.

3.1. The first order system. We introduce the auxiliary variables $\sigma$ and $\theta$ to write the balance equation (2) as a system of first order equations

$$\sigma - \nabla U = 0,$$

$$\theta - T(U, \sigma) = 0,$$

$$\partial_t U + \nabla \cdot F(U) + \nabla \cdot \theta + \nabla \cdot [A(U)\sigma] = S(U).$$

3.2. Traces and discrete spaces.

**Definition 3.1.** Let $\mathcal{T} = \{ E \}$ be a partition of $\Omega$ into polygons $E$ and let $E, E' \in \mathcal{T}$ with outer normals $\nu_E, \nu_{E'}$ share an edge $e := \partial E \cap \partial E'$. Furthermore, let $\phi$ be a function, which is smooth within $E$ and $E'$, but might be discontinuous across $e$, then the inner and outer trace of $\phi$ on $e$ with respect to $E$ is given by

$$\phi^+(x) := \lim_{\epsilon \rightarrow 0^+} \phi(x + \epsilon \nu_E), \quad \phi^-(x) := \lim_{\epsilon \rightarrow 0^-} \phi(x + \epsilon \nu_{E'}).$$

Furthermore, if we choose for every edge $e$ an unique normal $\nu e$, we can define the “left” and “right” traces by

$$\phi_L(x) := \lim_{\epsilon \rightarrow 0^+} \phi(x + \epsilon \nu e), \quad \phi_R(x) := \lim_{\epsilon \rightarrow 0^-} \phi(x - \epsilon \nu e).$$

**Definition 3.2.** Let $\mathcal{T}$ be a triangulation of $\Omega$ then we define the Discontinuous Galerkin Space by

$$V_h := \{ u \in L^2(\Omega) : u|_E \in \mathbb{P}_k \text{ for all } E \in \mathcal{T} \}$$

where $\mathbb{P}_k$ is the space of polynomials of degree $\leq K$.

**Definition 3.3.** The mean value of $\phi$ on the edge $e$ is defined by:

$$\{\phi\} = \frac{1}{2}(\phi^+ + \phi^-).$$
The jump operators on $e$ are given by:

$$
[u] = u^+ \otimes \nu^+ + u^- \otimes \nu^-,
$$
$$
[\sigma] = \sigma^+ \nu^+ + \sigma^- \nu^-.
$$

Where $u, \sigma$ take values in $\mathbb{R}^{d+2}, \mathbb{R}^{(d+2)\times d}$.

By multiplying equations (3)

$$
\sigma - \nabla U = 0,
\theta - T(U, \sigma) = 0,
\partial_t U + \nabla \cdot F(U) + \nabla \cdot \theta + \nabla \cdot A(U)\sigma = S(U).
$$

with test functions and integrating by parts on each element $E \in T$

$$
\int_E \sigma : \tau \, dx = -\int_e U \cdot \nabla \cdot \tau \, dx + \int_{\partial E} U \tau \cdot \nu \, ds,
\int_E \theta : \xi \, dx = \int_E T(U, \sigma) : \xi \, dx,
\int_E \partial_t U \cdot \psi \, dx = -\int_E [F(U) + \theta + A(U)\sigma] \cdot \nabla \psi - S(U)\psi \, dx
+ \int_{\partial E} [F(U) + \theta + A(U)\sigma] \cdot \nu \psi \, ds.
$$

where $\tau \in [V_h]^{(d+2)\times d}, \xi \in [V_h]^{(d+2)\times d}, \psi \in [V_h]^{d+2}$.

3.3. Numerical Fluxes. The numerical fluxes on the edge $e$ are given by:

$$
\sigma := \llbracket A(U)\sigma \rrbracket - \llbracket A(U)\sigma \rrbracket \otimes \beta - \frac{\alpha}{h} \llbracket A(U) \rrbracket [U],
\theta := \llbracket \theta \rrbracket - \llbracket \theta \rrbracket \otimes \beta - \frac{\alpha}{h} [U],
\hat{U} := \llbracket U \rrbracket + [U] \cdot \beta.
$$

where $h$ denotes the mesh size, $\alpha > 0$ and $\beta$ a switch function. For an edge $e$ with neighboring elements $E_e^+, E_e^-$ it holds $\beta = \frac{1}{2} \nu_{E_e^-} = -\frac{1}{2} \nu_{E_e^+}$. The numerical fluxes therfore read:

$$
\llbracket A(U)\sigma \rrbracket - \llbracket A(U)\sigma \rrbracket \otimes \beta = \begin{cases} (A(U)^+)^+ \sigma^+ \\ (A(U)^-)^- \sigma^- \end{cases},
\llbracket \theta \rrbracket - \llbracket \theta \rrbracket \otimes \beta = \begin{cases} \theta^+ \\ \theta^- \end{cases},
\llbracket U \rrbracket + [U] \cdot \beta = \begin{cases} U^- \\ U^+ \end{cases}.
$$

For the advective part a Lax-Friedrich flux is chosen:

$$
\tilde{F}(U^+, U^-) := \frac{1}{2} (F(U^+) + F(U^-)) \cdot \nu - \lambda(U^+ - U^-).
$$
The numerical solution $U_h$ can then be computed by solving the discrete system
\[
\int_E \sigma_h : \tau \, dx = \int_E U_h \cdot \nabla \cdot \tau \, dx - \int_{\partial E} \tilde{U} \cdot \tau \cdot \nu \, ds,
\]
\[
\int_E \theta_h : \xi \, dx = \int_E T(U_h, \sigma_h) \cdot \xi \, dx,
\]
\[
\int_E \partial_t U_h \cdot \psi = \int_E (F(U_h) + \theta + A(U_h)\sigma_h) \cdot \nabla \psi - S(U_h)\psi \, dx,
\]
\[
- \int_{\partial E} (\tilde{F} + \tilde{\theta} + \tilde{\sigma}) \psi \cdot \nu \, ds.
\]
for all $\tau \in [V_h]^{(d+2) \times d}$, $\xi \in [V_h]^{(d+2) \times d}$, $\psi \in [V_h]^{d+2}$.

By inverting the mass matrix on each element we can solve locally for $\sigma_h$, $\theta_h$ and $\partial_t U_h$. Written in operator form system (7) reads.
\[
\sigma_h = L_1[U_h],
\]
\[
\theta_h = L_2[\sigma_h, U_h],
\]
\[
\partial_t U_h = L_3[\theta_h, \sigma_h, U_h].
\]

We now have to solve the nonlinear system of ODEs
\[
\partial_t U_h = L[U_h] := L_3[L_2[L_1[U_h], U_h], L_1[U_h], U_h].
\]

For solving this ODE system we use implicit Runge-Kutta schemes where the nonlinear system for each step system is solved by an inexact Newton-method. For the inner solver a restarted GMRES solver is used. The use of of an implicit method was more effective than an explicit scheme due the stiff nonlinearity in the phase field equation and the diffusion term in the Navier Stokes equation and the phase field equation.

4. Numerical Example. The first numerical tests were performed for a rather artificial situation from the physical point of view. In the presented simulation we assumed an ideal gas law for the pure phases $p_i = a_i \rho, i = 1, 2$. The remaining parameters were:
\[
\delta = 0.01, a_1 = 2, a_2 = 1.2, \mu = 0.01, \lambda = -\mu.
\]
The initial configuration was rotational symmetric “bubble”, where the phasefield is close to 0 and density close to 0.9 in the center of the domain and is then smoothly connected to values close to $\phi = 1$ and $\rho = 2.1$ by using the $tanh$ function. The velocity is zero at the beginning of the simulation. We used $Q1$-elements on a Cartesian mesh with 200 cells in each direction. As shwon in Figure 1, the system evolves to a state where $\phi \approx 0$ in the whole domain, which was expected, because the doublewell function $W$ has it’s smaller minimum at 0.

5. Summary and Outlook. A simple test case showed that the model and the discretization shows the expected behavior. A next step would be to develop a scheme based on a nonconservative formulation of the governing equations. As pointed out in [10] and [6], conservative schemes suffer from a non decaying total energy, which leads unphysical velocity fields in the interfacial layer, so called parasitic currents, when the system approaches an equilibrium state. At the time of
Figure 1. Evolution of the phase field and the density
writing such a scheme is under development, using the following equivalent formulation of the problem (1)

$$\theta = \frac{\delta f}{\delta \phi} = f_\phi - \nabla \cdot (\rho \sigma_\phi)$$
$$\partial_t (\rho v) + \nabla \cdot (\rho v v) + \rho \nabla \frac{\delta f}{\delta \rho} - \theta \nabla \phi = \nabla \cdot (\mathcal{S}(\sigma_v))$$
$$\partial_t (\rho \phi) + \nabla \cdot (\rho \phi v) + \frac{\mu}{2} \theta = 0.$$

Note that in this formulation the contribution of $\nabla \phi$ to the momentum balance has the same structure as in the phase field equation. So no additional error will be introduced by an incompatible discretization of term involving $\nabla \phi$ in the phase field equation and the momentum balance.

REFERENCES


E-mail address: kraenkel@mathematik.uni-freiburg.de
EXISTENCE AND UNIQUENESS FOR SCALAR CONSERVATION LAWS ON MOVING HYPERSURFACES

GERHARD DZIUK, DIETMAR KRÖNER AND THOMAS MÜLLER
Abteilung für Angewandte Mathematik, Universität Freiburg
Hermann-Herder-Str. 10, D–79104 Freiburg, Germany

Abstract. We consider conservation laws on moving hypersurfaces. In this work the velocity of the surface is prescribed. We show existence and uniqueness for a scalar conservation law on the moving surface. We also present some numerical experiments. It turns out that in addition to the "Euclidean shocks" geometrically induced shocks may appear.

1. Introduction. In this work we study scalar conservation laws on moving hypersurfaces in $\mathbb{R}^{n+1}$. The motion of the surface is prescribed. Assume that $\Gamma(t)$ is a family of smooth and compact hypersurfaces without boundary which moves smoothly with time $t \in [0, T]$. By $G_T$ we denote the space time surface $G_T = \bigcup_{t \in (0, T)} \Gamma(t) \times \{t\}$. If the scalar material quantity $u = u(x,t)$, $x \in \Gamma(t)$, $t \in [0, T]$, is propagated with the surface and simultaneously transported according to a given flux $f = f((x,t), u)$ on the surface, then its evolution with respect to prescribed initial values $u_0$ is governed by the initial value problem

$$\dot{u} + u \nabla_{\Gamma} \cdot v + \nabla_{\Gamma} \cdot f(\cdot, u) = 0 \quad \text{on } G_T, \quad u(\cdot, 0) = u_0 \quad \text{on } \Gamma(0). \quad (1)$$

Here $v$ denotes the velocity of the surface $\Gamma$, and $\nabla_{\Gamma}$ is the surface gradient. The dot stands for a material derivative, which is defined for a differentiable function $g : G_T \to \mathbb{R}$ as

$$\dot{g} = \frac{\partial g}{\partial t} + v \cdot \nabla g. \quad (2)$$

Note that the material derivative only depends on the values of $g$ on the space-time surface $G_T$. The given flux function $f$ is assumed to be a tangent vector to the surface $\Gamma$ and to have vanishing tangential divergence. All quantities appearing in (1) are well defined for $u : G_T \to \mathbb{R}$ and do not depend on the ambient space. The derivation of the PDE is based on a conservation principle on moving surfaces (see [10]).

Let us briefly summarize the published results related to this topic. Total variation estimates for time independent Riemannian manifolds can be found in [14]. The existence proof of entropy solutions on time independent Riemannian manifolds is considered in [6]. Based on this paper Lengeler and one of the authors [18] are generalizing the results which we are going to prove in this contribution to the case of time dependent Riemannian manifolds. Convergence of finite volume schemes on time independent Riemannian manifolds can be found in [1]. In [17] LeFloch,
Okutmustur and Neves prove an error estimate of the form $||u - u_h||_{L^1} \leq ch^{\frac{2}{3}}$ for the scheme in [1]. This result was generalized to the time dependent case by Gieselsmann and Wiebe [12] under the assumption that an entropy solution exists, which we are going to prove in this paper. In [2] an error estimate for hyperbolic conservation laws on an $(N + 1)$ dimensional manifold (spacetime), whose flux is a field of differential forms of degree $N$, is shown. A wave propagation algorithm for hyperbolic systems on curved manifolds with application in relativistic hydrodynamics and magnetohydrodynamics have been developed and tested in [21], [4], [5], and finite volume schemes on spherical domains, partially with adaptive grid refinement in [7], [5].

For the proofs of the results in the following part of the paper we refer to [10].

2. Notations and basic relations for moving hypersurfaces.

Assumptions 2.1. Let $\Gamma_t = \Gamma(t) \subset \mathbb{R}^{n+1}$ for $t \in [0, T]$ be a time dependent, closed, smooth hypersurface. The initial surface $\Gamma_0$ is transported by the smooth function $\Phi : \Gamma_0 \times [0, T] \to \mathbb{R}^{n+1}$ with $\Phi(\Gamma_0, t) = \Gamma_t$ and $\Phi(\cdot, 0) = Id$. We assume that $\Phi(\cdot, t) : \Gamma_0 \to \Gamma_t$ is a diffeomorphism for every $t \in [0, T]$. The velocity $v$ of the material points is defined by $v(\Phi(\cdot, t), t) = \Phi_{t}(\cdot, t)$. The tangential flow of a conservative material quantity $u$ with $u(\cdot, t) : \Gamma_t \to \mathbb{R}$ is described by a flux function $f = f((x, t), u)$ which is a family of vector fields such that $f((x, t), u)$ is a tangent vector.

For the introduction of geometric relations we fix a time $t \in [0, T]$ and consider the hypersurface $\Gamma = \Gamma_t$ with normal vector field denoted by $\nu$. Furthermore we use the summation convention in the rest of the paper, i.e. we sum over doubly appearing indices.

Definition 2.2. For a differentiable function $g : \Gamma \to \mathbb{R}$ we define its tangential gradient as $\nabla_{\Gamma} g = \nabla g - \nabla g \cdot \nu \nu$, where $\overline{g}$ is an extension of $g$ to a neighborhood of $\Gamma$. We denote the components of the gradient by $\nabla_{\Gamma} g = (D_1 g, \ldots, D_{n+1} g)$. The Laplace-Beltrami operator then is given by $\Delta_{\Gamma} g = \nabla_{\Gamma} \cdot \nabla_{\Gamma} g = D_i D_j g$.

It is well known (see [8]) that the tangential gradient only depends on the values of $g$ on $\Gamma$. With the help of tangential gradients we can describe the geometric properties of $\Gamma$. The matrix

$$H = \nabla_{\Gamma} \nu, \quad H_{ij} = (\nabla_{\Gamma} \nu)_{ij} = D_i \nu_j = D_j \nu_i \quad (i, j = 1, \ldots, n + 1)$$

has a zero eigenvalue in normal direction: $H \nu = 0$. The remaining eigenvalues $\kappa_1, \ldots, \kappa_n$ are the principal curvatures of $\Gamma$. The mean curvature $H$ of $\Gamma$ is given as the trace of $H$: $H = H_{jj} = \sum_{j=1}^{n} \kappa_j$. Note, that this definition of the mean curvature differs from the common definition by a factor $\frac{1}{n}$. Integration by parts on a subset $K \subset \Gamma$ is given by (see [13])

$$\int_{K} \nabla_{\Gamma} g = \int_{K} g H \nu + \int_{\partial K} g \mu,$$

where $\mu$ denotes the conormal to $\partial K$.

The main difficulties in the proof of the existence theorem are due to the fact that the spatial (tangential) derivatives and (material) time derivatives do not commute.
Instead we have the following properties. For a function \( g \in C^2(\Gamma) \) we have for \( i, k = 1, \ldots, n + 1 \), that
\[
\mathcal{D}_i \mathcal{D}_k g = \mathcal{D}_k \mathcal{D}_i g + \mathcal{H}_{ki} \mathcal{D}_i g \nu_i - \mathcal{H}_{li} \mathcal{D}_i g \nu_k.
\]
and for \( g \in C^2(G_T) \)
\[
(\mathcal{D}_i g) = \mathcal{D}_i g - A_{ir}(v) \mathcal{D}_r g
\]
with the matrix \( A_{ir}(v) = \mathcal{D}_i v_r - \nu_i \eta \mathcal{D}_r \nu_s \) \( (l, r = 1, \ldots, n + 1) \).

3. **Definition of entropy solutions.** As in the Euclidean case classical solutions of (1) do not exist globally in time in general. Therefore we have to introduce the notion of a weak solution.

**Definition 3.1.** A function \( u \in L^\infty(G_T) \) is called a weak solution of (1) if
\[
\int_0^T \int_\Gamma u_\tau + f(\cdot, u) \cdot \nabla \Gamma \phi + \int_{\Gamma_0} u_0 \phi(\cdot, 0) = 0
\]
for all test functions \( \phi \in C^1(G_T) \) with \( \phi(\cdot, T) = 0 \).

In general weak solutions are not unique. Therefore we select the entropy solution which will be introduced in Definition 3.3. For the motivation of the entropy condition given in (9), let us consider the following Lemma.

**Lemma 3.2.** Let \( f = (f_1, \ldots, f_{n+1}) \), \( q = (q_1, \ldots, q_{n+1}) \), \( \eta \in C^2(\mathbb{R}) \), \( \eta'' \geq 0 \). Define \( q_l(\cdot, s) := \int_{s_0}^s \eta(t) f_{i_l}(\cdot, t) dt \) for \( l = 1, \ldots, n+1 \) and let \( u_0 \in L^\infty(\Gamma_0) \). Assume that \( u_\varepsilon \) is a smooth solution of
\[
\dot{u}_\varepsilon + u_\varepsilon \nabla \Gamma \cdot v + \nabla \Gamma \cdot f(\cdot, u_\varepsilon) - \varepsilon \Delta \Gamma u_\varepsilon = 0 \quad \text{on } G_T, \quad u_\varepsilon(\cdot, 0) = u_0 \varepsilon \quad \text{on } \Gamma_0.
\]
If \( u_\varepsilon \to u \) a.e. on \( G_T \) and \( u_0 \varepsilon \to u_0 \) a.e. on \( \Gamma_0 \) for \( \varepsilon \to 0 \) and \( u \in L^1(G_T) \), then \( u \) satisfies the entropy condition
\[
-\int_{\Gamma_0} \eta(u_0) \phi(\cdot, 0) + \int_{\Gamma_0}^T \int_{\Gamma_1}^T \left( -\eta(u) \phi(\cdot, t) - q(\cdot, u) \cdot \nabla \Gamma \phi + \nabla \Gamma \cdot v(\varepsilon(u) - \eta(u)) \phi \right) \leq 0
\]
for all test functions \( \phi \in H^1(G_T) \) with \( \phi \geq 0 \) and \( \phi(\cdot, T) = 0 \).

Now we use property (9) for the definition of an entropy solution.

**Definition 3.3.** Let \( \eta, q_l \) and \( u_0 \) be as in Lemma 3.2. Then \( u \in L^\infty(G_T) \) is an entropy solution of (1) if (9) holds for all test functions \( \phi \in H^1(G_T) \) with \( \phi \geq 0 \) and \( \phi(\cdot, T) = 0 \) and for all \( \eta \) and \( q \) with the properties, mentioned above.

The following definition of Kruzkov entropy solutions is equivalent to Definition 3.3 (see [15]). A function \( u \in L^\infty(G_T) \) is called Kruzkov entropy solution of (1) if
\[
\int_0^T \int_\Gamma |u - k| \dot{\phi} - \text{sign}(u - k) k \nabla \Gamma \cdot v \phi + \text{sign}(u - k)(f(\cdot, u) - f(\cdot, k)) \nabla \Gamma \phi
\]
\[
+ \int_{\Gamma_0} |u_0 - k| \phi(\cdot, 0) \geq 0
\]
for all \( k \in \mathbb{R} \) and all test functions \( \phi \in C^1(G_T) \) with \( \phi \geq 0 \) and \( \phi(\cdot, T) = 0 \).
The regularized problem.

In order to solve the conservation law (1) we solve the initial value problem (8) and consider \( u_\varepsilon \) for \( \varepsilon \to 0 \). For technical reasons let us consider, instead of (8), the following regularized PDE

\[
\dot{u}_\varepsilon + u_\varepsilon \nabla \Gamma \cdot v + \nabla \Gamma \cdot f(\cdot, u_\varepsilon) - \varepsilon \nabla \Gamma \cdot (B \nabla \Gamma u_\varepsilon) = 0
\]

on \( G_T \) with initial data \( u_\varepsilon(\cdot, 0) = u_{0\varepsilon} \) on \( \Gamma_0 \) with \( u_{0\varepsilon} \to u_0 \) a.e. on \( \Gamma_0 \) and (7).

Here \( B = B(x, t) \) is a symmetric diffusion matrix which maps the tangent space of \( \Gamma(t) \) into the tangent space at the point \( x \in \Gamma(t) \), so that we have \( Bv = 0 \) and \( \nu^* B = 0 \). Assume also that \( B \) is positive definite on the tangent space. Similarly as in Lemma 3.2 it can be shown that \( u \) is an entropy solution if \( u_\varepsilon \to u \) for \( \varepsilon \to 0 \).

The main purpose of the next Section is to prove a priori bounds for \( u_\varepsilon \) which are independent of \( \varepsilon \).

A Priori estimates for the regularized problem, existence and uniqueness.

We assume that the initial value problem (11) has a unique smooth solution. The aim of this section is the derivation of a priori estimates for \( u_\varepsilon, \dot{u}_\varepsilon \) and \( \nabla \Gamma u_\varepsilon \) which are independent of \( \varepsilon \). First we obtain (see [10])

\[
\sup_{t \in (0, T)} \|u_\varepsilon(\cdot, t)\|_{L^\infty(\Gamma(t))} \leq c
\]

with a constant \( c \) which is independent of \( \varepsilon \). This estimate is a consequence of a maximum principle for parabolic PDEs on moving hypersurfaces.

Now we estimate the spatial gradient uniformly in \( \varepsilon \). The proof in the case of conservation laws on manifolds is much more difficult than in the Euclidean case, since in local coordinates the coefficients depend on space and time. The main idea of the proof is to apply the spatial (tangential) derivatives and (material) time derivatives to the PDE (11). Then one has to take into account that they do not commute. Instead one has to use properties (4) and (5) in order to obtain

\[
\sup_{(0, T)} \int_\Gamma |\nabla \Gamma u_\varepsilon| \leq c
\]

with a constant \( c \) which does not depend on \( \varepsilon \).

In order to get an estimate for the time derivative we assume that the matrix \( B = (B_{ik})_{i,k=1,...,n+1} \) satisfies

\[
\dot{B} = BA(v) + A(v)^* B + \lambda B, \ B(\cdot, 0) = B_0
\]

where \( \lambda > 0 \) is a constant and \( B_0 \) is a symmetric and positive definite \((n+1) \times (n+1)\) matrix. It can be shown that there exists a symmetric and positive definite matrix \( B \) which solves (14) (see [10]). In this case we get

\[
\sup_{(0, T)} \int_\Gamma |\dot{u}_\varepsilon| \leq c
\]

with a constant \( c \) which does not depend on \( \varepsilon \). Using all these a priori estimates we obtain the following existence result.

**Theorem 5.1.** Assume Assumptions 2.1, (14) and let \( u_0 \in L^1(\Gamma_0) \cap L^\infty(\Gamma_0) \). Then there exists an entropy solution of (1). Furthermore if we assume \( \nabla \Gamma(\cdot) v \in L^\infty(G_T) \) and \( \dot{f} \) and \( \nabla \Gamma(\cdot) f \) are Lipschitz continuous with respect to \( u \), then the Kruzhkov entropy solution in the sense of (10) is unique.
6. Numerical algorithm. Now we are going to derive a finite volume scheme for the initial value problem (1). Up to our knowledge the first finite volume scheme on evolving surfaces was proposed by Lenz et al. [20] for the diffusion equation. We adapt this scheme to nonlinear scalar conservation laws on evolving surfaces. In a forthcoming paper Giesselmann and one of the authors [19] are proving error estimates for this scheme.

Following Dziuk and Elliot [9] the smooth initial surface $\Gamma_0$ is approximated by a triangulated surface $\Gamma_{0,h}$ which consists of a set of simplices (triangles for $n=2$) such that all its vertices $\{x_j^0\}_{j=1}^N$ sit on $\Gamma_0$. Such a set of simplices is called a triangulation $\mathcal{T}_h^0$ of $\Gamma_0$ and $h$ indicates the maximal diameter of a triangle on the whole family of triangulations. The triangulation $\mathcal{T}_h(t)$ and its $\Gamma(t)$ approximating surface $\Gamma_h(t)$ is defined by mapping the set of vertices $\{x_j^0\}_{j=1}^N$ with $\Phi(\cdot,t)$ onto $\Gamma(t)$, i.e. $x_j(t) := \Phi(x_j^0, t)$, i.e. they lie on motion trajectories. By this construction the set of simplices can be written as $\mathcal{T}_h(t) = \{T_j(t)| j = 1, \ldots, M\}$ for $t \in [0,T]$, where $M$ is the time independent number of simplices.

For the derivation of a finite volume scheme we introduce discrete time steps $t^k = k\tau$ where $\tau$ denotes the time step size and $k$ the time step index. To an arbitrary time step $t^k$ we assign a smooth surface $\Gamma^k := \Gamma(t^k)$, its approximation $\Gamma_h^k := \Gamma_h(t^k)$ and the corresponding triangulation $\mathcal{T}_h^k := \mathcal{T}_h(t^k)$ with simplices $T_j^k := T_j(t^k)$.

From [9] we know that for sufficiently small $h$ there is a uniquely defined lifting operator from $\Gamma_h^k$ onto $\Gamma^k$ via the orthogonal projection $\mathcal{P}^k$, defined by

$$x = \mathcal{P}^k(x) + d(x,t^k) \nu(\mathcal{P}^k(x),t^k) \quad \text{for } x \in \Gamma_h^k,$$

where $\nu(\cdot,t^k)$ denotes the surface normal on $\Gamma^k$ and $d(\cdot,t^k)$ the signed distance to $\Gamma^k$. For the comparison of quantities on $\Gamma^k$ and on $\Gamma_h^k$ we define curved simplices via the projection operator, i.e. $\mathcal{T}_j^k := \mathcal{P}^k T_j^k$. Such a projection $\mathcal{T}_j^k$ propagates during the $(k+1)$th time interval $[t^k,t^{k+1}]$. We set

$$\mathcal{T}_j^k(t) := \Phi((\Phi(\cdot,t^k))^{-1}(\mathcal{T}_j^k),t)$$

and denote by $V_j^k$ the $n$-dimensional measure of $T_j^k$. We define a finite volume scheme by

$$u_{j+1}^k V_j^{k+1} = V_j^k u_j^k - \tau \sum_{e \in \partial T_j^k} g_{e,T_j^k}(u_j^k, u_{l(j,e)}) \quad \text{for } u_{j}^{0} := \frac{1}{V_{j}^{0}} \int_{\mathcal{T}_j^0} u_0,$$

(16)

where $u_j^k \in \mathbb{R}$ represents the value of $u$ on $\mathcal{T}_j^k$, $l(j,e)$ is the index of the simplex which shares the edge $e$ with $T_j^k$ and the $g_{e,T_j^k}$ are some numerical flux functions.
Example 6.1 (Engquist-Osher Numerical Flux). Using the notations from above we define for a simplex \( T^k_j \in T^k \) and an edge \( e \in \partial T^k_j \)

\[
\begin{align*}
    c_{e,T^k_j}^k(u) &= \int_{e} f(\cdot, t^k, u) \cdot \mu_{\partial T^k_j}, \\
    c_{e,T^k_j}^{k,+}(u) &= c_{e,T^k_j}^k(0) + \int_{0}^{u} \max\{c_{e,T^k_j}^k(s), 0\} ds, \\
    c_{e,T^k_j}^{k,-}(u) &= \int_{0}^{u} \min\{c_{e,T^k_j}^k(s), 0\} ds.
\end{align*}
\]

Then an Engquist-Osher-flux is defined by

\[
g_{e,T^k_j}^{k,EO}(u, v) := c_{e,T^k_j}^{k,+}(u) + c_{e,T^k_j}^{k,-}(v). \tag{20}
\]

7. Numerical experiments. The finite volume scheme (16) is validated by numerical experiments. To this end we considered a linear transport problem on a sphere whose radius decreases exponentially in time. For this problem we know the exact solution and, thus, can compute some experimental orders of convergence (EOC), which is close to 0.6. For more details we refer to [10].

The results of three other experiments are illustrated in Figures 1, 2 and 3, respectively. Both have the function

\[
u_0(x_1, x_2, x_3) = \cos^2(\pi (x_1 + 2))\lambda_{\{x_1 < -3/2\}}(x_1)
\]
as initial values. Here \( \lambda \) denotes the characteristic function. For these experiments the flux function \( f \) is constructed by taking a constant vector field which is pointing in direction of the \( x_1 \)-axis and projecting it on the hypersurface \( \Gamma_\ell \). This flux function is not divergence-free. Figure 1 shows the numerical solution of a Burgers equation on an evolving ellipsoid. You can see a shock that moves from left to right. In Figure 2 the same equation is considered, but due to fast change of geometry in the middle of the ellipsoid, the mass is compressed so fast that a second shock riding on the first one is induced. Thus, this second shock is induced by the change of geometry. For the third experiment (see Figure 3 ) the same parameters as in the second one are chosen, only the flux function is different, which is chosen to be divergence-free (see [10]).

REFERENCES

Figure 1. Burgers like shock on an evolving ellipsoid for several time steps. Here, $T$ denotes the end time.

Figure 2. We see a Burgers like shock on an evolving ellipsoid. Caused by the deformation of the ellipsoid a second shock is produced and overtakes the first one. Here, $T$ denotes the end time.
Figure 3. As in Figure 2 a second shock is geometrically induced and overtakes the first one. Here, the flux function $f$ is divergence-free. $T$ denotes the end time.


E-mail address: gerd.dziuk@mathematik.uni-freiburg.de
E-mail address: dietmar.kroener@mathematik.uni-freiburg.de
E-mail address: mueller.thomas@mathematik.uni-freiburg.de
Abstract. We discuss universality properties of blowup of classical (smooth) solutions of conservation laws in one space dimension. It is shown that the renormalized wave profile tends to a universal function, which is independent both of initial conditions and of the form of a conservation law. This property is explained in terms of the renormalization group theory. A solitary wave appears in logarithmic coordinates of the Fourier space as a counterpart of this universality. As a numerical example, blowup in ideal polytropic gas is considered.

1. Introduction. It is well-known that, in inviscid conservation laws, smooth initial conditions typically give rise to a blowup (singularity) in finite time followed by formation of a shock wave in a weak solution. Development of such finite-time singularities is a classical subject described in every book on nonlinear waves, e.g., [11, 7, 1, 3]. Scalar conservation laws in one space dimension capture a lot of what happens in general systems. In this respect, the inviscid Burgers (also called Hopf) equation
\[ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0 \]
plays a special role. Solution of this equation is easily constructed by the method of characteristics. Each characteristic carries a constant value of \( u \), and the blowup occurs when characteristics cross.

In this paper we show that the wave profile \( u(t,x) \) creates a universal “core” just before the shock formation in a generic 1D conservation law. This core is described, up to scaling symmetry, by a function, which is independent of initial conditions as well as of the flux function. We explain the universality using the renormalization group approach, and discuss its relation to more sophisticated universal phenomena described by the theory of renormalization group [12, 6, 5]. Finally, we show that in logarithmic coordinates of the Fourier transformed function \( u(t,k) \), the blowup is mapped to a stable solitary wave moving with constant speed. As an example, we describe universal structure of shock formation in ideal polytropic gas.

2. Universal structure of blowup. First, let us consider the inviscid Burgers equation (1) with smooth initial condition \( u(t_0, x) = u_0(x) \). Its solution \( u(t, x) \) constructed by the method of characteristics has the implicit form
\[ x = x_0 + u_0(x_0)(t - t_0), \quad u = u_0(x_0), \]

2000 Mathematics Subject Classification. Primary: 35L67; Secondary: 35L65.
Key words and phrases. Blowup, conservation law, universality, renormalization group, soliton.
The work is supported by CNPq grants 477907/2011-3 and 305519/2012-3.
where \( x_0 \) is an auxiliary variable. For spatial derivative of \( u(t, x) \), we have

\[
\frac{\partial u}{\partial x} = \frac{\partial u}{\partial x_0} \frac{\partial x_0}{\partial x} = \frac{u_0'(x_0)}{1 + u_0'(x_0)(t - t_0)}.
\]

The denominator vanishes at \( t = t_0 = 1/u_0'(0) \). This yields the well-known result that the classical solution blows up along the characteristic with the minimum negative value of \( u_0'(x_0) \).

We choose the origin of time and space so that the blowup singularity appears at \( t = x = 0 \), and consider the classical solution in the interval \( t_0 \leq t < 0 \). Also, we can take \( u = 0 \) at the singularity, which can be achieved by the transformation \( x \mapsto x - u_0(0)(t - t_0) \) and \( u \mapsto u + u_0(0) \), which leaves (1) invariant. In this case, the initial condition \( u_0(x) \) satisfies

\[
u_0(0) = 0, \quad t_0 = 1/u_0'(0) < 0, \quad u_0''(0) = 0, \quad u_0'''(0) > 0,
\]

which are the blowup conditions at \( t = x = u = 0 \). Using (4) in (2), for small \( x_0 \), we obtain

\[
x = ut - \frac{u'''(0)}{6u'(0)} x_0^3 + o(x_0^3), \quad u = u_0'(0)x_0 + o(x_0).
\]

Equivalently,

\[
x = ut - cw^3 + o(u^3), \quad c = \frac{u'''(0)}{6(u'(0))^3} > 0.
\]

In particular, at \( t = 0 \), we obtain the well-known singular dependence \( u \sim -x^{1/3} \).

Consider now the function

\[
u_\lambda(t, x) = G_\lambda u(t, x) = \lambda^{1/3} u(\lambda^{-2/3} t, \lambda^{-1} x).
\]

It is easy to see that \( u_\lambda \) is a new (scaled) solution of (1), so (7) represents one of the symmetries of (1). Multiplying both sides of (6) by \( \lambda \) and making the substitution \( t \mapsto \lambda^{-2/3} t, \ x \mapsto \lambda^{-1} x \) yields the equation for \( u_\lambda(t, x) \) as

\[
x = u_\lambda t - cw^3 + \lambda o\left(\left(\lambda^{-1/3} u_\lambda\right)^3\right).
\]

The last (correction) term contains powers \( (\lambda^{-1/3} u_\lambda)^n \) with \( n > 3 \). Hence, it vanishes in the limit of large \( \lambda \), and (8) takes the exact form

\[
x = wt - cw^3
\]

for the limiting function

\[
w(t, x) = \lim_{\lambda \to \infty} u_\lambda(t, x).
\]

Equation (9) was found earlier in [10] and its linear stability analysis for the inviscid Burgers equation is carried out in [5]. This equation determines the universal function \( w(x, t) \) for \( t \leq 0 \), which is independent of initial conditions. Note that different values of the coefficient \( c \) correspond to \( w(t, x) \) determined up to scaling \( \sqrt{2}w(t, x/\sqrt{c}) \), which is a symmetry of the Burgers equation (1). The function \( w(t, x) \) is a self-similar solution,

\[
G_\lambda w = w.
\]

A numerical example of convergence of \( u(t, x) \) to the universal function \( w(t, x) \) is shown in Fig. 1a.

Now we are going to generalize the above theory by showing that the same limiting function \( w(t, x) \) appears in the blowup for a generic scalar conservation law

\[
\frac{\partial U}{\partial t} + \frac{\partial f(U)}{\partial x} = 0
\]
(a) Convergence of scaled profiles $u_\lambda(t_0, x)$ (thin black curves) to the universal function $w(t_0, x)$ (bold red curve) for the inviscid Burgers equation (1). The initial condition is $u_0(x) = -\frac{\sqrt{\pi}}{2} \text{erf} x$ at $t_0 = -1$, and $\lambda = 1, 10, 10^2, 10^3$. (b) Similar convergence results (15) for blowup of a simple wave in ideal polytropic gas. The bold red curve represents the scaled universal function $w(t_0, x)/f''(0)$, and thin black curves are the scaled profiles $U_\lambda(t_0, x) = G_\lambda U(t_0, x)$ corresponding to the density variation $U$ determined by (16), (18), (21).

with an analytic flux function $f(U)$ and initial condition $U_0(x)$ at $t = t_0$. This fact can be shown as follows. Solution of equation (12) has the form

$$x = x_0 + f'(U_0(x_0))(t - t_0), \quad U = U_0(x_0),$$

(13)

which reduces to solution (2) of the Burgers equation by the substitution $u = f'(U)$. We assume that the coordinates for time, space and state are chosen such that (4) holds for $u_0(x) = f'(U_0(x))$, and $u = f'(U)$ is locally invertible as $U = g(u)$ with $g(0) = f'(0) = 0$. This implies the blowup at $t = x = 0$ with $U = u = 0$. We will use the relation

$$G_\lambda u_\lambda^n(t, x) = \lambda^{1/3} u_\lambda^n(\lambda^{-2/3} t, \lambda^{-1} x) = \lambda^{(1-n)/3} u_\lambda^n(t, x).$$

(14)

Expanding $g(u)$ in Taylor series and using (10), (14), we see that all terms with $n > 1$ vanish for large $\lambda$. Since $g(0) = 0$ and $g'(0) = 1/f''(0)$, the remaining terms yield

$$\lim_{\lambda \to \infty} G_\lambda U(t, x) = \lim_{\lambda \to \infty} G_\lambda g(u(t, x)) = w(t, x)/f''(0).$$

(15)

Conditions (10) and (15) with the function $w(t, x)$ given by (9) demonstrate strong universal character of the blowup of a classical solution for a scalar 1D conservation law. We see that, when the singular point is approached, a part of the wave profile $u(t, x)$ becomes universal, i.e., independent both of the initial condition and of the flux function up to the scaling transformation.

As an example, let us consider formation of a shock wave in a simple wave solution for one-dimensional flow of ideal polytropic gas. The density $\rho(t, x)$ in this wave is described implicitly by

$$\frac{x - x_0}{t - t_0} = \frac{\gamma + 1}{\gamma - 1} \sqrt{A\gamma\rho_0^{(\gamma-1)/2}(x_0)}, \quad \rho = \rho_0(x_0),$$

(16)
where $\rho_0(x)$ is the initial condition at $t = t_0$, see, e.g., [2]. We will use the values \( \gamma = 5/3, A = 3/5, \) and $\rho_0(x) = 2 - \arctan x$. Then expressions (16) take the form
\[
\begin{align*}
x &= x_0 + 4(2 - \arctan x_0)^{1/3}(t - t_0), \\
\rho &= 2 - \arctan x_0,
\end{align*}
\] which can be written in the form (13) for the density variation in moving frame
\[
U(t, x) = \rho(t, x + x_1 + v_1(t - t_0)) - \rho_1
\] with
\[
\begin{align*}
f'(U) &= 4(U + \rho_1)^{1/3} - v_1, \\
U_0(x) &= 2 - \arctan(x + x_1) - \rho_1.
\end{align*}
\] The quantities $t_0, x_1, \rho_1$ and $v_1$ are chosen such that $f'(0) = 0$ and the function
\[
u_0(x) = f'(U_0(x)) = 4(2 - \arctan(x + x_1))^{1/3} - v_1
\] satisfies conditions (4) corresponding to blowup at $t = x = 0$ with $U = 0$. Values of these quantities (with first three decimal digits) are
\[
t_0 = -1.155, \ x_1 = 0.183, \ \rho_1 = 1.818, \ v_1 = 4.882.
\] Convergence (15) of the scaled wave profile $U_\lambda(t, x) = G_\lambda U(t, x)$ to the universal function is demonstrated in Fig. 1b.

3. **Renormalization group approach.** The universality just described can also be explained using the renormalization group approach. For this purpose, we consider $G_\lambda$ in (7) as an operator acting in the space of solutions of the Burgers equation (1) with initial conditions satisfying (4). Then the system dynamics can be seen as an action of $G_\lambda$ combined with the scaling of time, space and state. This operator defines a differentiable group with the property $G_\lambda G_\lambda_2 = G_{\lambda_1 + \lambda_2}$. The universal function $w(t, x)$ represents a stationary point (11) of the renormalization group operator. Our analysis showed that this stationary point (more precisely, a set of stationary points $w(t, x)$ determined by (9) up to a scaling constant $c$) is asymptotically stable in the sense of Lyapunov (for $\lambda$ considered as “time”). From this property, the universality result (10) follows. Note that the linear stability analysis was carried on explicitly in [5].

In the case of a general conservation law (12), the function $U_\lambda(t, x) = G_\lambda U(t, x)$ is a solution for a conservation law with a different flux function $f_\lambda(U) = \lambda^{2/3}f(\lambda^{-1/3}U)$. Thus, $G_\lambda$ is a renormalization group operator acting in the functional space $(U, f) \mapsto (U_\lambda, f_\lambda)$ of solutions and fluxes. The universality (15) of the blowup is explained by the fact that $G_\lambda$ has the asymptotically stable stationary point $(U, f) = (w(t, x), U^2/2)$ corresponding to the universal solution (9) of the Burgers equation (1).

The role of the renormalization group operator $G_\lambda$ is similar to those in other, much more sophisticated theories. For the inner scale, $x \sim ut \sim t^{5/2}$ in (6), the wave dynamics is governed by the universal function, which is a stationary point of $G_\lambda$. This is analogous, e.g., to the stationary point of the renormalization group operator, which determines critical phenomena in second-order phase transitions [12]. At larger spatial scales, there is no universality and the solution depends on the initial condition $u_0(x)$ as well as on the flux function $f(u)$. This is analogous, in turn, to the phenomenological Landau theory of second-order phase transitions valid at larger (though still small) deviations of temperature from a critical value [8].
4. **Blowup as a solitary wave in Fourier space.** Finite time blowup implies rapid increase of solution in a short wavelength range. We will see now that the universality of the blowup in $x$-space induces a solitary wave moving with constant speed to large wave numbers in logarithmic coordinates. Consider the Fourier transform of the solution, $u(t, k) = \mathcal{F}_x[u(t, x)]$. Since $u(t, x)$ is real, we have $u(t, -k) = u^*(t, k)$, and we assume $k \geq 0$ in the analysis that follows. For Fourier transformed functions, the relation (10) yields

$$w(t, k) = \lim_{\lambda \to \infty} u_{\lambda}(t, k),$$

with the $k$-space renormalization group operator

$$u_{\lambda}(t, k) = G_{\lambda} u(t, k) \equiv \lambda^{4/3} u(\lambda^{-2/3} t, \lambda k).$$

Expression (23) can be checked using Fourier transform of (7). Note that the function $w(t, x)$ grows as $w \sim x^{1/3}$ for large $x$, see (9), and its Fourier transform (regularized by adding a small imaginary part to $x$) behaves as $w \sim k^{-4/3}$ for small $k$.

Let us consider the wave profile $u(t, k)$ transformed to logarithmic coordinates

$$\tau = -\log(t/t_0), \quad \xi = \log k,$$

(24)

which we denote by the same letter $u(\tau, \xi)$. Here the blowup at $t = 0$ corresponds to $\tau \to \infty$. Using (23), we find

$$u_{\lambda}(\tau, \xi) = e^{4a/3} u \left( \tau + \frac{2}{3} a, \xi + a \right), \quad a = \log \lambda.$$

(25)

Because of (22), we have

$$e^{4a/3} u \left( \frac{2}{3} a, \xi + a \right) \to w(0, \xi) \quad \text{as} \quad a \to \infty,$$

(26)

where we put $\tau = 0$, and $w(\tau, \xi)$ denotes the universal function $w(t, k)$ written in coordinates (24). Now we substitute $a$ by $3\tau/2$ and write (26) in the form

$$u \left( \tau, \xi + \frac{3}{2} \tau \right) \to e^{-2\tau} w(0, \xi) \quad \text{as} \quad \tau \to \infty.$$

(27)

Therefore, $u(\tau, \xi)$ forms a wave in the $\xi$-space for large $\tau$ with the profile described by the universal function $w(0, \xi)$. This wave moves with the constant speed $3/2$ and decays as $e^{-2\tau}$.

Fourier transforms of derivatives $u^{(n)}(t, x) = \partial^n u / \partial x^n$ have the form $u^{(n)}(t, k) = (ik)^n u(t, k)$. Using logarithmic coordinates $(\tau, \xi)$, derivations similar to (24)–(27) yield

$$u^{(n)} \left( \tau, \xi + \frac{3}{2} \tau \right) \to \tau^n e^{(3n/2-2)\tau} e^{n\xi} w(0, \xi) \quad \text{as} \quad \tau \to \infty.$$

(28)

Recall that $w \sim k^{-4/3} = e^{-4\xi/3}$ for $\xi \to -\infty$ (small $k$). Thus, $e^{n\xi} w(0, \xi) \to 0$ as $\xi \to -\infty$ when $n > 4/3$. In this case the wave derivative $u^{(n)}$ forms a solitary wave in the space $(\tau, \xi)$, which moves with speed $3/2$, grows exponentially as $e^{(3n/2-2)\tau}$, and has the universal shape determined by $e^{n\xi} w(0, \xi)$. Formation of such a wave is shown in Fig. 2 for the second derivative, $n = 2$. Note that this solitary wave can be related to renormalized blowup solutions in shell models of turbulence, see [4, 9].
Figure 2. Profiles of the scaled second derivative $e^{-\tau} \left| \partial^2 u / \partial x^2 \right|$ in logarithmic Fourier coordinates $(\tau, \xi)$ defined by (24) for the solution shown in Fig. 1(a). Initial condition corresponds to $\tau = 0$, and the blowup corresponds to $\tau \to \infty$. The profile forms a solitary wave traveling with constant speed $3/2$, which reflects the universality of the blowup in $x$- and $k$-spaces. The bold red curve shows the exact limiting shape of the wave determined by the universal function $w(t, x)$.

5. Conclusion. We showed that solutions of conservation laws in one space dimension possess universal structure when approaching the finite-time singularity (blowup). The limiting wave profile is described by a scaled universal function in $x$-space, or by a solitary wave moving with constant speed in logarithmic coordinates of (wave number) $k$-space. The universal function is independent both of initial condition and of flux function. The universal properties are confirmed numerically in Figs. 1 and 2. It is interesting to extend the obtained results to systems of conservation laws and to higher dimensional spaces. In particular, the renormalization technique may be used as an effective tool for the analysis of finite-time singularities for inviscid conservation laws in 2D and 3D.

Acknowledgments. The author is grateful to C.M. Dafermos, E.A. Kuznetsov and D. Marchesin for useful discussions related to this paper.

REFERENCES


E-mail address: alexei@impa.br
CLASSIFICATION OF THE UMBILIC POINT IN IMMISCIBLE THREE-PHASE FLOW IN POROUS MEDIA

VÍTOR MATOS
Faculdade de Economia, Centro de Matemática, Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto, Portugal

PABLO CASTAÑEDA AND DAN MARCHESIN
Instituto Nacional de Matemática Pura e Aplicada
Estrada Dona Castorina 110
Rio de Janeiro, 22460-320 RJ, Brazil

Abstract. We consider the flow in a porous medium of three fluids that do not mix nor interchange mass. Under simplifying assumptions this is the case for oil, water and gas in a petroleum reservoir. For a simple geometry, the horizontal displacement of a pre-existent uniform mixture by another injected mixture gives rise to a Riemann problem for a system of two conservation laws. Such a system depends on laboratory-measured relative permeability functions for each of the three fluids. For Corey models each permeability depends solely on the saturation of the respective fluid, giving rise to systems containing an umbilic point in the interior of the saturation triangle. It has been conjectured that the structure of the Riemann solution in the saturation triangle is strongly influenced by the nature of the umbilic point, which is determined by the quadratic expansion of the flux function nearby. In 1987 it was proved that, for very general Corey permeabilities, umbilic points have types I or II in Schaeffer & Shearer’s classification.

In the current work we find precisely the boundaries where the transition occurs in the saturation triangle, which was not done in 1987. The novel tool is a constructive method for determining type I.

1. Introduction. In this work, we classify a special hyperbolic singularity, called umbilic point, which arises in state space for a class of conservation laws that describe oil recovery; the class is derived in Section 4. The umbilic point is classified according to Schaeffer and Shearer scheme, [8], which applies to systems of two conservation laws with fluxes that are well approximated, in some sense, by quadratic fluxes. In Appendix of [8] Schaeffer, Shearer, Marchesin and Paes-Leme, showed that the umbilic point of strongly convex permeability models has type I or type II. However, they did not determine parameters where the change from type I to II occurs. In Section 3, we detect the transition by means of a new simple way to identify type I singularities based on the quadratic expansion of the flux functions in

2000 Mathematics Subject Classification. Primary: 35L65; Secondary: 35Q86.

Key words and phrases. Conservation laws, Riemann problem, hyperbolic singularity, flow in porous media, Petroleum Engineering.

The first author is supported by grant PCI 302791/2011-6 and by grant FCT SFRH/BSAB/1164/2011, the second author is supported by CAPES/FAPERJ Postdoc Grant E-26/102.474/2010 and the third author thanks FAPERJ-E-26/102.965/2011, FAPERJ-E-26/111.416/2010, CNPq-472923/2010-2 and CNPq-301564/2009-4.
the conservation laws about the umbilic point. The classification is given in Section 5 and depends on the sign of a quartic function of two variables. Surprisingly, the quartic function turns out to be a product of four affine functions of two variables. Therefore, the change of sign occurs on straight lines and the fluxes are type I inside a small triangle in state space, and type II outside.

2. Background. Consider a $2 \times 2$ system of conservation laws in one space variable

$$\frac{\partial}{\partial t} U + \frac{\partial}{\partial x} F(U) = 0,$$

where $-\infty < x < \infty$, $t \geq 0$, $U = U(x, t) \in \mathcal{D}$, $F : \mathcal{D} \to \mathbb{R}^2$ is a $C^2$ function and $\mathcal{D}$ is an open set of $\mathbb{R}^2$. Equation (1) together with the initial conditions

$$U(x, 0) = \begin{cases} U_L & \text{if } x < 0 \\ U_R & \text{if } 0 < x \end{cases}$$

are the so called Riemann problem.

Both (1) and (2) may be rescaled by $(x, t) \to (cx, ct)$ for $c > 0$. Therefore, the solutions of (1) and (2) should depend only on $\xi = x/t$, thus, they are sequences of constant states and two kinds of solutions: rarefactions and shocks.

A rarefaction, also called rarefaction fan, is a solution of

$$(DF(U) - \xi I) U_\xi = 0,$$

for $\xi$ in an interval. If (1) is strictly hyperbolic the eigenvalues of $DF(U)$ are ordered, $\lambda_1(U) < \lambda_2(U)$. Hence, assuming strictly hyperbolicity, there are two kinds of rarefactions, each one associated to an eigenpair.

A shock is a discontinuous solution that moves with speed $s$ along which $U$ jumps form a left state $U_-$ to a right state $U_+$,

$$\lim_{x \to -st^-} U(x, t) = U_-, \quad \lim_{x \to st^+} U(x, t) = U_+,$$

and the well known Rankine-Hugoniot condition holds:

$$F(U_+) - F(U_-) - s(U_+ - U_-) = 0.$$  \hfill (3)

There are only three kinds of relevant shocks in [8], nonetheless, we define four kinds following [6, 7], which extended some results of [8].

Definition 2.1. A solution of (3) is a:

(i) slow shock iff $\lambda_1(U_+) < s < \lambda_1(U_-)$ and $s < \lambda_2(U_+);
(ii) fast shock iff $\lambda_2(U_+) < s < \lambda_2(U_-)$ and $\lambda_1(U_-) < s;
(iii) overcompressive shock iff $\lambda_2(U_+) < s < \lambda_1(U_-)$;
(iv) crossing discontinuity iff $\lambda_1(U_-) < s < \lambda_2(U_-)$ and $\lambda_1(U_+) < s < \lambda_2(U_+)$.

The oriented integral curves of each eigenvector field form rarefaction curves in state space. The rarefaction curves have the orientation of increasing eigenvalue. The solutions of (3) for a fixed $U_-$ that satisfy any of the condition of Definition 2.1 form shock curves in state space.

We give a definition of umbilic point slightly different from the one given in [8] by Schaeffer and Shearer.

Definition 2.2. Let $F$ be a $C^2$ function on an open set $\mathcal{D} \subseteq \mathbb{R}^2$, $F : \mathcal{D} \to \mathbb{R}^2$, $U \mapsto F(U)$ and $DF$ its derivative. Assume that $DF$ has two equal eigenvalues and is diagonalizable at $U^*$. If there exists a neighborhood $\mathcal{N}$ of $U^*$ such that $DF$ has distinct real eigenvalues for all $U \in \mathcal{N} \setminus U^*$ then $U^*$ is an umbilic point of $F$. 
Instead of saying that $U^*$ is an umbilic point of Eq. (1) with flux $F$, we will say just that $U^*$ is an umbilic point of $F$. (For a singularity where $\mathcal{N}$ intersects the hyperbolic and elliptic regions see [2].) The following simplifications are assumed.

**Simplification 2.3.** Let $F$ and $U^*$ be as in Definition 2.2, then:

(i) The flux $F$ has no constant terms;
(ii) The umbilic point is the origin, $U^* = (0,0)$;
(iii) The derivative $DF(U^*)$ is the null matrix.

The Simplification 2.3 holds without loss of generality since: (i) constant terms do not change the solution of conservation laws, thus they may be neglected; (ii) performing the translation $U \rightarrow U - U^*$ the umbilic point lies at the origin; (iii) changing the inertial frame, $(x,t) \rightarrow (x - \xi^* t, t)$, where $\xi^*$ is the eigenvalue of $DF(U^*)$, the derivative becomes $DF(U^*) - \xi^* I = 0$.

The hypothesis $\mathcal{H}$ of [8] motivates the following definition.

**Definition 2.4.** Let $F$ and $U^*$ be as in Definition 2.2 and $G$ be the second order Taylor expansion of $F$ around $U^*$. If $U^*$ is an umbilic point of $G$ then $F$ is an $\mathcal{H}$-flux.

We remark that if $U^*$ is an umbilic point of the quadratic function $G$ then it is also an umbilic point of $F$. However, the inverse implication is false.

Henceforth, $F$ will be an $\mathcal{H}$-flux, $G$ will be the second order Taylor expansion of $F$ (thus $G$ is an $\mathcal{H}$-flux too) and $U^*$ will be the umbilic point of $F$ and $G$.

**Remark 1.** Since we assume the Simplification 2.3 has been made on $F$, $G$ is an homogeneous quadratic $\mathcal{H}$-flux.

Schaeffer and Shearer in [8] classified generic $\mathcal{H}$-fluxes depending on the number and type of shock and rarefaction curves through the umbilic point. They did so in two steps. First, they established four robust topological configurations for homogeneous quadratic $\mathcal{H}$-fluxes (see Theorem 2.7 below). Then, for non degenerate homogeneous quadratic $\mathcal{H}$-flux, they proved the higher order terms do not affect the classification (see Theorem 2.10 below).

Palmeira and Marchesin, [6, 7], proved that higher order terms also do not affect the topological behavior of the shock and rarefaction curves in a neighborhood of $U^*$ (not only through $U^*$) – see Theorem 2.11 below. This result arises in the context of the Wave Manifold, see [4].

**Definition 2.5.** The fluxes $G_1$ and $G_2$ are equivalent if and only if there is a constant invertible $2 \times 2$ matrix $M$ such that

$$G_1(U) = M^{-1}G_2(MU).$$

**Lemma 2.6.** Equivalence preserves the structure of the shock and rarefaction curves. That is, the mapping $U \mapsto M^{-1}U$ maps shock and rarefaction curves to shock and rarefaction curves, respectively.

A version of the main result of Schaeffer and Shearer is presented in the following Theorems 2.7 and 2.10.

**Theorem 2.7.** Every $G$ is equivalent to the following normal flux

$$Q(u,v)^T = \begin{bmatrix} au^2 + 2buv + v^2 \\ bu^2 + 2uv \end{bmatrix}$$

(4)
with \( a \neq b^2 + 1 \). There are four different robust configurations for the shock and rarefaction curves of \( G \) through \( U^* \), labeled from I to IV. The flux \( G \) has:

(i) type I if and only if \( a < \frac{3}{4}b^2 \);
(ii) type II if and only if \( \frac{3}{4}b^2 < a < b^2 + 1 \);
(iii) type III if and only if \( b^2 + 1 < a < \Phi(b) \);
(iv) type IV if and only if \( \Phi(b) < a \).

**Definition 2.8.** The quadratic flux function \( G \) is degenerate if it is equivalent to \( Q \) with \( a = \frac{3}{4}b^2 \) or \( a = \Phi(b) \).

We remark that if \( a = b^2 + 1 \) then the \( Q \) flux has no umbilic point, rather it has a straight line of points with just one real eigenvalue. Thus, if \( a = b^2 + 1 \), \( Q \) is not a \( \mathcal{H} \)-flux and no \( G \) may be equivalent to \( Q \). (Schaeffer and Shearer define degenerate \( Q \) instead of degenerate \( G \); \( Q \) is degenerate also if \( a = b^2 + 1 \).)

The point (i) of the following Definition is natural, the point (ii) will be useful:

**Definition 2.9.** A given \( F \) is classified depending on its corresponding \( G \):

(i) if \( G \) has type I to IV then \( F \) has type I to IV, respectively;
(ii) if \( G \) is a border case between type I and type II (i.e., if \( G \) is equivalent to \( Q \) with \( a = \frac{3}{4}b^2 \) then \( F \) is of border-type I/II;

The classification of Schaeffer and Shearer depends on the shock and rarefaction curves through \( U^* \).

**Theorem 2.10.** If \( G \) is non degenerate, the shock and rarefaction curves of \( G \) through \( U^* \) are in one-to-one correspondence with shock and rarefaction curves of \( F \) through \( U^* \). Indeed every shock and rarefaction curve of \( F \) is tangent at \( U^* \) to a shock and rarefaction curve of \( G \).

Despite the fact that the classification depends only on the shock and rarefaction curves through \( U^* \), there is a one-to-one correspondence with shock and rarefaction curves of \( F \) around \( U^* \). The following Theorem summarizes the stability results from [6, 7].

**Theorem 2.11.** Shock and rarefaction curves near the umbilic point are structurally stable under \( C^3 \) perturbations of fluxes in the Whitney topology.

We remark that rarefactions curves are also structurally stable under \( C^2 \) perturbations of fluxes.

3. A constructive classification of type I fluxes. The work [8] provides the construction for a normal flux equivalent to a given \( \mathcal{H} \)-flux. Nevertheless, it is not always practical to use directly this construction in order to obtain the normal flux (4). The new Theorem 3.2 allows us to determine if a flux has type I.

For any \( G \) we have (recall Remark 1): (i) \( DG(U) \) is linear in \( u \) and \( v \); (ii) \( \det(DG(U)) \) is a quadratic form in \( u \) and \( v \). These facts motivate the following definition.

**Definition 3.1.** For a given \( G \) as in Definition 2.4 we associate the constant symmetric matrix \( N_G \) such that \( \det(DG(U)) = U^T N_G U \).

\[ \Phi(b) \text{ is given implicitly by } -32b^4 + b^2 \left( 27 + 36(a - 2) - 4(a - 2)^2 \right) + 4(a - 2)^3 = 0. \]
Theorem 3.2. Let $F$, $G$ and $N_G$ be as in Definitions 2.2, 2.4 and 3.1, then:
(i) $F$ has type I if and only if $\det (N_G) > 0$;
(ii) $F$ has border-type I/II if and only if $\det (N_G) = 0$;
(iii) $F$ has neither type I nor border-type I/II if and only if $\det (N_G) < 0$.

Proof. The derivative of the normal flux $Q$ from Equation (4) is
\[ DQ(U) = \begin{bmatrix} 2a + 2bv & 2bu + 2v \\ 2bu + 2v & 2u \end{bmatrix}, \]
thus the matrix $N_Q$ associated to $DQ$ is:
\[ N_Q = \begin{bmatrix} 4a - 4b^2 & -2b \\ -2b & -4 \end{bmatrix}. \]

Then $\det (N_Q) = -16a + 12b^2$, thus:
(i) $Q$ has type I if and only if $\det (N_Q) > 0$;
(ii) $Q$ has border-type I/II if and only if $\det (N_Q) = 0$;
(iii) $Q$ has neither type I nor border-type I/II if and only if $\det (N_Q) < 0$.

If $Q$ and $G$ are equivalent then there exists an invertible constant matrix $M$ such
that $Q(U) = M^{-1}G(MU)$ for all $U \in \mathbb{R}^2$, thus
\[ DQ(U) = M^{-1}DG(MU)M \Rightarrow \det (DQ(U)) = \det(DG(MU)) \Leftrightarrow \]
\[ \Leftrightarrow U^TN_QU = U^TM^TN_GMU \Leftrightarrow N_Q = M^TN_GM \Rightarrow \]
\[ \Rightarrow \det (N_Q) = \det(M^TN_GM) \Leftrightarrow \det(N_Q) = \det^2(M)\det(N_G). \]
Therefore, if $G$ and $Q$ are equivalent $\det(N_Q)$ and $\det(N_G)$ have the same sign. Since $F$ is classified depending on $G$, as in Definition 2.9, the proof is complete. $\square$

Remark 2. If $G$ is non degenerate, Theorems 2.10 and 2.11 ensure that shock and rarefaction curves around $U^*$ for $F$ and $G$ are topologically equivalent. If $G$ is degenerate, for instance, $F$ is border-type I/II, then nothing is known about the relation between these shock and rarefaction curves of $F$ and $G$.

4. The general immiscible three-phase flow in porous media. In this section we give a brief derivation of the model we studied, sometimes named Corey with general power permeabilities, which arises in Petroleum Engineering. We consider a one-dimensional, horizontal (i.e., with negligable gravitational effects), incompressible flow in a homogeneous porous media filled with three immiscible phases (e.g., water, oil and gas) with no mass exchange between phases. The fraction of the porous volume occupied by each phase is called saturation and is denoted by $s_i$, for $i \in \{w, o, g\}$. (The subscripts $w$, $o$ and $g$ stand for water, oil and gas, respectively.) Admitting no unflled pore space and no other phase we have $\sum_j s_j = 1$.

Definition 4.1. The domain of $(s_w, s_o)$ such that $s_w + s_o < 1$, $0 < s_w$ and $0 < s_o$ is called the saturation triangle $\Delta$.

We also define: $\phi$, the porosity of the medium; $\rho_i$, the constant density of phase $i$; $v$, total seepage velocity of all fluids; $f_i$, fractional flow of phase $i$, so $\sum_j f_j = 1$.

The mass conservation of each of the three phases is given by:
\[ \frac{\partial}{\partial t} (\phi \rho_i s_i) + \frac{\partial}{\partial x} (\rho_i v f_i) = 0, \quad \text{for} \quad i \in \{w, o, g\}. \quad (5) \]
Assuming that \( v \) is a non zero constant, determined by boundary conditions, it is possible to eliminate one equation and to nondimensionalize Eq. (5) to obtain

\[
\frac{\partial}{\partial t} s_i + \frac{\partial}{\partial x} f_i = 0, \quad \text{for} \quad i \in \{w, o\},
\]

\[
s_g = 1 - s_w - s_o \quad \text{and} \quad f_g = 1 - f_w - f_o.
\]

The fractional flow is given by \( f_i = \sum_{m=1}^{m_i} \) where \( m_i = b_i s_i^{a_i}, \) \( a_i > 1 \) and \( b_i > 0, \) is the mobility of each phase. Therefore, we have

\[
f_w(s_w, s_o) = \frac{b_w s_w^{a_w}}{b_o s_o^{a_o} + b_g (1-s_w-s_o)^{a_g}},
\]

\[
f_o(s_w, s_o) = \frac{b_o s_o^{a_o}}{b_w s_w^{a_w} + b_o s_o^{a_o} + b_g (1-s_w-s_o)^{a_g}}.
\]

We will study (1) with the flux function \( F: \Delta \rightarrow \mathbb{R}^2, \) \( (s_w, s_o) \mapsto (f_w, f_o)^T \) with \( f_w \) and \( f_o \) as in Eq. (7) – see Definition 4.1.

5. Classification of the umbilic point for general immiscible three-phase flow in porous media. In [3] and [8] it was proved that the system (6) with flux (7) has an umbilic point at \( (s_w, s_o) \) satisfying

\[
\frac{d m_{w} (s_w)}{ds_w} = \frac{d m_{o} (s_o)}{ds_o} = \frac{d m_{g} (s_g)}{ds_g}.
\]

We call \( \alpha, \beta \) the coordinates of the umbilic point and \( \gamma = 1 - \alpha - \beta. \)

In Appendix of [8] Schaeffer, Shearer, Marchesin and Paes-Leme showed, in a more general case, that the flux \( F \) of Eq. (7) is type I or II but they did not determine the location of the border between I and II. The main result of our paper (Theorem 5.2) is the classification of \( F \) depending on the position of the umbilic point \( (\alpha, \beta) \) on the saturation triangle. To do that we will use Theorem 3.2.

Expanding \( f_w \) and \( f_o \) of Eq. (7) around \( (\alpha, \beta) \) up to second order, \( s_w = u + \alpha, \) \( s_o = v + \beta, \) \( (s_g = w + \gamma), \) and performing Simplification 2.3 we obtain a homogeneous quadratic \( \mathcal{H} \)-flux that allows us to classify the flux \( F: \)

\[
g_w(u, v) = \frac{q_w (\bar{m}_o + \bar{m}_g) - q_g \bar{m}_w u^2}{\bar{m}^2} - \frac{q_g \bar{m}_w v^2}{\bar{m}^2} + \frac{(q_o + q_g) \bar{m}_w}{2} - \frac{(q_o + q_g) \bar{m}_w}{2},
\]

\[
g_o(u, v) = -\frac{(q_w + q_g) \bar{m}_o u^2}{2} - \frac{q_g \bar{m}_w + q_o \bar{m}_o}{2} + \frac{q_o (\bar{m}_w - \bar{m}_o) - q_g \bar{m}_w v^2}{2},
\]

where:

\[
\bar{m}_w = b_w a_w^{a_w}, \quad \bar{m}_o = b_o a_o^{a_o}, \quad \bar{m}_g = b_g a_g^{a_g},
\]

\[
q_w = A_w b_w a_w^{a_w-2}, \quad q_o = A_o b_o a_o^{a_o-2}, \quad q_g = A_g b_g a_g^{a_g-2},
\]

\[
m = \bar{m}_w + \bar{m}_o + \bar{m}_g, \quad A_i = a_i(a_i - 1) \quad \text{for} \quad i \in \{w, o, g\}.
\]

For simplicity we write \( A_w = A, \) \( A_o = B \) and \( A_g = C. \) We define \( G(u, v) = (g_w(u, v), g_o(u, v))^T \) where \( g_w(u, v) \) and \( g_o(u, v) \) are given by Eq. (8). The following triangle is used in the classification of \( F. \)

**Definition 5.1.** Let us define the triangle \( \mathcal{T} \) in the plane \( (\alpha, \beta) \) with vertices:

\[
\left( \frac{AB}{AB+BC}, 0 \right), \left( 0, \frac{AB}{CA+AB} \right) \quad \text{and} \quad \left( \frac{CA}{BC+CA}, \frac{BC}{BC+CA} \right);
\]

along the sides of the triangle \( \Delta \) in Fig. 1. The corresponding \( \gamma \) coordinates are

\[
\frac{BC}{AB+BC}, \frac{CA}{CA+AB} \quad \text{and} \quad 0.
\]
CLASSIFICATION OF THE UMBILIC POINT

Figure 1. The triangle $T$ inside of $\Delta$ and the vertices of $T$ as centers of mass.

Figure 2. The plane $(\alpha,\beta)$ and the classification on the triangle $\Delta$ in type I and II. (Example for $A^2 = 6$, $B^2 = 2$ and $C^2 = 20$.)

Assuming that $A$, $B$, $C$ are all different, each straight line crosses each of the three axes $\alpha = 0$, $\beta = 0$ and $\gamma = 0$ once. The intersection points with the axes are the same for each pair of straight lines (see Fig. 2). There exists a curious way to compute $T$. Let us put masses on the vertices of $\Delta$: (i) a mass $BC$ at $\alpha = 1$; (ii) a mass $CA$ at $\beta = 1$; (iii) a mass of $AB$ at $\gamma = 1$; (see Fig. 1). Then, in order to calculate the vertex of $T$ that lies at one edge of $\Delta$, we must swap the masses on the vertices of that edge and then calculate the center of mass.

Theorem 5.2. The flux $F$ of Eq. (7) has:

(i) type I if and only if $(\alpha,\beta)$ lies inside $T$;
(ii) type II if and only if $(\alpha,\beta)$ lies outside $T$;
(iii) border-type I/II if and only if $(\alpha,\beta)$ lies on the edges of $T$. 

Proof. We classify $F$ using Theorem 3.2. The second order expansion of $F$ is $G$ in Eq. (8). Straightforward calculations show the matrix $N_G$ is
\[
N_G = \frac{1}{2m^3} \begin{bmatrix}
-m_oq_gq_w & q_wm_gq_o - g_qm_wq_o - m_oq_gq_w \\
q_wm_gq_o - q_gm_wq_o - m_oq_gq_w & -g_qm_wq_o
\end{bmatrix},
\]
and its determinant $\det(N_G)$ is the product of the negative constant $-4A^2B^3C^2\alpha^2\beta^2\gamma^2\bar{m}^6$ by the product of four linear functions:
\[
+BC\alpha + AC\beta - AB\gamma; \quad +BC\alpha - AC\beta + AB\gamma;
\]
\[
-BC\alpha + AC\beta + AB\gamma; \quad +BC\alpha + AC\beta + AB\gamma.
\]
(9)
Therefore, the zero set of $\det(N_G)$ is formed by four straight lines, on which $F$ has border-type I/II. Substituting any of $\alpha$, $\beta$, $\gamma$ by zero in (9) we find the vertices of $T$. We may check the sign of $\det(N_G)$ to prove that $F$ has type I only inside of $T$. Since in [8] it was proved that $F$ has type I or II, it follows that $F$ has type II outside of $T$.\]

In the special case of equal mobilities, $a_w = a_o = a_g$, $A = B = C$, the triangles $\Delta$ and $T$ have parallel edges, the vertices of $T$ are $(0, \frac{1}{2})$, $(\frac{1}{2}, 0)$ and $(\frac{1}{2}, \frac{1}{2})$. The case of exponent 2 is studied by Asakura in [1] and by Isaacson et al. in [5].

6. Conclusion. In this work we classify the singularities that arise from a model of Petroleum Engineering. The model depends on parameters, $a_i$, $b_i$ for $i \in \{w, o, g\}$. The location of the singularity, the umbilic point, depends on the same parameters. However, the subdivision of the saturation triangle in type I or type II umbilic points depends only on the parameters $A$, $B$ and $C$; i.e., on the mobility powers $a_w$, $a_o$ and $a_g$, in a very geometrical way.

REFERENCES


*E-mail address:* vmatos@fep.up.pt

*E-mail address:* castaneda@impa.br

*E-mail address:* marches@impa.br
WELL-BALANCED SIMULATION OF GEOPHYSICAL FLOWS VIA THE SHALLOW WATER EQUATIONS WITH BOTTOM TOPOGRAPHY: CONSISTENCY AND NUMERICAL COMPUTATION

THOMAS MÜLLER AND AXEL PFEIFFER
Abteilung für Angewandte Mathematik, Universität Freiburg
Hermann-Herder-Str. 10, D-79104 Freiburg, Germany

Abstract. In this contribution we review a well-balancing modification for the shallow water equations and give a new result on its order of consistency, where dry regions are included. Furthermore, we combine the well-balancing modification with an existing wetting and drying approach. We show numerical results including the application to equations posed on surfaces including a realistic bottom topography. For this, we propose a finite volume scheme that is able to deal with a variety of surfaces.

1. Introduction. The shallow water equations are a simplified mathematical model for incompressible free-surface flows that can be used for the simulation of geophysical problems (tsunamis, avalanches etc.). The system of conservation/balance laws is expressed in terms of the water height $h$ and the depth-averaged horizontal momentum $hv$ and reads

$$\partial_t u + \text{div} F(u) = S(u, b) \quad \text{in } \Omega \times [0, T].$$

where

$$u := \left( \frac{h}{hv}, \right), \quad F(u) := \left( hv \otimes \frac{hv}{h} + g h^2 \text{Id} \right), \quad S(u, b) := \left( 0 - gh \nabla b \right)$$

and $s = s(x, t)$ denotes the position of the free surface, $b = b(x)$ the bottom profile, $h = h(x, t) = s(x, t) - b(x)$ the water height, $v = v(x, t)$ the velocity and $g$ the gravitational constant. Additionally, we need to impose initial values

$$u(\cdot, 0) = u_0(\cdot) \quad \text{in } \Omega.\quad (2)$$

Usually, the computational domain $\Omega$ is a subset of the horizontal plane ($\mathbb{R}^2$ or $\mathbb{R}$), but we also consider the shallow water equations on the sphere $S^2$ as a model for global atmospheric flows (e.g. weather forecast, propagation of tsunamis). Since classical schemes do not preserve the stable lake at rest state ($s =$-const and $v \equiv 0$) for a non-flat bottom topography, the well-balancing strategy described in [2] is applied to the finite volume scheme. This approach is based on a modification of the arguments of the numerical flux function (hydrostatic reconstruction). One
main focus of this talk is to show that under certain assumptions the order of consistency of the first order scheme (in 1D) is not affected by this modification, cf. Section 3.

In our simulations with real bottom topography, previously wet regions may drain with time or, conversely, dry regions may be filled with water. Since this can lead to instabilities/unphysical water heights, we employ, additionally to the well-balancing, a wetting and drying strategy proposed by [4], cf. Section 4.2.

For the shallow water equations on the rotating sphere, one has to take into account that the momentum should be tangential to the sphere. We propose a finite volume scheme that is based on a discrete parallel transport of the momentum before being put into the numerical flux function. For our simulations we use the real bottom topography of the earth [1]. The generic concept of the implementation allows us to replace both the governing system with another hyperbolic system (e.g. Euler equations of gas dynamics) and the computational domain with other closed surfaces (e.g. torus), cf. Section 5.

2. Finite Volume Scheme. In this section, as a basis for the subsequent well-balancing and wetting and drying modifications, we introduce the standard finite volume (fv) scheme. For the sake of a simple presentation we restrict ourselves to the one-dimensional case. The eigenvalues of the one-dimensional shallow water equations are given by \( \lambda_1(u) = \frac{hv}{h} - \sqrt{gh} \) and \( \lambda_2(u) = \frac{hv}{h} + \sqrt{gh} \).

Notation. We fix the following notation for the discretization of the computational domain \( \Omega = [a, b] \). We divide \( \Omega \) into grid cells \( C_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \) for \( i \in I \) with width \( \Delta x \). The midpoint of a cell \( C_i \) is \( x_i = a + (i + \frac{1}{2})\Delta x \) while the boundary points are \( x_{i\pm\frac{1}{2}} = x_i \pm \frac{\Delta x}{2} \). The time step size is denoted by \( \Delta t \) and time steps by \( t^n = n\Delta t \).

**Definition 2.1** (Finite Volume Scheme without bottom topography). Let \( \mathcal{U}_0 := \{(h, hv) \in \mathbb{R}^2 | h > 0 \text{ or } (h = 0, \frac{hv}{h} := 0) \} \) denote the space of physical values for \( u \). For a trivial bottom topography \( b \equiv 0 \) and initial datum \( u_0 \in L^\infty(\mathbb{R}, \mathcal{U}_0) \) we define the fv scheme by

\[
\begin{align*}
    u_i^{n+1} &:= Qu_i^n := u_i^n - \frac{\Delta t}{\Delta x} (g(u_i^n, u_{i+1}^n) - g(u_{i-1}^n, u_i^n)), \\
    u_0^i &:= \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_0(x) \, dx,
\end{align*}
\]

(3)

(4)

where \( g \) denotes a numerical flux function that is consistent with \( f \). In order to get a stable scheme we need to impose the following CFL condition:

\[
\Delta t \max_i \left( \max_{k=1,2} |\lambda_k(u_i)| \right) \leq \Delta x.
\]

(5)

The numerical solution \( u_{\Delta x} \) is defined by \( u_{\Delta x}(x, t) := u_i^n \) for \( x \in C_i \) and \( t^n \leq t < t^{n+1} \).

3. Well-Balancing. For a bottom topography \( b \) of class \( C^1 \) the (stationary) lake at rest function \( u(x, t) = (h(x, t), hv(x, t))^T := (s - b(x), 0)^T \) is an exact solution of the shallow water equations (1). For a numerical scheme it is desirable to preserve such stationary solutions, but in general fv schemes do not preserve lake at rest solutions exactly.
A modification of numerical flux functions that ensures the preservation of such stationary solutions is called well-balancing modification. In the following we describe a well-balancing ansatz due to Audusse et al. [2]. This approach is based on the so-called hydrostatic reconstruction of the arguments of the numerical flux functions. The main part of this section is devoted to proving that the well-balancing modification does not affect the order of consistency of the numerical scheme.

Audusse et al. obtain a well-balanced scheme $Q^*$ for the shallow water equations with bottom topography as follows. First, the discrete initial values $u^0_i$ are given by equation (4). Then, for given discrete values $u^n_i$, the reconstructed water heights, two at each interface $x_{i+\frac{1}{2}}$, are defined by

$$
\begin{align*}
 h^n_{i+\frac{1}{2}} &= \max(0, h^n_i + b_i - \max(b_i, b_{i+1})), \\
 h^n_{i+\frac{3}{2}} &= \max(0, h^n_{i+1} + b_{i+1} - \max(b_i, b_{i+1})), \\
 b_i &= \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} b(x) dx.
\end{align*}
$$

The reconstructed values for the corresponding interface are then defined by

$$
\begin{align*}
 u^n_{i+\frac{1}{2}}^- &= \begin{cases} 
 h^n_{i+\frac{1}{2}} - h_i^n & h_i^n > 0 \\
 0 & h_i^n = 0
\end{cases}, \\
 S^n_i &= \frac{1}{\Delta x} \left( \frac{h^n_{i+\frac{1}{2}}}{2} (h^n_{i+\frac{1}{2}} - 2 h^n_{i+1})^2 - \frac{h^n_{i+1}}{2} (h^n_{i+\frac{1}{2}} + h^n_{i+1})^2 \right), \\
 u^n_{i+\frac{3}{2}}^+ &= \begin{cases} 
 h^n_{i+\frac{3}{2}} + h^n_{i+1} & h^n_{i+1} > 0 \\
 0 & h^n_{i+1} = 0
\end{cases},
\end{align*}
$$

where definition of the discrete source term $S^n_i$ is motivated by the balance $\partial_x (\frac{g}{2} h^2) = -gh\partial_x b$ which is equivalent to $gh\partial_x s = 0$ and obviously holds for lake-at-rest situations.

**Definition 3.1** (Well-Balanced Finite Volume Scheme [2]). The well-balanced fv scheme operator $Q^*$ is defined by

$$
Q^* u^n_i := u^{n+1}_i := u^n_i - \Delta t \left( \frac{g}{\Delta x} (u^n_{i+\frac{1}{2}}, u^n_{i+\frac{1}{2}}) - g(u^n_{i-\frac{1}{2}}, u^n_{i-\frac{1}{2}}) \right) + \Delta t S^n_i.
$$

**Remark 1.** The operator $Q^*$ preserves lake-at-rest situations exactly and does not impose a stronger CFL condition as (5).

In the following we are going to prove consistency of order one for the well-balanced scheme $Q^*$. To this end let us assume that an exact solution $w$ of (1)-(2) is given and define the pointwise evaluation $w^n_i := w(x_i, t^n)$. Furthermore, since we will apply $Q^*$ on $w^n_i$, all discrete quantities in the rest of this section are defined on the basis of the $w^n_i$’s. From [2, 5] it follows that for a Lipschitz numerical flux function $g$, we have

$$
w^{n+1}_i - Q^* w^n_i = O(\Delta t).
$$

Let us define wet regions $M_{wet} := \{(x, t) \in \Omega \times \mathbb{R}^+ \mid h(x, t) > 0\}$ and dry regions $M_{dry} := \{(x, t) \in \Omega \times \mathbb{R}^+ \mid h(x, t) = 0\}$ for a given water height function $h$.

**Theorem 3.2.** Let $\Omega = [a, b] \subset \mathbb{R}$ be bounded and $g \in C^2([0,1], \mathbb{R})$ and assume that all second order derivatives of $g$ are bounded for occurring parameter values. Let the domain $\Omega \times \mathbb{R}^+$ be separated by a finite number of smooth curves (wet-dry-interfaces) $S_k : t \mapsto (x_k(t), t), \quad k = 1, \ldots, N_1$ into wet and dry regions. For $b \in C^2(\Omega)$ with $||b||_{L^\infty} \leq c_1$ for some constant $c_1$ we consider a Lipschitz-continuous weak solution

$$
\frac{\partial h}{\partial t} + \frac{\partial}{\partial x} \left( \frac{1}{2} (\sqrt{1 + \frac{g}{h}})^2 \frac{\partial h}{\partial x} \right) = 0
$$

for a given water height function $h$. It is consistent of order one.
\(w = (h, hv)^T \in C^{0,1}(\Omega \times \mathbb{R}^+, \mathcal{U}_0)\) of the shallow water equations (1)-(2) with the assumptions \(M_{\text{wet}} \cap M_{\text{dry}} = \{M_k(t), t \in \mathbb{R}^+, \ k = 1, \ldots, N_1\}\) and

\[w \in C^2(M_{\text{wet}}, \mathcal{U}_0), \quad v = \left(\frac{hv}{h}\right) \in C^2(M_{\text{wet}}, \mathbb{R}), \quad \|\partial_x^2 v\|_{L^\infty(M_{\text{wet}})} < c_3.\]

Furthermore, let there be constants \(\varepsilon > 0, c_2 > 0\) such that for all \((x_M, t_M) \in M_{\text{wet}} \cap M_{\text{dry}}\) if \((x, t) \in M_{\text{wet}}\) with \(|x - x_M| \leq \varepsilon\) it follows that \(h(x, t_M) \geq c_2|x - x_M|\). With the velocities \(\sigma'_i\) of the wetting-drying interface we need to adapt the time step size such that additionally

\[
\max_{k=1, \ldots, N_1} \|\sigma'_k\|_{L^\infty} \Delta t \leq \Delta x.
\]

Then, for the well-balanced fv scheme operator \(Q^*\) and \(w^n(x, t) := w^n_i\) for \(x \in C_i\) and \(t^n \leq t < t^{n+1}\) the following consistency estimate holds.

\[
\|w^{n+1} - Q^* w^n\|_{L^1(\Omega)} = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta t\Delta x)
\]

For the proof of Theorem 3.2 we need the following two lemmata.

**Lemma 3.3.** If \(b \in C^2(\Omega), \|\partial_x^2 b\|_{L^\infty} \leq c_1\) for some constant \(c_1\) and \(w = (h, hv)^T \in C^{0,1}\), then we obtain for the discrete source term \(S^n_i\) which is defined according to (6), but on the basis of the \(w^n_i\)'s, that

\[S^n_i = \begin{pmatrix} 0 \\ -g h(x_i, t^n) \partial_x b(x_i) \end{pmatrix} + \mathcal{O}(\Delta x).\]

**Lemma 3.4.** Let \(L := \max(||\partial_x^2 b||_{L^\infty}, \text{Lip}(h))\) and the assumptions of Theorem 3.2 hold, then for all \(i \in I\) with \(h^n_i > 2L\Delta x\) the values \(w^n_{i\pm 12}\), which are reconstructed from \(w^n_{i-1, i+1, i+1, i+1}\), fulfill

\[w^n_{i+\frac{1}{2}} - w^n_{i-\frac{1}{2}} = w^n_i - w^n_{i-1} + \mathcal{O}(\Delta x^2), \quad w^n_{i+\frac{1}{2}} - w^n_{i-\frac{1}{2}} = w^n_{i+1} - w^n_i + \mathcal{O}(\Delta x^2).\]

**Proof of Lemma 3.3.** The key step in this proof is a case differentiation between the cases A: \(h^n_i > 2L\Delta x\) and B: \(h^n_i \leq 2L\Delta x\), where \(L := ||\partial_x b||_{L^\infty}\). This is necessary in order to resolve the maxima in the definition of the reconstructed values. In case A we consider the subcases A.1: \(\partial_x b(x_i) \neq 0\) and A.2: \(\partial_x b(x_i) = 0\) and apply Taylor expansion in order to prove the statement. In case B we immediately get the statement because all terms are of the order \(\mathcal{O}(\Delta x)\).

**Proof of Lemma 3.4.** Condition \(h^n_i > 2L\Delta x\) guarantees positive water height, and thus class \(C^2\), for the Lipschitz-continuous water height \(h(\cdot, t^n)\) on the neighbouring grid cells. Consequently, the denominators that occur in the definition of the reconstructed values are positive, as well. We consider the cases A: \(\partial_x b > 0\) or \(\partial_x b < 0\) on \(C_{i-1} \cup C_i \cup C_{i+1}\) and B: \(\exists x^* \in C_{i-1} \cup C_i \cup C_{i+1}\) with \(\partial_x b(x^*) = 0\). Due to the monotonicity in case A we can resolve the inner maxima, Taylor expansion of \(b\) resolves the outer maxima of the reconstructed values. Applying Taylor expansion and the boundedness of \(\partial_x b\) and \(\partial_x v\) lead to the statement in both cases.

**Proof of Theorem 3.2.** As in the proof of Lemma 3.3, this proof is based on differentiating between the cases A: \(h^n_i > 2L\Delta x\) and B: \(h^n_i \leq 2L\Delta x\), where \(L := \max(||\partial_x^2 b||_{L^\infty}, \text{Lip}(h))\). We start with case A, where obviously \(h > \frac{1}{2}L\Delta x\) and consequently \(h\) is of class \(C^2\) on \([x_i-\frac{1}{2}, x_i+\frac{1}{2}]\). From condition (7) we deduce that \(w\) is
of class $C^2$ on $[x_{i-1/2}, x_{i+1/2}] \times [t^n, t^{n+1}]$. By standard argumentation (see e.g. [9]), Lemma 3.3 and the fact that $w^n_{i+1/2} = w_i^n + O(\Delta x)$ it follows that

$$g(u^n_{i+1/2}, w^n_{i+1/2}) - g(u^n_{i-1/2}, w^n_{i-1/2}) = \nabla g(w^n_i, w^n_i) \cdot (w^n_i - w^n_{i-1})$$

$$+ \nabla g(w^n_i, w^n_i) \cdot (w^n_{i+1} - w^n_i) + O(\Delta x^2).$$

Invoking the consistency of the numerical flux function $g$ we conclude

$$g(u^n_{i+1/2}, w^n_{i+1/2}) - g(u^n_{i-1/2}, w^n_{i-1/2}) = \Delta x \frac{\partial F}{\partial x}(w^n_i) + O(\Delta x^2).$$

On the other hand, due to the regularity of $w$, we can combine the fact that $w^{n+1}_i - w^n_i = \partial_t w(x_i, t^n) \Delta t + O(\Delta t^2)$, the shallow water equations\footnote{Indeed, in wet regions equations (1)-(2) hold exactly since $w$ is $C^2$ here and a weak solution.} and Lemma 3.3 to get

$$w^{n+1}_i - Q^* w^n_i = O(\Delta x^2)$$

and thus $\|w^{n+1}_i - Q^* w^n_i\|_{L^1(C_i)} = |C_i| O(\Delta x^2)$. In case B, due to the smallness of $h^n_i$ and consequently of all relevant discrete quantities, it is easy to deduce that $\|w^{n+1}_i - Q^* w^n_i\|_{L^1(C_i)} = \|w^{n+1}_i - w^n_i + O(\Delta x)\|_{L^1(C_i)} = O(\Delta x^2)$. A combination of both cases and summation over $i \in I$ completes the proof. Note that for case B we use the fact that the number of cells which correspond to this case is bounded uniformly w.r.t. $\Delta x$ due to our assumptions.

\[\square\]

**Remark 2.** In [3] the well-balanced fv scheme is generalized to the case of unstructured dual grids in 2D and to second order. In [14] this approach is transferred to two-dimensional unstructured grids, which are not dual grids. The generalization of the method from [2] to fv schemes of arbitrary order of accuracy was done in [13].

4. **Wetting and Drying.** The shallow water equations are derived for wet domains, but the case $h = 0$ is of particular practical interest (e.g. dambreaks, tsunamis). From a computational point of view there are two resulting problems. Firstly, there may be very small values for the water height which occur as denominators in the computation of velocity and thus of the eigenvalues. Secondly, negative water heights might occur. Eventually, we seek a fv scheme which incorporates wetting and drying and is well-balanced, as well.

4.1. **Desingularized Water Heights** [10]. In order to deal with small water heights in nearly dry areas one can instead of computing the velocity by

\[v = \frac{hv}{h + v}\]

which might become singular for $h << 1$, use a so-called desingularized velocity

$$v^d := \frac{\sqrt{2h(\frac{h}{h + v})}}{\sqrt{h + \max(h, \epsilon)}}$$

for a fixed small $\epsilon > 0$. This approach is due to [10].

4.2. **Wetting & Drying Algorithm** [4]. A wetting and drying approach which guarantees nonnegativity of the discrete water height and is applicable to the well-balancing modification from [2] is introduced in [4]. The basic idea is to restrict the water outflow of each grid cell according to the amount of water that is inside the cell at the current time step. This is realized by introducing local time step sizes which are defined in the way that the cell would run dry exactly if there was no water inflow.

For a grid cell $C_i$ with numerical fluxes $g_{i+1/2}^\pm := g(u^n_{i+1/2}, w^n_{i+1/2})$, the outflow (of water) is obviously given by positive $g^h_{i+1/2}^+$ or negative $g^h_{i+1/2}^-$, where $g^h_{i+1/2}^+$ denotes
the first (water height) component of $g_{i+\frac{1}{2}+}$. The local time step sizes $\Delta t_i$ (for the $i$-th grid cell) and $\Delta t_{i+\frac{1}{2}}$ (for the interface $x_{i+\frac{1}{2}}$) are defined by

$$\Delta t_i := \begin{cases} \Delta t & \text{if } Q^+ u_i^n \geq 0, \\ \frac{\Delta t}{|C_i|^n h_i^n} & \text{else}, \end{cases} \quad \Delta t_{i+\frac{1}{2}} := \begin{cases} \Delta t_i & \text{if } g_{i+\frac{1}{2}+}^h \geq 0, \\ \Delta t_{i+1} & \text{if } g_{i+\frac{1}{2}+}^h < 0. \end{cases}$$

where $g^\text{outflow} := \max(g_{i+\frac{1}{2}+}^h, 0) - \min(g_{i-\frac{1}{2}+}^h, 0)$.

**Definition 4.1** (Well-Balanced Finite Volume Scheme with Wetting & Drying). The well-balanced fv scheme operator $Q^+$ with wetting and drying is defined by

$$Q^+ u_i^n := u_i^{n+1} := u_i^n - \frac{1}{\Delta x} \left( \Delta t_{i+\frac{1}{2}} g_{i+\frac{1}{2}+}^h - \Delta t_{i-\frac{1}{2}} g_{i-\frac{1}{2}+}^h \right) + \Delta t S_i^n.$$

Note that $\Delta t_{i+\frac{1}{2}}$ only differs from $\Delta t$ if one of the neighbouring cells would otherwise develop a negative water height.

**Remark 3.** By construction the operator $Q^+$ preserves nonnegativity of the water height, is well-balanced and conservative.

5. **Numerical Experiments.** For the presentation of the well-balancing and wetting and drying modifications we restrict ourselves to the onedimensionl case and to fv schemes of first order. However, a second order scheme in 2D was implemented and validated, cf. Subsection 5.1. In Subsection 5.2 we will present a fv scheme on closed surfaces and apply it to the shallow water equations on the sphere.

5.1. **Validation.** For the validation of the fv scheme we compare the numerical solution which is defined according to the generalization of $Q^+$ to second order schemes and two dimensions (cf. [14] for details) to the exact solution of the shallow water equations in a parabolic bowl [15]. The experimental orders of convergence are shown in Table 1.

5.2. **Application to Shallow Water Equations on the Rotating Sphere with Realistic Ocean Bathimetry.** Systems of conservation laws usually consist of conservation of scalar quantities (e.g. mass, energy, height) and conservation of momentum. While it is not trivial to define conservation of momentum on surfaces\(^3\), conservation of scalar quantities on surfaces was treated in [7, 11, 12].

From a numerical point of view a conservation property for the numerical flux function of the form $g_{e,C} (u_j^n, u_{l(j,e)}^n) = -g_{e,C_{l(j,e)}} (u_{l(j,e)}^n, u_j^n)$ for numerical flux\(^3\)

\[^3\] Note that at any place the momentum vector should lay in the respective tangent plane.
functions $g$ that are defined for an edge $e \subset \partial C_j$ of a grid cell $C_j$ with neighbouring grid cell $C_{l(j,e)}$ cannot be expected. This is due to the fact that the cells $C_i$ and $C_{l(j,e)}$ have different tangent planes and the discrete momentum should be contained in the respective tangent plane. Instead of taking into account the fact that the momentum is forced to be tangential is by Lagrange multipliers $[8]$ we propose the following approach, which is applicable to a large class of surfaces, even whose boundary might not be $C^1$.

1. Have an orthonormal basis of the tangent space for every grid cell.
2. Save momentum approximation in coordinates of the cell dependent basis.
3. For the flux computation we
   (a) rotate the tangent plane of the neighbouring cell in the plane of the current cell,
   (b) compute neighbour’s momentum components in the basis of the current grid cell’s tangent space,
   (c) apply a standard numerical flux function (e.g. Lax-Friedrichs, HLLE).

In the following we want to show the application of the described fv scheme for surfaces to the case of the shallow water equations on a rotating sphere with realistic bottom topography. The rotation is taken into account by adding a Coriolis force term on the right hand side. The realistic earth topography and ocean bathimetry is taken from $[1]$. The initial values, that describe the effect of an asteroid impact, are due to $[16]$. The simulation of a tsunami caused by an asteroid impact near the chilean coast is shown in Figure 1. For this simulation we added adaptivity to the numerical scheme with an heuristic adaption criterion that is based on the magnitude of jumps in water height, momentum and bottom topography.

5.3. Implementation Aspects. All Simulations have been performed within the DUNE-FEM module which is based on the Distributed and Unified Numerics Environment (DUNE) $[6]$. Most of the simulation have been performed in parallel, especially the simulations including the wetting and drying algorithm from 4.2. For more details on the implementation, confer $[14]$.

Acknowledgments. We would like to thank Christoph Gersbacher for writing a parser which transforms the ETOPO data $[1]$ into a DUNE-FEM-readable format.

REFERENCES

Figure 1. Asteroid impact tsunami near the Chiliean cost with adaptive mesh refinement and realistic bottom topography.

Approximation of the Effective Hamiltonian through a Degenerate Elliptic Problem

Martin Nolte
Abteilung für Angewandte Mathematik
Universität Freiburg
Hermann-Herder-Straße 10
79104 Freiburg, Germany

Abstract. For strongly convex Hamiltonians, an interesting approximation to the effective Hamiltonian through a highly nonlinear elliptic problem was introduced in [2]. It is very appealing to use this approximation to build numerical schemes. Numerical results indicate, however, that it is more promising to devise numerical methods for the viscosity solution of the limiting equation.

1. Introduction. The effective Hamiltonian plays an important role in the homogenization of nonlinear PDEs, see e.g., [5]. For a Hamiltonian $H \in C^0(T^d, \mathbb{R}^d)$, its effective Hamiltonian $\bar{H} : \mathbb{R}^d \rightarrow \mathbb{R}$ evaluated at some point $P \in \mathbb{R}^d$ is defined as the unique constant $\bar{H}(P)$ such that the cell problem

$$H(x, P + \nabla u) = \bar{H}(P) \quad \text{in } T^d$$

has a viscosity solution $u \in C^0(T^d)$.

Unfortunately, analytical expressions of the effective Hamiltonian are only known in rare cases. Therefore, its numerical approximation has seen quite some activity, see, e.g., [3, 4, 6, 7, 8, 9].

In this article, we will restrict ourselves to strongly convex $C^2$-Hamiltonians, i.e., there is a constant $\alpha > 0$ such that $\xi^T H_{pp}(x, p) \xi \geq \alpha |\xi|^2$ for all $x \in T^d$ and all $p, \xi \in \mathbb{R}^d$. As shown in [1], the effective Hamiltonian $\bar{H}$ is then given by the formula

$$\bar{H}(P) = \inf_{u \in C^\infty(T^d)} \max_{x \in T^d} H(x, P + \nabla u(x)). \quad (1)$$

A straight-forward discretization of (1) was applied in [3], leading to a large minimax problem. Here, we follow a different approach introduced by Evans in [2], where (1) is approximated by the regularized minimization problem

$$\bar{H}_k(P) := \inf_{u \in C^\infty(T^d)} \frac{1}{k} \ln \left[ \int_{T^d} e^{kH(x, P + \nabla u(x))} \, dx \right]. \quad (2)$$

It was shown that $\bar{H}_k(P) \rightarrow \bar{H}(P)$ as $k \rightarrow \infty$. Indeed, a minimizer in (2) exists within the Sobolev space $H^1(T^d)$ and it is a weak solution of the Euler-Lagrange equation

$$\nabla \cdot \left( e^{kH(x, P + \nabla u_k)} H_p(x, P + \nabla u_k) \right) = 0 \quad \text{in } T^d. \quad (3)$$

Notice that the shifted Hamiltonian $(x, p) \mapsto H(x, P + p)$ also satisfies our assumptions. In the following, we will, therefore, assume without loss of generality that $P = 0$ and omit the argument of the effective Hamiltonian.
The structure of this paper is as follows: In Section 2 we briefly describe a finite element discretization of (3). Using a finite difference approximation, we get rid of the exponential $e^{kH(x,\nabla u)}$ in Section 3. To circumvent a restriction of type $k \leq C(\Delta x)$, we stabilize this scheme in Section 4 and briefly discuss the limit $k \to \infty$. Section 5 then provides numerical evidence of the superiority of the stabilized finite difference scheme over the direct discretizations of (3). We finish with some conclusions in Section 6.

2. Standard Finite Elements for the Euler-Lagrange Equation. Today, the standard approximation of weak solutions to (3) are Lagrangian finite elements. For some $N \in \mathbb{N}$, $\Delta x = \frac{1}{N}$ and $I_{i-\frac{1}{2}} := [(i-1)\Delta x, i\Delta x]$, we use a periodic, Cartesian grid $\mathcal{G} = \{I_{i-\frac{1}{2}} \times I_{j-\frac{1}{2}} \mid i, j \in \mathbb{Z}/NZ\}$ of the flat torus $\mathbb{T}^d$ together with the finite element space

$$X_G = \{u \in C^0(\mathbb{T}^d) \mid u|_E \in Q^1(E) \ \forall E \in \mathcal{G}, (u,1)_{L^2(\mathbb{T}^d)} = 0\},$$

where $Q^1(E)$ denotes the space of bilinear polynomials on $E$. We then seek a discrete solution $u_G \in X_G$ satisfying

$$a(u_G, \varphi) := \int_{\mathbb{T}^d} e^{kH(\cdot, \nabla u_G)} P_p(\varphi, \nabla u_G) = 0 \ \forall \varphi \in X_G. \quad (4)$$

The approximate effective Hamiltonian $\tilde{H}_{k,\Delta x}$ is then given by

$$\tilde{H}_{k,\Delta x}(P) = \tilde{H}_{k,G}(P) = \frac{1}{k} \ln \int_{\mathbb{T}^d} e^{kH(\cdot, \nabla u_G) - z} + z,$$

where $z \in \mathbb{R}$ is arbitrary.

The linearization $\delta a[u]$ of $a$ at some point $u \in X_G$ reads

$$\delta a[u](v, \varphi) = \int_{\mathbb{T}^d} e^{kH(\cdot, \nabla u_G)} \nabla^T v \left(k H_p(\cdot, \nabla u_G) H_p^T(\cdot, \nabla u_G) + H_{pp}(\cdot, \nabla u_G)\right) \nabla \varphi.$$

Applying Newton’s method, we need to solve the linear system

$$\delta a[u^n](\delta u^n, \varphi) = a(u^n, \varphi) \ \forall \varphi \in X_G$$

in each Newton step. This system is ill-conditioned due to the factor $e^{kH(\cdot, \nabla u_G)}$ and the scheme usually fails to converge for $k \geq 10$ for an initial guess of $u^0 \equiv 0$.

Stability can be improved by solving (4) for increasing $k_0 < \ldots < k_M$, each time using the previous solution as initial guess for the Newton scheme. For large $k$, however, the steps $k_{m+1} - k_m$ have to be chosen extremely small.

3. A Finite Difference Approximation to the Euler-Lagrange Equation. In [9], a finite difference discretization of (3) is applied to simple Hamiltonians of the form

$$H(x, p) = \frac{1}{2} |p|^2 + E(x). \quad (5)$$

In this section, we extend this scheme to general strongly convex $C^2$-Hamiltonians. For smooth solutions $u$, the Euler-Lagrange equation (3) is equivalent to

$$(H_p(x, \nabla u) H_p^T(x, \nabla u) + \frac{1}{k} H_{pp}(x, \nabla u)) \colon \nabla^2 u$$

$$+ H_s(x, \nabla u) H_p(x, \nabla u) + \frac{1}{k} \text{tr} H_{xp}(x, \nabla u) = 0 \quad \text{in } \mathbb{T}^d, \quad (6)$$

where $\nabla^2 u$ denotes the Hessian of $u$, $\text{tr} A = \sum_i A_{ii}$ denotes the trace of a matrix $A$ and $A : B = \sum_{ij} A_{ij} B_{ij}$ denotes the Frobenius inner product of two matrices $A$ and $B$. 
and $B$. Note that for Hamiltonians of the special form (5), $H_{xp}$ vanishes and (6) reduces to
\[(\nabla u \nabla u^T + \frac{1}{k} \mathbb{I}) : \nabla^2 u + E(x) \nabla u = 0.\]

For some $N \in \mathbb{N}$, we denote the mesh width $\Delta x = \frac{1}{N}$ and the mesh points $x_{ij} := (i\Delta x, j\Delta x)$. To obtain a finite difference approximation $u_{ij} \approx u(x_{ij})$ of (6), we approximate the derivatives $\nabla_{\Delta x} u_{ij} \approx \nabla u(x_{ij})$ and $\nabla_{\Delta x}^2 u_{ij} \approx \nabla^2 u(x_{ij})$ by the finite differences
\[
\begin{align*}
\partial_{x}^{\Delta x} &= \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x}, \\
\partial_{x}^{2\Delta x} &= \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{\Delta x^2}, \\
\partial_{y}^{\Delta x} &= \frac{u_{i+1,j+1} - u_{i+1,j-1} - u_{i-1,j+1} + u_{i-1,j-1}}{4\Delta x^2}.
\end{align*}
\]

Using the notation
\[
\begin{align*}
A(x, p) &:= H_p(x, p) H_p^T(x, p) + \frac{1}{k} H_{pp}(x, p), \\
b(x, p) &:= H_x^T(x, p) H_p(x, p) + \frac{1}{k} \text{tr} H_{xp}(x, p),
\end{align*}
\]
we obtain the following discrete form of (6):
\[
A(x_{ij}, \nabla_{\Delta x} u_{ij}) : \nabla_{\Delta x}^2 u_{ij} + b(x_{ij}, \nabla_{\Delta x} u_{ij}) = 0 \quad \text{for all } i, j \in \mathbb{Z}/N\mathbb{Z}.
\]

As only the diagonal of $\nabla_{\Delta x}^2 u_{ij}$ depends on $u_{ij}$, we can use (7) to express $u_{ij}$ in terms of its 8 neighboring values $u_{i\pm 1, j}, u_{i, j\pm 1},$ and $u_{i\pm 1, j\pm 1}$. Starting from initial values $u^0_{ij}$, we can then use this relation in a fixed point iteration:
\[
u_{ij}^{n+1} := \frac{\Delta x}{2\text{tr} A(x_{ij}, \nabla_{\Delta x} u_{ij})} \left[ A(x_{ij}, \nabla_{\Delta x} u_{ij}^n) : \left( \nabla_{\Delta x}^2 u_{ij}^n + \frac{2u_{ij}^n}{\Delta x^2} \mathbb{I} \right) + b(x_{ij}, \nabla_{\Delta x} u_{ij}^n) \right].
\]

If this sequence converges to some $\bar{u}_{ij}$, an approximate effective Hamiltonian can be obtained by:
\[
\bar{H}_{k, \Delta x} := \bar{z} + \frac{1}{k} \ln \left( \Delta x^2 \sum_{i,j=1}^N e^{k(H(x_{ij}, \nabla_{\Delta x} \bar{u}_{ij}) - \bar{z})} \right),
\]
where $\bar{z} = \max_{i,j} H(x_{ij}, \nabla_{\Delta x} \bar{u}_{ij})$. This formula is also used in [9]. Notice that $\bar{z}$ itself can also be used as an approximation to the effective Hamiltonian $\bar{H}_k$.

Unfortunately, numerical experiments show that this scheme is unstable unless $k \leq C(\Delta x)$. As an example, consider the modified uncoupled penduli problem introduced in 5.1. The distance
\[\|u^{n+1} - u^n\| := \max_{i,j \in \mathbb{Z}/N\mathbb{Z}} |u_{ij}^{n+1} - u_{ij}^n|\]
between successive iterates for different values of $k$ on a fixed mesh is displayed in Figure 1. As we see, the sequence $(u_{ij}^n)_{n \in \mathbb{N}}$ fails to converge for large values of $k$.

4. Stabilizing the Finite Difference Scheme. In this section, we exploit another result obtained in [2] to stabilize the scheme: The minimizers $u_k$ in (2) are continuous and converge to some $u \in C(\mathbb{T}^d)$ as $k \to \infty$. The limit $u$ is a viscosity solution of the equation
\[-H_p(x, \nabla u) H_p^T(x, \nabla u) : \nabla^2 u = H_x^T(x, \nabla u) H_p(x, \nabla u) \quad \text{in } \mathbb{T}^d.\]
This indicates that we might stabilize the scheme by simply adding some mesh-dependent viscosity term to (7):

$$\left(A(x_{ij}, \nabla_{\Delta x} u_{ij}) + \frac{\Delta x}{2} \beta_{ij} \right): \nabla_{\Delta x}^2 u_{ij} + b(x_{ij}, \nabla_{\Delta x} u_{ij}) = 0,$$

where

$$\beta_{ij} = \max \{ \beta_{\text{min}}, |\nabla_{\Delta x} (H_T(x_{ij}, \nabla_{\Delta x} u_{ij}) H_p(x_{ij}, \nabla_{\Delta x} u_{ij}))| \}.$$  

In our numerical experiments, we choose $\beta_{\text{min}} = 10^{-3}$.

It turns out that this additional viscosity term does indeed stabilize the scheme and we can pass to the limit $k \to \infty$ in the discretization, using the formula

$$\bar{H}_{\Delta x} = \max_{i,j \in \mathbb{Z}/N} H(x_{ij}, \nabla_{\Delta x} u_{ij})$$

to approximate the effective Hamiltonian.

5. Numerical Results. In this section we apply the stabilized scheme to two benchmark problems, where the effective Hamiltonian is known analytically: A slight variation of the Hamiltonian for two uncoupled penduli and a Hamiltonian introduced in [5] that is not of the form (5).

5.1. The Modified Uncoupled Penduli Problem. A frequently used test problem for the approximation of the effective Hamiltonian is the Hamiltonian for two uncoupled penduli:

$$H(x, p) = \frac{1}{2} |p|^2 - \cos(2\pi x_1) - \cos(2\pi x_2) = H_{1d}(x_1, p_1) + H_{1d}(x_2, p_2),$$

where $H_{1d}(x, p) = \frac{1}{4} |p|^2 - \cos(2\pi x)$. The effective Hamiltonian inherits this structure, i.e., $H(p) = \bar{H}_{1d}(p_1) + \bar{H}_{1d}(p_2)$. The effective Hamiltonian $\bar{H}_{1d}$ is plotted in Figure 2. It is a continuous function with $\bar{H}_{1d}(P) = \bar{H}_{1d}(-P)$, differentiable except in $P = \pm \frac{\pi}{4}$ and constantly 1 in $[-\frac{\pi}{4}, \frac{\pi}{4}]$. An analytical expression for $\bar{H}_{1d}$ is, e.g., given in [3].

Unfortunately, minimizers $u \in C^{0,1}(\mathbb{T}^2)$ in (1) also inherit this alignment to the coordinate axes; they are of the form $u(x) = u_1(x_1) + u_2(x_2)$. As our finite
difference approximations share this alignment to the coordinate axes, this can have an undesired positive effect on the convergence.

To circumvent this difficulty, we use a slight modification of (10):

\[ H(x, p) = \frac{1}{10} |p|^2 - \cos(2\pi(2x_1 + x_2)) - \cos(2\pi(2x_2 - x_1)) \]

The exact effective Hamiltonian is then given by

\[ H(p) = \tilde{H}_d\left(\frac{1}{\delta}(2p_1 + p_2)\right) + \tilde{H}_d\left(\frac{1}{\delta}(2p_2 - p_1)\right). \]

Let us start the discussion by taking a look at the approximation error caused by the finite element approximation (4) for fixed, fairly large \( k = 640 \) in \( P_1 = \left(\frac{12}{5}, \frac{4}{5}\right) \) (left) and \( P_2 = \left(\frac{24}{5}, \frac{8}{5}\right) \) (right):

<table>
<thead>
<tr>
<th>( N )</th>
<th>( \tilde{H}_{k,\Delta x} )</th>
<th>Error</th>
<th>EOC_{\Delta x}</th>
<th>( N )</th>
<th>( \tilde{H}_{k,\Delta x} )</th>
<th>Error</th>
<th>EOC_{\Delta x}</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2.112740</td>
<td>1.13 · 10^{-1}</td>
<td>0.01</td>
<td>8</td>
<td>3.486407</td>
<td>7.20 · 10^{-1}</td>
<td>--</td>
</tr>
<tr>
<td>16</td>
<td>1.983583</td>
<td>1.64 · 10^{-2}</td>
<td>0.01</td>
<td>16</td>
<td>3.150795</td>
<td>3.85 · 10^{-1}</td>
<td>0.91</td>
</tr>
<tr>
<td>32</td>
<td>1.986904</td>
<td>1.31 · 10^{-2}</td>
<td>0.03</td>
<td>32</td>
<td>2.960169</td>
<td>1.94 · 10^{-1}</td>
<td>0.99</td>
</tr>
<tr>
<td>64</td>
<td>1.987030</td>
<td>1.30 · 10^{-2}</td>
<td>0.01</td>
<td>64</td>
<td>2.860932</td>
<td>9.47 · 10^{-2}</td>
<td>1.03</td>
</tr>
</tbody>
</table>

The tables show the approximate effective Hamiltonian \( \tilde{H}_{k,\Delta x} \), the error \( e_{k,\Delta x} := |\tilde{H}_{k,\Delta x} - \tilde{H}_d| \), and its experimental order of convergence (w.r.t. \( \Delta x \)) \( \text{EOC}_{\Delta x} = \ln \frac{e_{k,\Delta x}}{e_{k,\Delta x/2}} / \ln 2 \). In the point \( P_1 \), which resides in the flat region of \( \tilde{H}_d \), the error is dominated by the approximation in \( k \), while in the point \( P_2 \), residing outside this flat region, we observe first order convergence, i.e., the error is dominated by the spatial approximation.

The approximation error using the finite difference approximation (7) on a fixed mesh with \( \Delta x = \frac{1}{32} \) in \( P_1 \) (left) and \( P_2 \) (right) are displayed in the following tables:

<table>
<thead>
<tr>
<th>( k )</th>
<th>Its</th>
<th>( \tilde{H}_{k,\Delta x} )</th>
<th>Error</th>
<th>EOC_{\Delta x}</th>
<th>( k )</th>
<th>Its</th>
<th>( \tilde{H}_{k,\Delta x} )</th>
<th>Error</th>
<th>EOC_{\Delta x}</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>950</td>
<td>1.366458</td>
<td>6.34 · 10^{-1}</td>
<td>--</td>
<td>5</td>
<td>1064</td>
<td>2.749102</td>
<td>1.71 · 10^{-2}</td>
<td>--</td>
</tr>
<tr>
<td>10</td>
<td>21 220</td>
<td>1.594824</td>
<td>4.05 · 10^{-1}</td>
<td>0.64</td>
<td>10</td>
<td>1607</td>
<td>2.758863</td>
<td>7.37 · 10^{-3}</td>
<td>1.22</td>
</tr>
<tr>
<td>20</td>
<td>3 932</td>
<td>1.759641</td>
<td>2.40 · 10^{-1}</td>
<td>0.75</td>
<td>20</td>
<td>2234</td>
<td>2.766147</td>
<td>8.54 · 10^{-5}</td>
<td>6.43</td>
</tr>
<tr>
<td>40</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>40</td>
<td>2 783</td>
<td>2.773524</td>
<td>7.29 · 10^{-3}</td>
<td>--6.42</td>
</tr>
<tr>
<td>80</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>80</td>
<td>3 154</td>
<td>2.782377</td>
<td>1.61 · 10^{-2}</td>
<td>--1.15</td>
</tr>
</tbody>
</table>

They show the number of iterations, the approximate effective Hamiltonian \( \tilde{H}_{k,\Delta x} \), the error \( e_{k,\Delta x} := |\tilde{H}_{k,\Delta x} - \tilde{H}_d| \), and its experimental order of convergence (w.r.t. \( \Delta x \)).
(10) $\text{EOC}_k = \ln \frac{e_k \Delta x}{e_{k/2} \Delta x} / \ln 2$. In contrast to the finite element approximation, the scheme converges only for small values of $k$ within the flat region of $\bar{H}_{1d}$. Outside this flat region, the approximation errors seem comparable to those of the finite element discretization.

Using the stabilized finite difference approximation (9) with $k = \infty$, we obtain the following values in $P_1$ (left) and $P_2$ (right):

<table>
<thead>
<tr>
<th>N</th>
<th>Its</th>
<th>$\bar{H}_{\Delta x}$</th>
<th>Error</th>
<th>EOC$_{\Delta x}$</th>
<th>N</th>
<th>Its</th>
<th>$\bar{H}_{\Delta x}$</th>
<th>Error</th>
<th>EOC$_{\Delta x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>154</td>
<td>2.176541</td>
<td>1.77 $\cdot 10^{-1}$</td>
<td>1.64</td>
<td>8</td>
<td>240</td>
<td>3.478439</td>
<td>7.12 $\cdot 10^{-1}$</td>
<td>1.05</td>
</tr>
<tr>
<td>16</td>
<td>480</td>
<td>2.023333</td>
<td>2.33 $\cdot 10^{-2}$</td>
<td>2.92</td>
<td>16</td>
<td>761</td>
<td>3.128889</td>
<td>3.63 $\cdot 10^{-1}$</td>
<td>0.97</td>
</tr>
<tr>
<td>32</td>
<td>3071</td>
<td>2.000075</td>
<td>7.48 $\cdot 10^{-5}$</td>
<td>8.28</td>
<td>32</td>
<td>1750</td>
<td>2.950862</td>
<td>1.85 $\cdot 10^{-1}$</td>
<td>0.97</td>
</tr>
<tr>
<td>64</td>
<td>391208</td>
<td>2.000000</td>
<td>1.98 $\cdot 10^{-10}$</td>
<td>18.53</td>
<td>64</td>
<td>5945</td>
<td>2.859892</td>
<td>9.37 $\cdot 10^{-2}$</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Here, we observe the opposite effect: The approximation within the flat region is very good, while it is only acceptable outside this region. On the other hand, a huge number of iterations is required in $P_1$. In both cases, however, we observe at least linear convergence, indicating the robustness of the scheme.

![Figure 3](image_url)

**Figure 3.** Approximate effective Hamiltonian for the modified uncoupled penduli problem obtained by (9) with $k = \infty$ and $\Delta x = \frac{1}{32}$: contour lines of $\bar{H}_{\Delta x}$ (left), the error $|\bar{H}_{\Delta x} - \bar{H}|$ (right).

The approximation to the effective Hamiltonian given by the stabilized finite difference method (9) with $k = \infty$ on a mesh with $\Delta x = \frac{1}{32}$ is displayed in Figure 3. The dashed lines indicate the position of the kinks in the two one-dimensional effective Hamiltonians $\bar{H}_{1d}$ and the points $P_1$ and $P_2$ are shown. In the lower left part, the function should be constant, in the lower right and upper left parts, the contours should be straight lines, and in the upper right part they are curves connecting these lines. Observe that this structure is reproduced quite well; the maximum error of approximately 0.25 is located at the edge of the flat region. Both discretizations of the Euler-Lagrange equation (3) do not achieve this level of accuracy (see [6]).

5.2. A Simple Shear Flow. In [5], homogenization of an advection-diffusion-reaction equation with nonlinear reaction leads to Hamiltonians of the form

$$H(x, p) = \kappa |p|^2 + v(x) \cdot p$$
for some continuous velocity field $v : \mathbb{R}^2 \to \mathbb{R}^2$. For a simple shear flow of the form $v(x) = \bar{v} + v_s(x_2)e_1$, the effective Hamiltonian can be calculated analytically (see also [5]).

For our numerical experiments, we choose $\kappa = 1$, $\bar{v} = \frac{1}{5} (1, 3)^T$, and $v_s(y) = \frac{1}{5} \cos(2\pi y)$, resulting in the Hamiltonian

$$H(x, p) = |p|^2 + \frac{1}{5} \left( (1 + \cos 2\pi x_2) p_1 + 3 p_2 \right).$$

Figure 4. Approximate effective Hamiltonian for the simple shear flow problem obtained by (9) with $k = \infty$ and $\Delta x = \frac{1}{32}$: contour lines of $\bar{H}_{\Delta x}$ (left), the error $|\bar{H}_{\Delta x} - H|$ (right).

The approximate effective Hamiltonian obtained by the stabilized finite difference method (9) with $k = \infty$ on a mesh with $\Delta x = \frac{1}{32}$ is displayed in Figure 4. Again, the dashed line indicates the position of a kink in the exact effective Hamiltonian. As before, the structure is approximated very well; the kinks are clearly visible in the contour lines of $\bar{H}_{\Delta x}$.

6. Conclusions. It is indeed possible to apply schemes based on (3) to approximate the effective Hamiltonian. The presented numerical results indicate, however, that it is more promising to device efficient numerical schemes for the limiting equation (8). While the stabilized scheme described above seems to be quite robust, it also converges rather slowly and there seems to be room for improvement. We also need to keep in mind that it is still unclear, whether a viscosity solution to (8) is indeed a minimizer for the right hand side in (1).

REFERENCES

E-mail address: nolte@mathematik.uni-freiburg.de
NONLINEAR STABILITY OF BOUNDARY LAYER SOLUTIONS
TO THE EULER-POISSON EQUATIONS IN PLASMA PHYSICS

Shinya Nishibata\textsuperscript{1}, Masashi Ohnawa\textsuperscript{2} and Masahiro Suzuki\textsuperscript{1}

\textsuperscript{1}Department of Mathematical and Computing Sciences
Tokyo Institute of Technology
Tokyo 152-8552, Japan

\textsuperscript{2}Research Institute of Nonlinear Partial Differential Equations
Organization for University Research Initiatives, Waseda University
Tokyo 169-8555, Japan

Abstract. The main concern of the present paper is to study a boundary layer, called sheath, which appears over a material in contact with plasma. The Bohm criterion in plasma physics claims the velocity of positive ions should be faster than a certain value for the formation of a sheath. The behavior of the plasma is governed by the Euler-Poisson equations. Mathematically, we define the sheath by a monotone stationary solution to the system over a half space. We first derive conditions for the existence of the stationary solution and observe that the Bohm criterion with certain physically reasonable conditions are sufficient for the existence of the stationary solution. Then we discuss its asymptotic stability under the Bohm criterion by using the weighted energy method. We also obtain convergence rates subject to the spatial decay rates of the initial perturbation.

1. Introduction. We study the asymptotic stability of a boundary layer, which is called a sheath in plasma physics. We begin with a brief explanation of the formation process of the sheath. Plasma is often said to be the fourth state other than solid, liquid or gas, and is characterized by its high energy and sparse density. Due to its high energy, plasma is ionized into positively charged ions and negatively charged electrons. In the situations where there are no external forces, nor obstacles in contact with plasma, the plasma is believed to be electrically neutral. Now we consider the situation in which a wall is put in a plasma and is negatively charged. Once particles hit the wall, they accumulate on it. Because of the difference in the thermal velocities, more electrons accumulate on the wall compared to ions, leaving ion-rich region in the rest of the space. The potential gradient attracts positive ions to the wall while repelling electrons, acting to confine the region of biased distribution of two species close to the wall. If the fluxes of the electron and the ion to the wall coincide, the situation is stationary. This stationary boundary layer is called a sheath in plasma physics. Its typical thickness is given by the Debye length, which in most cases, is smaller than the characteristic length scales determined by such quantities as mean free path, the ionization length or the geometry of the system. Hence the sheath is considered to be collision free.

\textsuperscript{2000 Mathematics Subject Classification.} Primary: 35B40, 35M10, 76K05; Secondary: 35B35, 35B45.

\textit{Key words and phrases.} Bohm criterion, sheath, boundary layer, asymptotic behavior, convergence rate, weighted energy method.
The research of the sheath is initiated in the 1920s, and in the early pioneering work by Langmuir in [14], it is already recognized that the velocity of ions directing toward the sheath attains a certain amount of positive value. Later, Bohm in [4] studies the stationary problem and obtains the necessary condition for the formation of the sheath. This condition is now called the Bohm criterion. It claims that \( |u_i| \), the mean velocity of the incoming flow of positive ions to the sheath from the inner region, is equal to or greater than the ion sound speed \( c_s := \sqrt{(T_e + \gamma_i T_i)/m_i} \) in dimensional form, where \( \gamma_i \) is the ion polytropic coefficient, \( m_i \) is the ion mass, and \( T_e \) and \( T_i \) denote the temperature of electrons and ions, respectively.

The sheath is intensively studied in plasma physics since the process of sheath formation is important for nearly all applications where a plasma is in contact with material surfaces. In spite of its importance, the definition of the sheath has been left ambiguous because of the difficulty in the consistent understanding of plasma and sheath. Basic physical problems remain unanswered such as where the sheath-plasma boundary is located, how they are connected or how the ions are accelerated before entering the sheath. Readers are referred to [1, 12] for reviews and new insights concerning the sheath.

Few mathematical works have been done until in the middle of 1990’s. The earliest mathematical works [5, 6] present traveling wave analysis associated with the Euler-Poisson system. Later, [8, 9] research the dynamics of this system in the whole space around constant states for the fixed ion density model and the Boltzmann relation model (1) for electron density, respectively.

Concerning the effect of boundaries, Ambroso, Méhats and P.-A. Raviart in [3] show the existence of the monotone stationary solution to (1) under (9) over one-dimensional bounded domain. Later Ambroso in [2] numerically shows that the solution to (1) approaches the stationary solution as time tends to infinity in the same setting as [3]. Suzuki in [13] interprets the sheath as a monotone stationary solution to the system of Euler-Poisson equation (1) for one-dimensional half space and shows that the Bohm criterion together with the physically natural boundary condition on the electric potential is sufficient for the unique existence of a monotone stationary solution. Asymptotic stability of the stationary solution under (9) and also under (10) are proved in [11], giving mathematical justification to the Bohm criterion. The objective of the present paper is to summarize the results obtained in these two papers [11, 13].

2. Main results. The isothermal flow of positive ions is governed by the Euler-Poisson equations:

\[
\begin{align*}
\rho_t + \text{div}(\rho u) &= 0, \\
(\rho u)_t + \text{div}(\rho u \otimes u) + K \nabla \rho + \rho \nabla \phi &= 0, \\
-\Delta \phi &= \rho - e^\phi,
\end{align*}
\]

where unknown functions \( \rho, u \) and \( \phi \) stand for the density of positive ions, the velocity of positive ions and the electrostatic potential, respectively. The positive constant \( K \) corresponds to the temperature of ions.

The first equation describes the conservation of mass, and the second one is the equation of momentum in which the pressure gradient and the electrostatic potential gradient are taken into account. The third equation describes the relation between the potential, the ion density and the electron density. It is obtained by assuming the Boltzmann relation, that is, the electron density is given by \( \rho_e = e^\phi \).
We study the initial boundary value problem to (1) in the \( N \)-dimensional half space \( \mathbb{R}^N := \{ (x_1, \ldots, x_N) \in \mathbb{R}^N \mid x_1 > 0 \} \) for \( N = 1, 2, 3 \). The space coordinate is denoted by \( x = (x_1, \ldots, x_N) = (x_1, x') \), where \( x_1 \) and \( x' = (x_2, \ldots, x_N) \) are the normal and the tangential components, respectively. The initial and the boundary data are prescribed as
\[
(\rho, u)(0, x) = (\rho_0, u_0)(x) \quad \text{with} \quad \inf_{x \in \mathbb{R}^N} \rho_0(x) > 0, \quad (2)
\]
\[
\lim_{x_1 \to \infty} (\rho_0, u_0)(x_1, x') = (\rho_+, u_+, 0, \ldots, 0) \in \mathbb{R}^{N+1} \quad \text{for an arbitrary} \quad x' \in \mathbb{R}^{N-1}, \quad (3)
\]
where \( \rho_+, u_+ \) and \( \phi_0 \) are constants. We take a reference point of the potential \( \phi \) as \( x_1 = \infty \), that is,
\[
\lim_{x_1 \to \infty} \phi(t, x_1, x') = 0 \quad \text{for an arbitrary} \quad x' \in \mathbb{R}^{N-1}. \quad (5)
\]
The well-posedness of the problem is addressed later. To construct a classical solution of the Poisson equation (1c), the condition
\[
\rho_+ = 1 \quad (6)
\]
is necessary. Owing to conditions (5) and (6), the quasi-neutrality \( \rho = \rho_c \) holds as \( x_1 \to \infty \) since \( \lim_{x_1 \to \infty} \rho_0 = \rho_c = e^0 = \rho_+ \).

The planar stationary solution \((\tilde{\rho}, \tilde{U}, \tilde{\phi})(x_1)\), \( \tilde{U} = (\tilde{u}, 0, \ldots, 0) \in \mathbb{R}^N \) is a solution to (1) independent of the time variable \( t \) and of the tangential coordinates \( x' \) and it satisfies
\[
\begin{align*}
(\tilde{\rho} \tilde{u})_{x_1} &= 0, \quad (7a) \\
(\tilde{\rho} \tilde{u}^2 + K \tilde{\rho})_{x_1} + \tilde{\rho} \tilde{\phi}_{x_1} &= 0, \quad (7b) \\
-\tilde{\phi}_{x_1} &= \tilde{\rho} - e^\tilde{\phi} \quad (7c)
\end{align*}
\]
under conditions (2)–(6), that is,
\[
\inf_{x_1 \in \mathbb{R}^+} \tilde{\rho}(x_1) > 0, \quad \lim_{x_1 \to \infty} (\tilde{\rho}, \tilde{u}, \tilde{\phi})(x_1) = (\rho_+, u_+, 0), \quad \tilde{\phi}(0) = \phi_0. \quad (8)
\]
In the discussion of the existence of the stationary solution, the Sagdeev potential
\[
V(\phi) := \int_0^\phi \left[ h^{-1}(\xi) - e^\xi \right] d\xi, \quad h(\xi) := \frac{u_+^2}{2} \left( 1 - \zeta^{-2} \right) - K \log \zeta
\]
plays crucial roles (for graphs of \( V \), see Figure 1). Here the inverse function \( h^{-1} \) is defined by adopting the branch which contains \( \zeta = 1 \) (see [13] for details). The unique existence of the monotone stationary solution is obtained in [13].

**Theorem 2.1.** ([11, 13]) (i) Let \( u_+ \) be a constant satisfying either \( u_+^2 \leq K \) or \( u_+^2 \geq K + 1 \). Then the stationary problem (7) and (8) has a unique monotone solution \((\tilde{\rho}, \tilde{u}, \tilde{\phi})(x_1)\) verifying
\[
\rho, \tilde{u}, \tilde{\phi} \in C(\mathbb{R}^+) \quad \text{and} \quad \rho, \tilde{u}, \tilde{\phi}, \tilde{\phi}_{x_1} \in C^1(\mathbb{R}^+)
\]
if and only if the boundary data \( \phi_0 \) satisfies conditions
\[
V(\phi_0) \geq 0 \quad \text{and} \quad \phi_0 \leq h(M_+)
\]
where \( M_+ := |u_+|/\sqrt{K} \). Moreover, in the case that \( u_+^2 \geq K + 1 \) and \( \phi_0 \neq h(M_+) \), the stationary solution belongs to \( C^\infty(\mathbb{R}^+) \) and verifies the supersonic condition
\[
\inf_{x_1 \in \mathbb{R}^+} (\tilde{u}^2 - K)(x_1) > 0.
\]
Let $u_+^2$ be a constant satisfying $K < u_+^2 < K + 1$. If $\phi_0 \neq 0$, then the stationary problem (7) and (8) does not admit any solutions in the function space $C^1(\mathbb{R}^+)$. If $\phi_0 = 0$, then a constant state $(\hat{\rho}, \hat{u}, \hat{\phi}) = (1, u_+, 0)$ is the unique solution.

By Theorem 2.1, we see that the condition

$$u_+^2 > K + 1, \quad u_+ < 0$$

(9)

together with $|\phi_0| \ll 1$ or

$$u_+^2 = K + 1, \quad u_+ < 0$$

(10)

together with $\phi_0 \leq 0$ gives a sufficient condition for the unique existence of the monotone stationary solution. The condition

$$u_+^2 \geq K + 1, \quad u_+ < 0$$

(11)

is called the Bohm criterion in [12]. Note that some textbooks as [7] drop the equality in (11) and its manner is adopted in [13]. In the present paper, considering the physical interest as seen in [1, 12], we call (11) the Bohm criterion and define the sheath by the monotone stationary solution to (1) under (11).

To study the asymptotic stability of the sheath, we introduce unknown functions $v := \log \rho$, $\hat{v} := \log \hat{\rho}$ and the perturbation

$$(\psi, \eta, \sigma)(t, x_1, x') := (v, u, \phi)(t, x_1, x') - (\hat{v}, \hat{U}, \hat{\phi})(x_1).$$

Subtracting (7) from (1), we have the governing equations for $(\psi, \eta, \sigma)$ as

$$\psi_t + u \cdot \nabla \psi + \text{div} \eta + \eta \hat{v} x_1 = 0,$$

$$\eta_t + u \cdot \nabla \eta + K \nabla \psi + \nabla \sigma + \eta \hat{U} x_1 = 0,$$

$$-\Delta \sigma = e^{\psi+\hat{\phi}} - e^\psi - e^{\sigma+\hat{\phi}} + e^{\hat{\phi}}.\quad (12c)$$

The initial and the boundary data to (12) are obtained from (3)–(5) and (8):

$$(\psi, \eta)(0, x) = (\psi_0, \eta_0)(x) := (\log \rho_0 - \log \hat{\rho}, u_0 - \hat{U}),$$

$$\sigma(t, 0, x') = 0 \quad \text{for an arbitrary } x' \in \mathbb{R}^{N-1}.\quad (14)$$
If the perturbations \((\psi, \eta, \sigma)\) are sufficiently small, all characteristics in \(x_1\)-direction of hyperbolic system (12a) and (12b) are negative owing to the supersonic condition \(\inf_{x_1 \in \mathbb{R}_+} (\ddot{u}^2 - K(x_1)) > 0\). Namely,
\[
\lambda_1 := \eta_1 + \ddot{u} - \sqrt{K} < 0, \quad \lambda_2 := \eta_1 + \ddot{u} + \sqrt{K} < 0,
\]
\[
\lambda_{i} = \eta_1 + \ddot{u} < 0 \quad \text{for } i = 3, \ldots, N + 1.
\]

Hence no boundary conditions for the hyperbolic system (12a) and (12b) are necessary for the well-posedness of the initial boundary value problem (12)–(14). Consequently the boundary condition (4) is necessary and sufficient.

The results are stated in Theorem 2.2 and Theorem 2.3. In Theorem 2.2, the Bohm criterion is validated by showing the unique existence of global-in-time solution to (1) and its stability under (9). In Theorem 2.3, we show similar results under the degenerate condition of (10). In both results, we obtain convergence rates subject to the spatial decay rates of the initial perturbation by using the weighted energy method. Here we introduce a function space
\[
X^i_j([0, T]) := \bigcap_{k=0}^i C^k([0, T]; H^{j+k}([\mathbb{R}^N_+])), \quad i, j \in \mathbb{Z}, \ i, j \geq 0,
\]
where \(H^1(\mathbb{R}^N_+)\) denotes the \(l\)-th order Sobolev space in the \(L^2\) sense.

**Theorem 2.2.** ([11]) For \(N = 1, 2, 3\), let \(m = [N/2] + 2\). Assume the condition (9).

(i) Suppose \(e^{\lambda x_1/2} \psi_0, e^{\lambda x_1/2} \eta_0 \in (H^m(\mathbb{R}^N_+))^{N+1}\) for a certain positive constant \(\lambda\). Then there exists a positive constant \(\delta\) such that if \(\beta \in (0, \lambda]\) and \(\beta + \|e^{\beta x_1/2} \psi_0 - e^{\beta x_1/2} \eta_0\|_{H^m} / \beta \leq \delta\) are satisfied, the initial boundary value problem (12)–(14) has a unique global-in-time solution \((\psi, \eta, \sigma)\) as \(e^{\beta x_1/2} \psi, e^{\beta x_1/2} \eta, e^{\beta x_1/2} \sigma \in X^0_m([0, \infty)) \times (X^0_m([0, \infty)))^N \times X^2_m([0, \infty))\). Moreover, the solution verifies the decay estimate
\[
\|(e^{\beta x_1/2} \psi, e^{\beta x_1/2} \eta)(t)\|^2_{H^m} + \|(e^{\beta x_1/2} \sigma(t))\|^2_{H^{m+2}} \leq C\|(e^{\beta x_1/2} \psi_0, e^{\beta x_1/2} \eta_0)\|^2_{H^m} e^{-\alpha t}
\]
for certain positive constants \(C\) and \(\alpha\) which are independent of the time variable \(t\).

(ii) Suppose \((1 + \beta x_1)^{\lambda/2} \psi_0, (1 + \beta x_1)^{\lambda/2} \eta_0 \in (H^m(\mathbb{R}^N_+))^{N+1}\) for certain constants \(\lambda\) and \(\beta\) satisfying \(\lambda \geq 2\) and \(\beta > 0\). Then for an arbitrary \(\alpha \in (0, \lambda]\) there exists a positive constant \(\delta\) such that if \(\beta + \|((1 + \beta x_1)^{\lambda/2} \psi_0, (1 + \beta x_1)^{\lambda/2} \eta_0)\|_{H^m} / \beta \leq \delta\) are satisfied, the initial boundary value problem (12)–(14) uniquely has a global-in-time solution \((\psi, \eta, \sigma)\) as \((1 + \beta x_1)^{\alpha/2} \psi, (1 + \beta x_1)^{\alpha/2} \eta, (1 + \beta x_1)^{\alpha/2} \sigma \in X^0_m([0, \infty)) \times (X^0_m([0, \infty)))^N \times X^2_m([0, \infty))\). Moreover, the solution verifies the decay estimate
\[
\|(1 + \beta x_1)^{\alpha/2} \psi(t), (1 + \beta x_1)^{\alpha/2} \eta(t)\|^2_{H^m} + \|(1 + \beta x_1)^{\alpha/2} \sigma(t)\|^2_{H^{m+2}} \leq C\|(1 + \beta x_1)^{\lambda/2} \psi_0, (1 + \beta x_1)^{\lambda/2} \eta_0\|^2_{H^m} (1 + \beta t)^{-(\lambda - \alpha)},
\]
where \(C\) is a positive constant determined by \(\alpha\).

Theorem 2.2 is an extension of the theorem obtained in [13], in which the spatial dimension is limited only to one and more crucially, strictly stronger conditions than (9) is assumed.

**Theorem 2.3.** ([11]) For \(N = 1, 2, 3\), let \(m = [N/2] + 2\). Assume the condition (10). Set \(\lambda_0 = 5.5693\ldots\) be the unique real solution to the equation \(\lambda_0(\lambda_0 - 1)(\lambda_0 - 2) - 12(\lambda_0 + 2) = 0\). Suppose \((1 + \beta x_1)^{\lambda/2} \psi_0, (1 + \beta x_1)^{\lambda/2} \eta_0 \in (H^m(\mathbb{R}^N_+))^{N+1}\).
for certain constants $\lambda$ and $\beta$ satisfying $\lambda \in [4, \lambda_0)$ and $\beta > 0$. Then for arbitrary $\alpha$, $\theta$ satisfying $\alpha \in (0, \lambda)$, $\theta \in (0, 1]$, there exists a positive constant $\delta$ such that if $\phi_0 \in (-\delta, 0)$, $\beta/((K + 1)|\phi_0|/6)^{1/2} \in [\theta, 1]$ and $||(1 + \beta x_1)^{\lambda/2}\psi_0, (1 + \beta x_1)^{\lambda/2}\eta_0)||_{H^\infty} / \beta^2 \leq \delta$ are satisfied, the initial boundary value problem (12)–(14) has a unique global-in-time solution $(\psi, \eta, \sigma)$ as $((1 + \beta x_1)^{\lambda/2} \psi, (1 + \beta x_1)^{\lambda/2} \eta, (1 + \beta x_1)^{\alpha/2} \sigma) \in C^0_0([0, \infty)) \times (C^0_0([0, \infty)))^N \times C^0_0([0, \infty))$. Moreover, the solution verifies the decay estimate
\[
||(1 + \beta x_1)^{\alpha/2}\psi, (1 + \beta x_1)^{\alpha/2}\eta)(t)||_{H^m}^2 + ||(1 + \beta x_1)^{\alpha/2}\sigma(t)||_{H^{m+2}}^2 \leq C||(1 + \beta x_1)^{\lambda/2}\psi_0, (1 + \beta x_1)^{\lambda/2}\eta_0)||_{H^m}^2 (1 + \beta t)^{-(\lambda - \alpha)/3},
\]
where $C$ is a positive constant determined by $\alpha$ and $\theta$.

REFERENCES


E-mail address: shinya@is.titech.ac.jp
E-mail address: ohnawa@aoni.waseda.jp
E-mail address: masahiro@is.titech.ac.jp
A WELL-BALANCED DG SCHEME WITH UNCONDITIONALLY POSITIVE IMPLICIT TIME INTEGRATION

SIGRUN ORTLEB AND ANDREAS MEISTER
Fachbereich Mathematik und Naturwissenschaften
University of Kassel, Heinrich Plett Str. 40
34132 Kassel, Germany

Abstract. This paper presents a new well-balanced and positivity preserving DG scheme with modal filtering on unstructured triangular grids using implicit time integration. The novel approach is based on the so-called Patankar trick and guarantees non-negativity of the water height for any time step size while still preserving conservativity. Due to this modification, the implicit scheme can now take full advantage of larger time steps and is therefore able to beat explicit time stepping in terms of CPU time.

1. Introduction. The shallow water equations (SWE) represent an important model in many scientific and engineering applications. They can be used to provide realistic simulations of flows in rivers, lakes or coastal areas, where the incorporation of arbitrary non-flat bottom topography is absolutely necessary. If the bottom topography is assumed to be constant with respect to time, the SWE are given by

\[
\begin{align*}
\partial_t \varphi + \partial_{x_1} (\varphi v_1) + \partial_{x_2} (\varphi v_2) &= 0, \\
\partial_t (\varphi v_1) + \partial_{x_1} \left( \varphi v_1^2 + \frac{1}{2} \varphi^2 \right) + \partial_{x_2} (\varphi v_1 v_2) &= -g \varphi \partial_{x_1} b, \\
\partial_t (\varphi v_2) + \partial_{x_1} (\varphi v_1 v_2) + \partial_{x_2} \left( \varphi v_2^2 + \frac{1}{2} \varphi^2 \right) &= -g \varphi \partial_{x_2} b,
\end{align*}
\]

where \( b \) denotes the bottom elevation, the geopotential \( \varphi = gH \) is composed of the water height \( H \) above the bottom and the gravitational constant \( g = 9.812 \) and \( v = (v_1, v_2)^T \) denotes the velocity vector. In this contribution, we solve this system of balance laws with the discontinuous Galerkin (DG) method [5, 10]. In order to introduce a small but sufficient amount of numerical dissipation to the DG scheme in case of non-smooth solutions, we use a novel damping strategy consisting in modal filters. These filters are based on the formulation of spectral viscosity due to Tadmor [17] and are described in [12, 13]. An important challenge is posed by steady states with non-zero flux gradients that are exactly balanced by the non-zero source term. In addition, the DG scheme has to guarantee non-negativity of the water height \( H \). Hence a positivity preserving, well-balanced numerical scheme is required. In this contribution, we apply the positivity preserving (PP) and well-balanced (WB) method of Xing, Zhang and Shu [19] to unstructured triangular

2000 Mathematics Subject Classification. Primary: 65M60, 35L65.
Key words and phrases. DG method, shallow water equations, implicit time integration, well-balancedness, unconditional positivity, Patankar schemes.

The first author is supported by DFG grant ME 1889/3-1.

823
2. A Positivity-Preserving, Well-Balanced DG Scheme with Adaptive Filtering. We rewrite the shallow water equations (1) in the more compact form

$$\frac{\partial}{\partial t} u(x, t) + \nabla \cdot F(u(x, t)) = s(u(x, t), x),$$

(2)

for $(x, t) \in \Omega \times \mathbb{R}^+$, where $\Omega \subset \mathbb{R}^2$ is an open polygonal domain and the conservative variables are now collected in $u = (\varphi, \varphi v_1, \varphi v_2)^T$, while $F$ contains the fluxes and $s$ the sources due to bottom topography. Let now $T^h$ be a conforming triangulation of $\Omega$ and let $W^h$ be the piecewise polynomial space defined by $W^h = \{ w_h \in L^\infty(\Omega) \mid w_h|_{\tau_i} \in P^N(\tau_i) \quad \forall \tau_i \in T^h \}$, where $P^N(\tau_i)$ denotes the space of all polynomials on $\tau_i$ of degree $\leq N$. The well-balanced DG-approximation $u_h : \Omega \times \mathbb{R}^+ \rightarrow \mathbb{R}^3$, $u_h(\cdot, t) \in (W^h)^3$ is now constructed according to the approach of Xing, Zhang and Shu in [19], i.e. we solve

$$\frac{d}{dt} \int_{\tau_i} u_h \cdot w \, dx = \int_{\partial \tau_i} F(u_h) \cdot \nabla w \, dx - \int_{\partial \tau_i} F^B(u^+, u^-, \varphi_i^+, \varphi_i^-) \cdot w \, d\sigma$$

$$+ \int_{\tau_i} s_h(u_h, x) \cdot w \, dx,$$

(3)

for any $\tau_i \in T^h$, $w \in (W^h)^3$, where $F^B$ is the well-balanced correction of a suitable numerical flux function to be specified below and $u_i^-$, $u_i^+$ denote the approximate solution within $\tau_i$ and an adjacent element, respectively. Furthermore, the source term is discretized by $s_h(u_h, x) = -g \cdot (0, \varphi_h \cdot \partial_{x_1} b_h, \varphi_h \cdot \partial_{x_2} b_h)^T$, where $b_h$ is the projection of the bottom $b$ to $W^h$. Given a numerical flux $F^{num}$, the well-balanced flux $F^B$ is

$$F^B(u^-, u^+, \varphi_i^+, \varphi_i^-) = F^{num}(u^-, u^+, \varphi_i^+, \varphi_i^-) + \left( \begin{array}{c}
\frac{n_1}{2} ((\varphi_i^-)^2 - (\varphi_i^*)^2) \\
\frac{n_2}{2} ((\varphi_i^-)^2 - (\varphi_i^*)^2)
\end{array} \right),$$

where the modified (starred) left and right states are obtained from a hydrostatic reconstruction according to [1], i.e.

$$u^\pm_{i,*} = \left( (\varphi^\pm_{i,*}, \varphi^\mp_{i,*} \cdot (v_1)^\pm_{i,*}, \varphi^\pm_{i,*} \cdot (v_2)^\pm_{i,*})^T \right),$$

$$\varphi^\pm_{i,*} = \max \left\{ 0, \varphi_i^+ + g \left( b_i^+ - \max \left\{ b_i^-, b_i^+ \right\} \right) \right\}.$$
More precisely, in our computations, we chose the HLL flux \([8]\) for \(F^{num}\). We shortly remark that although we obtain a local truncation error of order \(N+1\) if the integrals in \((3)\) are approximated by quadrature formulae of order \(2N\) on elements and of order \(2N+1\) on edges, the well-balanced property in general requires a higher degree of exactness. More precisely, in the case of no discharge, \(\varphi v\) and a constant water surface \(b_h + H_h = const\), all integrals in \((3)\) have to be evaluated exactly, see \([18]\), where the well-balancedness of this approach is proven. The resulting system of ordinary differential equations, denoted by \(\frac{d}{dt} U = L_h(U, t)\), may then be solved by appropriate time integration schemes that are compatible with the property of positivity preservation. In case of explicit time integration, to achieve non-negativity of the water height, we use the PP limiter described in \([19]\). This technique is based on quadrature rules on the triangle \(\tau_i\) with positive weights that exactly integrate polynomials of degree \(N\) and whose set of nodes \(X^N\) also contains the points that are used for the numerical integration over the element boundary \(\partial \tau_i\) within the DG scheme. Quadrature rules with these properties were constructed in \([20]\). We now assume that \((3)\) is solved by the forward Euler scheme and denote by \(u^i_n(x)\) and \(u^{i+1}_n(x)\) the approximate solution at the time levels \(t^n\) and \(t^{n+1} = t^n + \Delta t\), respectively. Furthermore, let the vector of cell averages at time \(t^n\) be denoted by \(\bar{u}^n_i = \frac{1}{|\tau_i|} \int_{\tau_i} u^n_i(x) dx\). The result by Xing, Zhang and Shu then asserts that the cell average \(\bar{H}^{n+1}_i\) of the water height at time \(t^{n+1}\) is non-negative under the premises that a suitable CFL-type restriction on \(\Delta t\) is fulfilled and that \(H^n_i(x)\) is non-negative at each quadrature point in \(X^N_i\). To enforce this last condition, Xing and Shu use a simple scaling around the cell average, i.e. \(u^n_i := u^n_i|_{\tau_i}\) is modified to \(\tilde{u}^n_i\) by \(\tilde{u}^n_i(x) = u^n_i + \theta (u^n_i(x) - \bar{u}^n_i)\), where \(\theta = \min \left\{1, \frac{H^n_i - \min_{x \in \chi^N_i} H^n_i(x)}{H^n_i}\right\}\). A direct computation of the velocities as \(v_i = (H \nu) / H_i\) will eventually lead to numerical instabilities for small water height even for bounded discharges. Similar to the approaches of Bryson et al. \([2]\) and Xing et al. \([19]\), we therefore compute damped velocities within the DG scheme whenever pointwise values are required. More precisely, we set \(v_i = 0\) if \(H_i < \epsilon\) and \(v_i = \frac{2H_i(H \nu)}{H^2 + \max(H_i, \epsilon)}\) otherwise, where \(\epsilon = 10^{-6}\) was chosen in computations. In addition, positivity preservation with respect to the forward Euler time integration can be extended to higher order SSP schemes, see \([20]\). In our computations, we used the third order Shu-Osher TVD Runge-Kutta time discretization \([16]\) as the representative explicit scheme. The representative implicit time integration method is obtained from an SDIRK3 scheme which is modified via the so-called Patankar trick as described in Section 4.

3. Modal Filtering. To represent the approximate solution on each triangle, we use the well-known Proriol-Koornwinder-Dubiner (PKD) basis \([10, 6]\) given by

\[
\Phi_{lm}(r,s) = P_{l}^{0,0} \left( \frac{1 + r - 1}{s - 1} \right) \left( \frac{1 - s}{2} \right)^{l} P_{m}^{2l+1,0}(s), \quad l, m \in \mathbb{N}_0, \quad 0 \leq l + m \leq N
\]

on the reference element \(\mathcal{T} = \{(r,s) \in \mathbb{R}^2 \mid -1 \leq r, s, r + s \leq 0\}\), where \(P_{n}^{\alpha, \beta}\) denotes the one-dimensional Jacobi polynomial of degree \(n\) associated to the weight function \(w(x) = (1-x)^{\alpha}(1+x)^{\beta}\). Precisely the corresponding PKD coefficients on each triangle, denoted by \(\tilde{u}_{i,lm}\), are evolved in time and these quantities can easily be accessed. Therefore, additional numerical dissipation can be efficiently implemented if it is directly based on these coefficients. In our shock capturing
strategy, the basic DG scheme described above is hence supplemented by a modal filter. In [13], dealing with the solution of the Euler equations of gas dynamics, this filter is obtained componentwise for each conservative variable by an extension of the spectral viscosity method to multidomain PKD expansions. Applying operator-splitting to the viscous formulation then results in a modification of the vector of PKD coefficients by an exponential filter of the form

\[
\tilde{u}_{i,lm}^{\text{mod}} = \exp \left(-\alpha_i s_i \eta^{2p}\right) \tilde{u}_{i,lm}, \quad \eta = \frac{l + m}{N + 1},
\]

with shock indicator \( s_i \). This filter only depends on the filter order \( 2p \) and the filter strength \( \alpha_i = C_p \frac{N \Delta t}{h_i} \), where \( \Delta t \) and \( h_i \) denote the time step size and the shortest height of \( \tau_i \), respectively. The shock indicator \( s_i \) in (4) is based on the decay rate of PKD coefficients, similar to the definition in [15]. In order not to interfere with well-balancedness and to introduce more stability in nearly dry regions, we set

\[
s_i = (\tilde{H}_i^*)^{-1} \min \left\{ 1, \frac{1000(5N^4 + 1)}{\sum_{l+m=N} (\tilde{\mu}_{i,lm})^2} \left( \sum_{l+m<N} (\tilde{\mu}_{i,lm})^2 + \tilde{\epsilon} \right)^{-1} \right\},
\]

with a small regularization parameter \( \tilde{\epsilon} = 10^{-10} \), modified cell means of water height \( \tilde{H}_i^* = \max\{\tilde{H}_i^n, \tilde{\epsilon}\} \) and well-balanced indicators \( \tilde{\mu}_{i,lm} = \tilde{\varphi}_{i,lm} + \tilde{g}_{i,lm} \).

4. The Patankar Trick for Unconditional Positivity. Neglecting boundary terms, in the DG scheme for cell means of water height, \( \tilde{H}_i(t) = \frac{1}{\tau_i} \int_{\Gamma_i} \tilde{H}_{h}(x, t) \, dx \), we now distinguish between positive and negative flux contributions over element boundaries. Collecting the indices of neighbor elements in the set \( N(\tau_i) \), we have

\[
\frac{d}{dt} \left[ |\tau_i| \tilde{H}_i \right] = - \sum_{j \in N(\tau_i)} \int_{\Gamma_{ij}} F_{ij}^{WB}(\mathbf{u}_-, \mathbf{u}_+, \mathbf{n}) \, d\sigma = \sum_{j \in N(\tau_i)} p_{ij} - \sum_{j \in N(\tau_i)} d_{ij},
\]

where the properties \( p_{ij} - d_{ij} = -\int_{\Gamma_{ij}} F_{ij}^{WB}(\mathbf{u}_-, \mathbf{u}_+, \mathbf{n}) \, d\sigma \) and \( p_{ij} = d_{ij} \) are guaranteed by choosing the production term \( p_{ij} = \max \left\{ 0, -\int_{\Gamma_{ij}} F_{ij}^{WB}(\mathbf{u}_-, \mathbf{u}_+, \mathbf{n}) \, d\sigma \right\} \) and the destruction term \( d_{ij} = \max \left\{ 0, \int_{\Gamma_{ij}} F_{ij}^{WB}(\mathbf{u}_-, \mathbf{u}_+, \mathbf{n}) \, d\sigma \right\} \).

In the similar context of production-destruction equations Burchard, Deleersnijder and Meister [3] developed the so-called modified Patankar-Euler (MPE) scheme based on the non-conservative Patankar scheme [14], which is of first order in time. For the cell means of water height in the DG SWE code it has the form

\[
|\tau_i| \tilde{H}_i^{n+1} = |\tau_i| \tilde{H}_i^n + \Delta t \left( \sum_{i=1}^{l} p_{ij}^{n} \tilde{H}_j^{n+1} - \sum_{i=1}^{l} d_{ij}^{n} \tilde{H}_j^{n+1} \right),
\]

In [3], a second order modified Patankar scheme was constructed as well. Both modified Patankar schemes are positivity preserving and conservative for any time step size, see [3]. Our new modified Patankar SDIRK3 (MPSDIRK3) scheme is now a modification of the method of Cash [4] which is given by

\[
\begin{align*}
U^{(1)} &= U^n + \gamma \Delta t L_h \left( U^{(1)} \right) \\
U^{(2)} &= U^n + \delta \Delta t L_h \left( U^{(1)} \right) + \gamma \Delta t L_h \left( U^{(2)} \right) \\
U^{n+1} &= U^n + \alpha \Delta t L_h \left( U^{(1)} \right) + \beta \Delta t L_h \left( U^{(2)} \right) + \gamma \Delta t L_h \left( U^{n+1} \right)
\end{align*}
\]
with $\alpha = 1.2084966491760101, \beta = -0.6443631706844691, \gamma = 0.4358665215084580$ and $\delta = 0.2820667392457705$. As the implicit Euler scheme is unconditionally positive, see [9], we can safely apply the first stage in combination with the PP limiter by Xing and Shu. As $\delta < \gamma$, also the second stage yields non-negative water height. Only in the last stage, we need to modify the vector

$$\mathbf{S} = \mathbf{U}^n + \alpha \Delta t L_h \left( \mathbf{U}^{(1)} \right) + \beta \Delta t L_h \left( \mathbf{U}^{(2)} \right) = \begin{bmatrix} g \bar{H}^S_{i,lm} \\ g(\mathbf{Hv})_{i,lm} \end{bmatrix}$$

by a vector $\tilde{\mathbf{S}} = \begin{bmatrix} g \bar{H}^S_{i,lm} \\ g(\mathbf{Hv})_{i,lm} \end{bmatrix}$ with non-negative water-heigh. For this purpose, in analogy to the modified Patankar schemes in [3], we modify the cell means $\bar{H}^S_i = \bar{H}^S_{i,00}$ by

$$\bar{H}^S_i = \bar{H}^n_i + \frac{\Delta t}{|\tau_i|} \sum_{j \in N(\tau_i)} \left( p_{ij} \frac{\bar{H}^S_j}{\bar{H}^n_j} - d_{ij} \bar{H}^S_i \right),$$

where $p_{ij} = \max \left\{ 0, -\alpha \int_{\Gamma_{ij}} F^W_B \left( \mathbf{U}^{(1)}, \mathbf{n} \right) d\sigma - \beta \int_{\Gamma_{ij}} F^W_E \left( \mathbf{U}^{(2)}, \mathbf{n} \right) d\sigma \right\} = d_{ij}$ and $\bar{H}^* = \bar{H}^n$ if $\bar{H}^* > \epsilon$ whereas otherwise $\bar{H}^* = \bar{H}^n$. By setting $\bar{H}^* = \bar{H}^S$ for moderate values of water height, we intend to reduce the influence of the Patankar modification in wet areas away from wet-dry fronts. In analogy to the computation of damped velocities in case of a very small water height, the production and destruction terms are again modified to

$$p_{ij} = \frac{d_{ij}}{\bar{H}^n_j} = \begin{cases} 0, & \text{if } \bar{H}^* < \epsilon, \\ 2 \bar{H}^* \cdot p_{ij} / \left( (\bar{H}^*)^2 + \max\{ (\bar{H}^*)^2, \epsilon \} \right), & \text{otherwise}. \end{cases}$$

The higher order coefficients in the DG scheme are now modified as follows. We set $\bar{H}^S_{i,lm} = (Hv_1)_{i,lm} = (Hv_2)_{i,lm} = 0$ if $\bar{H}^S_i < 0$, otherwise the higher order coefficients remain unchanged. We remark that the cell means of the discharges will not be modified by this procedure.

5. **Numerical Experiments for Shallow Water Flows.** We will now present numerical results for two classical test cases demonstrating the behaviour of our well-balanced and positivity preserving DG scheme with modal filtering for a polynomial degree of $N = 2$. The nonlinear systems of equations arising due to the implicit time integration were solved using a Jacobian-free Newton-GMRES scheme. In order prevent the error in time to dominate the error in space, the CFL number for implicit time integration was set to $\frac{\Delta t}{\max\{|v|, |\sqrt{g\bar{h}}|\}} = 1$, while the time step choice for the explicit scheme was based on the positivity requirement.

Figure 1 depicts the DG solution for a small perturbation of a steady state originally considered by LeVeque, see [11], on a computational grid of $K = 46360$ elements. Filter parameters where set to $p = 1$ and $C_p = 10$. Here, the main purpose is to show that our filtering technique produces correct results and a very detailed resolution of the small perturbation. The basic features of the results agree very well with those presented in [18]. Furthermore, a better resolution of the wave structures due to less dissipative shock capturing can be perceived. A clear gain thanks to the higher order scheme is also visible in comparison with the second-order computations on a much finer grid in [2]. Figure 2 shows the
DG solution for the oscillating-lake test proposed in [7] on a computational grid consisting of $K = 23138$ elements. Filter parameters where set to $p = 1$ and $C_p = 10$. The results are in very good agreement with those presented in [19]. Hence, one can conclude that the wetting and drying treatment suggested in that work may also be combined with shock capturing by modal filtering and implicit time integration. In Table 1, we compare the CPU times of our MPSDIRK3 scheme to those obtained by the TVD-RK3 scheme of Shu and Osher for the oscillating lake test on increasingly stiff computational grids. The stiffness of the grids is increased via local refinement as depicted in Figure 3. We use the stiffness measure $S = \frac{\max_{\tau \in \Delta h} |\tau|}{\min_{\tau \in \Delta h} |\tau|}$. According to the results, the implicit scheme beats the explicit one by a factor up to 1.5. Table 1 furthermore lists the mass conservation errors committed by the implicit scheme. Full conservation can obviously only be achieved if the accuracy within the iterative solver is set to zero, which is neglected due to practical reasons as usual. However, the results show that for the specific tolerances chosen in this study, the corresponding conservation error can be neglected. So far, no preconditioner was used within the Newton-GMRES solver. Hence, we expect an additional speed-up when preconditioning strategies are used. The specific design of these preconditioners is left to future work.

**REFERENCES**


AN IMPLICIT WB AND PP DG SCHEME FOR SWE

Figure 2. Oscillating lake. Water surface $w = H + b$. Output times: top left at $T = T_{\text{per}}/6$, top right at $T = T_{\text{per}}/3$, bottom left at $T = T_{\text{per}}/2$, bottom right at $T = T_{\text{per}} = 2\pi/\sqrt{0.2g}$.

Figure 3. Stiffest computational grid with $S = 1654.6$.

<table>
<thead>
<tr>
<th>Stiffness $S$</th>
<th>Avg. $\Delta t_{EX}$</th>
<th>Avg. $\Delta t_{IM}$</th>
<th>$\frac{CPU_{EX}}{CPU_{IM}}$</th>
<th>$err_{cons}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>2.99e-4</td>
<td>1.07e-2</td>
<td>0.65</td>
<td>2.31e-14</td>
</tr>
<tr>
<td>25.9</td>
<td>1.51e-4</td>
<td>5.42e-3</td>
<td>0.82</td>
<td>1.11e-14</td>
</tr>
<tr>
<td>103.4</td>
<td>7.55e-5</td>
<td>2.71e-3</td>
<td>1.29</td>
<td>8.88e-15</td>
</tr>
<tr>
<td>413.7</td>
<td>3.77e-5</td>
<td>1.36e-3</td>
<td>1.52</td>
<td>6.93e-14</td>
</tr>
<tr>
<td>1654.6</td>
<td>1.89e-5</td>
<td>6.79e-4</td>
<td>1.34</td>
<td>2.25e-13</td>
</tr>
</tbody>
</table>

Table 1. CPU time comparison and conservation error of implicit scheme.


E-mail address: ortleb@mathematik.uni-kassel.de
E-mail address: meister@mathematik.uni-kassel.de
ON DECAY OF PERIODIC RENORMALIZED SOLUTIONS TO
SCALAR CONSERVATION LAWS

Evgeniy Yu. Panov
41, B. Sankt-Peterburgskaya St
173003 Veliky Novgorod, Russia

Abstract. We establish a necessary and sufficient condition for decay of period-
ic renormalized solutions to a multidimensional conservation law with merely
continuous flux vector.

1. Introduction. In the half-space $\Pi = \mathbb{R}_+ \times \mathbb{R}^n$, $\mathbb{R}_+ = (0, +\infty)$, we consider the
Cauchy problem for a first order multidimensional conservation law
\begin{equation}
    u_t + \text{div}_x \varphi(u) = 0 \quad (1)
\end{equation}
with initial data
\begin{equation}
    u(0, x) = u_0(x). \quad (2)
\end{equation}
The flux vector $\varphi(u)$ is supposed to be only continuous:
\begin{equation}
    \varphi(u) = (\varphi_1(u), \ldots, \varphi_n(u)) \in C(\mathbb{R}, \mathbb{R}^n).
\end{equation}
We assume that initial function $u_0(x)$ is periodic, that is, $u_0(x + e_i) = u_0(x)$ for
almost all $x \in \mathbb{R}^n$ and all $i = 1, \ldots, n$, where $\{e_i\}_{i=1}^n$ is a basis of periods in $\mathbb{R}^n$.
Denote by $P$ the corresponding fundamental parallelepiped
\begin{equation}
    P = \{ x = \sum_{i=1}^n \alpha_i e_i \mid \alpha_i \in [0, 1), \ i = 1, \ldots, n \}.
\end{equation}

If $u_0(x) \in L^\infty(\mathbb{R}^n)$ then the notion of entropy solution of (1), (2) in the sense of
S.N. Kruzhkov [5] is well-defined.

Definition 1.1. A bounded measurable function $u = u(t, x) \in L^\infty(\Pi)$ is called an
entropy solution (e.s. for short) of (1), (2) if for all $k \in \mathbb{R}$
\begin{equation}
    |u - k|_t + \text{div}_x [\text{sign}(u - k)(\varphi(u) - \varphi(k))] \leq 0 \quad (3)
\end{equation}
in the sense of distributions on $\Pi$ (in $\mathcal{D}'(\Pi)$);
\begin{equation}
    \text{ess lim}_{t \to 0} u(t, \cdot) = u_0 \quad \text{in } L^1_{\text{loc}}(\mathbb{R}^n).
\end{equation}

Condition (3) means that for all non-negative test functions $f = f(t, x) \in C^1_0(\Pi)$
\begin{equation}
    \int_{\Pi} [|u - k| f_t + \text{sign}(u - k)(\varphi(u) - \varphi(k)) \cdot \nabla_x f] dtdx \geq 0
\end{equation}
(here $\cdot$ denotes the inner product in $\mathbb{R}^n$).

2000 Mathematics Subject Classification. Primary: 35L65, 35B10; Secondary: 35B35.
Key words and phrases. Periodic renormalized solutions, decay property, $H$-measures.
Supported by Russian Foundation for Basic Research, grant no. 12-01-00230-a.
In the case under consideration when the flux functions are merely continuous, the effect of infinite speed of propagation for initial perturbations appears, which leads even to the nonuniqueness of e.s. to problem (1), (2) if \( n > 1 \) (see examples in [6, 7]). But, if initial function is periodic (at least in \( n - 1 \) independent directions), the uniqueness holds: an e.s. of (1), (2) is unique and space-periodic, see the proof in [14, 15].

In the present paper we assume that the initial function \( u_0 \in L^1(P) \) and may be unbounded. In this general situation even the natural requirement \( \varphi(u) \in L^1_{loc}(\Pi_T, \mathbb{R}^n) \) turns out to be too restrictive. However, if we reject these assumption, we cannot consider entropy conditions (and even the equation itself) within the framework of the theory of distributions. To define such solutions \( u = u(t, x) \) (called renormalized), one uses entropy conditions for superpositions \( s(u) \), where \( s \) are bounded functions of special form (cut-off functions). Renormalized entropy solutions to the problem (1), (2) with summable initial data were first introduced in [1], where the existence and uniqueness of such solutions were also established. The results of [1] were generalized in [8] to the case of arbitrary measurable initial data. The notion of renormalized entropy solution was later modified in [9] for the periodic case.

Recall the corresponding definition. We denote by \( s_{a,b}(u) = \max(a, \min(b, u)) \) the cut-off function at levels \( a \) and \( b \), where \( a, b \in \mathbb{R}, a \leq b \).

**Definition 1.2.** A \( x \)-periodic measurable function \( u = u(t, x) \) is called a renormalized entropy solution (r.e.s. for short) of (1), (2) if for all \( a, b \in \mathbb{R}, a \leq b \)

\[
(s_{a,b}(u))_t + \text{div}_x(\varphi(s_{a,b}(u))) = \mu_b - \mu_a \quad \text{in } D'(\Pi),
\]

where \( \mu_p, p \in \mathbb{R}, \) is a family of \( x \)-periodic nonnegative locally finite measures on \( \Pi (\mu_p \in M_{loc}(\Pi), \mu_p \geq 0) \) such that \( \lim_{p \to \infty} \mu_p((0,T) \times P) = 0 \) for all \( T > 0 \), and

\[
\text{ess lim}_{t \to 0} |s_{a,b}(u(t, \cdot)) - s_{a,b}(u_0)| = 0 \quad \text{in } L^1(P).
\]

In the case of bounded \( u, u_0 \) the notions of r.e.s. and e.s. coincide. Moreover, in this case the defect measures \( \mu_p \) satisfy the representation

\[
\mu_p = -\frac{1}{2} \{ |u - p|_t + \text{div}_x[\text{sign}(u - p)(\varphi(u) - \varphi(p))'] \}.
\]

As was shown in [9], for each \( u_0 \in L^1(P) \) there exists a unique r.e.s. \( u = u(t, x) \) of problem (1), (2). Moreover, the following contraction property holds in \( L^1(P) \) (see [9, Corollary 3.3]):

**Proposition 1.** Let \( u(t, x) \) and \( v(t, x) \) be r.e.s. to the problem (1), (2) with the initial data \( u_0(x) \) and \( v_0(x) \) (which are supposed to be merely measurable functions), respectively. Then for almost all \( t > 0 \)

\[
\int_P |u(t, x) - v(t, x)|dx \leq \int_P |u_0(x) - v_0(x)|dx.
\]

As was established by G.-Q. Chen and H. Frid [2], under the conditions \( \varphi(u) \in C^2(\mathbb{R}, \mathbb{R}^n) \) and

\[
\forall (\tau, \xi) \in \mathbb{R}^{n+1}, (\tau, \xi) \neq 0, \quad \text{meas} \{ u \in \mathbb{R} | \tau + \varphi'(u) \cdot \xi = 0 \} = 0,
\]

\[\text{for all } (\tau, \xi) \in \mathbb{R}^{n+1}, (\tau, \xi) \neq 0, \quad \text{meas} \{ u \in \mathbb{R} | \tau + \varphi'(u) \cdot \xi = 0 \} = 0,
\]

\[
\mu_p = -\frac{1}{2} \{ |u - p|_t + \text{div}_x[\text{sign}(u - p)(\varphi(u) - \varphi(p))'] \}.
\]

As was shown in [9], for each \( u_0 \in L^1(P) \) there exists a unique r.e.s. \( u = u(t, x) \) of problem (1), (2). Moreover, the following contraction property holds in \( L^1(P) \) (see [9, Corollary 3.3]):

**Proposition 1.** Let \( u(t, x) \) and \( v(t, x) \) be r.e.s. to the problem (1), (2) with the initial data \( u_0(x) \) and \( v_0(x) \) (which are supposed to be merely measurable functions), respectively. Then for almost all \( t > 0 \)

\[
\int_P |u(t, x) - v(t, x)|dx \leq \int_P |u_0(x) - v_0(x)|dx.
\]

As was established by G.-Q. Chen and H. Frid [2], under the conditions \( \varphi(u) \in C^2(\mathbb{R}, \mathbb{R}^n) \) and

\[
\forall (\tau, \xi) \in \mathbb{R}^{n+1}, (\tau, \xi) \neq 0, \quad \text{meas} \{ u \in \mathbb{R} | \tau + \varphi'(u) \cdot \xi = 0 \} = 0,
\]
the following decay property holds for bounded space-periodic entropy solutions $u(t,x)$ of (1), (2):

$$\text{ess lim } u(t,\cdot) = \text{const} = \frac{1}{|P|} \int_{P} u_0(x)dx \quad \text{in } L^1(P).$$

(6)

Here $|P|$ denotes the Lebesgue measure of $P$.

In the present paper we generalize this result to the case of renormalized entropy solutions of (1), (2) and propose the following necessary and sufficient condition for the decay property

$$\forall \xi \in L', \xi \neq 0, \text{ the function } u \to \varphi(u) \cdot \xi \text{ is not affine on non-empty intervals},$$

(7)

where $L' = \{ x \in \mathbb{R}^n \mid x \cdot e_i \in \mathbb{Z} \ \forall i = 1, \ldots, n \}$ is the dual lattice to the lattice of periods $L = \{ x = \sum_{i=1}^{n} k_i e_i \mid k_i \in \mathbb{Z}, \ i = 1, \ldots, n \}$, $\mathbb{Z}$ being the set of integers.

Thus, our main result is the following theorem.

**Theorem 1.3.** Every r.e.s. of equation (1) satisfies the decay property (6) if and only if condition (7) holds.

In the case when the basis of periods is not fixed and may depend on a solution, the statement of Theorem 1.3 remains valid after replacement of condition (7) by the following stronger one:

$$\forall \xi \in \mathbb{R}^n, \xi \neq 0, \text{ the function } u \to \varphi(u) \cdot \xi \text{ is not affine on non-empty intervals}. \quad (8)$$

Obviously, condition (8) is strictly weaker than (5) even in the case of smooth flux $\varphi(u)$.

2. **Preliminaries.** Using methods developed in papers [16, 17], we can claim that, after a possible correction on a set of null Lebesgue measure, a r.e.s. of equation (1) is a continuous map $t \to u(t,\cdot) \in L^1(P)$, that is, $u \in C([0, +\infty), L^1(P))$. We will assume below that this property is always satisfied.

Let $u^+ = \max(u,0)$, $u^- = \max(-u,0)$. We will need some simple properties of r.e.s.

**Lemma 2.1.** If $u(t,x)$ is a r.e.s. of (1), (2) with initial function $u_0 \in L^1(P)$ then for every $t > 0$ $u(t,\cdot) \in L^1(P)$ and

$$\forall R \geq 0 \int_P (|u(t,x)| - R)^+ dx \leq \int_P (|u_0(x)| - R)^+ dx,$$

$$\int_P u(t,x)dx = \int_P u_0(x)dx.$$  

(9)

Observe that the first inequality in (9) readily follows from [9, Theorem 3.2].

**Lemma 2.2.** Let $u = u(t,x)$ be a r.e.s. of (1), (2), and $\mu_p \in M_{\text{loc}}(\Pi)$, $p \in \mathbb{R}$ be the family of defect measures from condition (4). Then

$$\mu_p(\mathbb{R}^+ \times P) \leq \int_P (|u_0(x)| - |p|)^+ dx.$$

If $u(t,x)$ is a r.e.s. of (1), (2) then for each $h \in \mathbb{R}^n$ the function $u(t,x+h)$ is a r.e.s. of (1), (2) with initial function $u_0(x+h)$. By Proposition 1 for all $t > 0 \int_P |u(t,x+h) - u(t,x)|dx \leq \int_P |u_0(x+h) - u_0(x)|dx$, which implies that the family of functions $u(t,\cdot), t > 0$, is precompact in $L^1(P)$. It follows that for a bounded
Lipschitz function \( s(u) \) the family of compositions \( s(u(t, \cdot)) \), \( t > 0 \), is compact in \( L^2(P) \). This easily implies the following result

**Lemma 2.3.** Let \( s(u) \) be a bounded Lipschitz function, \( v(t, x) = s(u(t, x)) \), and

\[
v(t, x) = \sum_{\kappa \in \mathbb{L}'} a_\kappa(t) e^{2\pi i \kappa \cdot x}
\]

be the Fourier series of \( v(t, \cdot) \) in \( L^2(\mathbb{R}^n) \), so that

\[
a_\kappa(t) = |P|^{-1} \int_{\mathbb{R}^n} e^{-2\pi i \kappa \cdot x} v(t, x) dx.
\]

Then this series converges to \( v(t, \cdot) \) in \( L^2(P) \) uniformly with respect to \( t \), that is, for each \( \varepsilon > 0 \) there exists \( N \in \mathbb{N} \) such that

\[
\sum_{|\kappa| > N} |a_\kappa(t)|^2 < \varepsilon^2 \quad \forall t > 0.
\]

To prove Theorem 1.3, we use, as in [2], the strong pre-compactness property for the self-similar scaling sequence \( u_k = u(kt, kx) \), \( k \in \mathbb{N} \). This pre-compactness property will be obtained under condition (7) with the help of localization principles for \( H \)-measures with “continuous indexes”, introduced in [11, 12].

First, we recall the original concept of \( H \)-measure invented by L. Tartar [22] and, independently, by P. Gerard [4]. Let \( F(u)(\xi), \xi \in \mathbb{R}^N, \) be the Fourier transform of a function \( u(x) \in L^2(\mathbb{R}^N) \), \( S = S^{N-1} = \{ \xi \in \mathbb{R}^N \mid |\xi| = 1 \} \) be the unit sphere in \( \mathbb{R}^N \). Denote by \( u \to \pi, \; u \in \mathbb{C} \) the complex conjugation.

Let \( \Omega \) be an open domain in \( \mathbb{R}^N \), and let \( U_k(x) \in L^2_{loc}(\Omega) \) be a sequence weakly convergent to the zero function.

**Proposition 2** (see Theorem 1.1 in [22]). There exists a nonnegative Borel measure \( \mu \) in \( \Omega \times S \) and a subsequence \( U_r(x) = U_k(x), \; k = k_r, \) such that

\[
\langle \mu, \Phi_1(x)\bar{\Phi}_2(x)\psi(\xi) \rangle = \lim_{r \to \infty} \int_{\mathbb{R}^N} F(\Phi_1 U_r)(\xi)F(\Phi_2 U_r)(\xi)\psi \left( \frac{\xi}{|\xi|} \right) d\xi
\]

for all \( \Phi_1(x), \Phi_2(x) \in C_0(\Omega) \) and \( \psi(\xi) \in C(S) \).

The measure \( \mu \) is called the Tartar \( H \)-measure corresponding to \( U_r(x) \).

We need also the concept of measure valued functions (Young measures). Recall (see [3, 21]) that a measure-valued function on a domain \( \Omega \subset \mathbb{R}^N \) is weakly measurable map \( x \to \nu_x \) of \( \Omega \) into the space \( \text{Prob}_0(\mathbb{R}) \) of probability Borel measures with compact support in \( \mathbb{R} \).

The weak measurability of \( \nu_x \) means that for each continuous function \( g(\lambda) \) the function \( x \to \langle \nu_x, g(\lambda) \rangle = \int g(\lambda) d\nu_x(\lambda) \) is measurable on \( \Omega \).

Measure-valued functions of the kind \( \nu_x(\lambda) = \delta(\lambda - u(x)) \), where \( u(x) \in L^\infty(\Omega) \) and \( \delta(\lambda - u^*) \) is the Dirac measure at \( u^* \in \mathbb{R} \), are called regular. We identify these measure-valued functions and the corresponding functions \( u(x) \), so that there is a natural embedding of \( L^\infty(\Omega) \) into the set \( \text{MV}(\Omega) \) of measure-valued functions on \( \Omega \).

Measure-valued functions naturally arise as weak limits of bounded sequences in \( L^\infty(\Omega) \) in the sense of the following theorem by L. Tartar [21].

**Theorem 2.4.** Let \( u_m(x) \in L^\infty(\Omega), \; m \in \mathbb{N} \), be a bounded sequence. Then there exist a subsequence (we keep the notation \( u_m(x) \) for this subsequence) and a measure valued function \( \nu_x \in \text{MV}(\Omega) \) such that

\[
\forall g(\lambda) \in C(\mathbb{R}) \quad g(u_m) \xrightarrow{m \to \infty} \langle \nu_x, g(\lambda) \rangle \quad \text{weakly-* in } L^\infty(\Omega).
\]
Besides, \( \nu_x \) is regular, i.e., \( \nu_x(\lambda) = \delta(\lambda - u(x)) \) if and only if \( u_m(x) \xrightarrow{m \to \infty} u(x) \) in \( L^1_{loc}(\Omega) \) (strongly).

In [11] the new concept of \( H \)-measures with “continuous indexes” was introduced, corresponding to sequences of measure valued functions. We describe this concept in the particular case of “usual” sequences in \( L^\infty(\Omega) \). Let \( u_m(x) \) be a bounded sequence in \( L^\infty(\Omega) \). Passing to a subsequence if necessary, we can suppose that this sequence converges to a measure valued function \( \nu \) and the corresponding distribution functions \( U_m(x,p) = \gamma^m p, +\infty) \), \( u_0(x,p) = \nu_x(p, +\infty) \) on \( \Omega \times \mathbb{R} \). Observe that \( U_m(x,p), u_0(x,p) \in L^\infty(\Omega) \) for all \( p \in \mathbb{R} \), see [11, Lemma 2]. We define the set

\[
E = E(\nu) = \left\{ p_0 \in \mathbb{R} \mid u_0(x,p) \xrightarrow{p \to p_0} u_0(x,p_0) \text{ in } L^1_{loc}(\Omega) \right\}.
\]

As was shown in [11, Lemma 4], the complement \( \mathbb{R} \setminus E \) is at most countable and if \( p \in E \) then \( U_m(x,p) \xrightarrow{m \to \infty} 0 \) weakly-* in \( L^\infty(\Omega) \).

The next result, similar to Proposition 2, has been established in [11, Theorem 3], [13, Proposition 2, Lemma 2].

**Proposition 3.** 1) There exists a family of locally finite complex Borel measures \( \{\mu^{pq}\}_{p,q \in E} \) in \( \Omega \times S \) and a subsequence \( U_r(x,p) = U_{m_r}(x,p) \) such that for all \( \Phi_1(x), \Phi_2(x) \in C_0(\Omega) \) and \( \psi(\xi) \in C(S) \)

\[
\langle \mu^{pq}, \Phi_1(x)\Phi_2(x)\psi(\xi) \rangle = \lim_{r \to \infty} \int_{\mathbb{R}^N} F(\Phi_1 U_r(\cdot,p))(\xi)\overline{F(\Phi_2 U_r(\cdot,q))(\xi)} \psi \left( \frac{\xi}{|\xi|} \right) d\xi;
\]

(11)

2) The correspondence \( (p,q) \to \mu^{pq} \) is a continuous map from \( E \times E \) into the space \( M_{loc}(\Omega \times S) \) of locally finite Borel measures on \( \Omega \times S \) (with the standard locally convex topology);

3) For any \( p_1, \ldots, p_l \in E \) the matrix \( \{\mu^{p,p_j}\}_{i,j=1}^l \) is Hermitian and positive semidefinite, that is, for all \( \zeta_1, \ldots, \zeta_l \in \mathbb{C} \) the measure

\[
\sum_{i,j=1}^l \mu^{p_i,p_j} \zeta_i \overline{\zeta_j} \geq 0.
\]

Notice that assertion 3) readily follows from relation (11).

We call the family of measures \( \{\mu^{pq}\}_{p,q \in E} \) the \( H \)-measure corresponding to the subsequence \( u_r(x) = u_{m_r}(x) \).

As was demonstrated in [11], the \( H \)-measure \( \mu^{pq} = 0 \) for all \( p,q \in E \) if and only if the subsequence \( u_r(x) \) converges as \( r \to \infty \) strongly (in \( L^1_{loc}(\Omega) \)). Observe also that assertion 3) in Proposition 3 implies that measures \( \mu^{pp} \geq 0 \) for all \( p \in E \), and that

\[
|\mu^{pq}(A)| \leq \sqrt{\mu^{pp}(A)\mu^{qq}(A)}
\]

(12)

for any Borel set \( A \subset \Omega \times S \) and all \( p,q \in E \).

3. **Main results.** We fix a periodic r.e.s. \( u = u(t,x) \) of (1), (2).

Let \( s(u) \) be a bounded Lipschitz function, \( v(t,x) = s(u(t,x)) \), and

\[
v(t,x) = \sum_{\kappa \in L'} a_\kappa(t)e^{2\pi i \kappa \cdot x}
\]

(13)
be the Fourier series of \( v(t, \cdot) \) in \( L^2(\mathbb{R}^n) \). Then
\[
v_k(t, x) = v(k t, k x) = \sum_{\kappa \in \mathbb{L}'} a_{\kappa}(k t) e^{2 \pi i \kappa x},
\]
which implies that, may be after extraction of a subsequence, \( v_k \to v^* \) as \( k \to \infty \) weakly-* in \( L^\infty(\Pi) \), where \( v^* = v^*(t) \) being the weak limit of the coefficient \( a_0(k t) \).
Let \( \hat{\mu} \) be the Tartar’s \( H \)-measure corresponding to the sequence \( v_r - v^* \), where \( v_r = v_{k_r}(t, x) \) is a subsequence of \( v_k \).

**Lemma 3.1.** The following inclusion holds: \( \text{supp} \hat{\mu} \subset \Pi \times S_0 \), where
\[
S_0 = \left\{ \xi/|\xi| \in S \mid \xi = (\tau, \xi) \neq 0, \tau \in \mathbb{R}, \xi \in L' \right\}.
\]

To prove Lemma 3.1, we use arguments similar to ones in [22, Example 2.1] and take into account the uniform convergence of series (13) (which holds by Lemma 2.3).

We fix \( l \in \mathbb{N} \) and consider the \( H \)-measure \( \{\mu_{pq}\}_{p,q \in E} \) corresponding to a subsequence \( v_r = v_{k_r}(t, x) \) of the sequence \( v_k = s_{-l,l}(u(k t, k x)) \), \( k \in \mathbb{N} \), defined in accordance with Proposition 3. Let \( s(u) \in C^1(\mathbb{R}) \). As easily follows from the definitions, the Tartar \( H \)-measure \( \hat{\mu} \) corresponding to the sequence \( s(v_r) - v^* \) (\( v^* \) being the weak limit of \( s(v_r) \)) satisfies the equality
\[
\hat{\mu} = \int_{[-l,l]^2} \mu_{pq} s(p) s'(q) dp dq.
\]

Using this representation, the statement of Lemma 3.1 and the continuity of \( \mu_{pq} \) with respect to \((p, q)\), we derive the following first localization principle.

**Theorem 3.2.** For every \( p, q \in E \) \( \text{supp} \mu_{pq} \subset \Pi \times S_0 \).

Observe that \( s_{a,b}(v_k) = s_{a,b}(u_k) \) for \( a, b \in [-l, l] \), \( a \leq b \) and by (4)
\[
(s_{a,b}(u_k))_t + \text{div}_x \varphi(s_{a,b}(v_k)) = \mu^k - \mu^k_a \quad \text{in} \quad \mathcal{D}'(\Pi),
\]
where \( \mu^k = \kappa \mu_p(k t, k x) \) in \( \mathcal{D}'(\Pi) \), that is, \( \{\mu^k, f(t, x)\} = k^{-n} \langle \mu_p, f(t/k, x/k) \rangle \) for each \( f(t, x) \) in \( C_0(\Pi) \), \( p \in \mathbb{R} \). By the periodicity of \( \mu_p \) this implies that
\[
\mu^k_p(\mathbb{R}+ \times X) = \int_{\mathbb{R}+ \times P} (|\varphi(x)| - |p|)^+ dx \leq C_p \int_P (|u_0(x)| - |p|)^+ dx,
\]
in view of Lemma 2.2, and we find that sequence of measures \( \mu^k - \mu^k_a \) is bounded in \( M(\mathbb{R}+ \times P) \). By the Murat interpolation lemma [10] this sequence is precompact in the Sobolev space \( H^{-1}_{loc}(\Pi) \). Then, by the results of [18, Theorems 4-6] the following second localization principle holds.

**Theorem 3.3.** Let \( X = X(p) \subset \mathbb{R}^{n+1} \) be the minimal linear subspace such that \( \text{supp} \mu^{pp} \subset \Pi \times X \). Then there exists \( \delta > 0 \) such that the function \( u \to \tau u + \xi \cdot \varphi(u) \) is constant on the interval \((p - \delta, p + \delta)\) for all \( \xi = (\tau, \xi) \in X \).

Now we are ready to prove our main Theorem 1.3. As follows from Theorems 3.2, 3.3, if \( \dim L > 0 \) then there exists nonzero vector \( (\tau, \xi) \in X \cap (\mathbb{R} \times L') \) such that the function \( \tau u + \xi \cdot \varphi(u) \) is constant on some interval \((p - \delta, p + \delta)\). Obviously then \( \xi \neq 0 \) and \( \xi \cdot \varphi(u) \) is affine on \((p - \delta, p + \delta)\). But this contradicts to nondegeneracy condition (7). We conclude that \( X = \{0\} \) and, therefore, \( \mu^{pp} = 0 \) for all \( p \in E \). In view of (12) the \( H \)-measure \( \mu^{pq} \) is trivial and the sequence \( v_r = s_{-l,l}(u_r) \) converges strongly in \( L^1_{loc}(\Pi) \). Since \( l \in \mathbb{N} \) is arbitrary we see that the same is true.
for the sequence \( u_r : u_r \to u^* \) in \( L^1_{loc}(\Omega) \). Here we also take into account that, in view of Lemma 2.1, the sequence \( u_r \) cannot produce concentration effects as \( r \to \infty \).

Besides, the limit function \( u^* = u^*(t) \) depends only on \( t \). Passing to the limit in the relation
\[
(s_{-l,t}(u_r))_t + \text{div}_x \varphi(s_{-l,t}(u_r)) = \mu^r_t - \mu^L_l
\]
as \( r \to \infty \) and then as \( l \to \infty \) (also taking into account that
\[
\mu^r_t(\mathbb{R}^+ \times P) + \mu^L_l(\mathbb{R}^+ \times P) \leq 2C_l \to 0,
\]
we arrive at the identity \( u^*_t = 0 \) showing that \( u^* = c = \text{const} \). The relation \( u_r(t,x) \to c \) in \( L^1_{loc}(\Omega) \) implies that (after possible extraction of a subsequence) for \( \text{a.e. } t > 0 \) \( u_r(t,x) \to c \) in \( L^1_{loc}(\mathbb{R}^n) \). By the periodicity, this reads
\[
\int_P |u(k_t,t,k_xx) - c| dx \to 0.
\]
Making the change of variables \( y = k_xx \) and using the space periodicity of \( u \), we find that for \( \text{a.e. } t > 0 \)
\[
\int_P |u(k_t,t,y) - c| dy = \int_P |u(k_t,t,k_xx) - c| dx \to 0.
\]
(14)

We fix such \( t = t_0 > 0 \). Then by Proposition 1 for each \( t > k_t \)
\[
\int_P |u(t,y) - c| dy \leq \int_P |u(k_t,t_0,y) - c| dy.
\]
(15)

In view of (14) it follows from (15) that \( \lim_{t \to \infty} u(t,x) = c \) in \( L^1(P) \). Besides, in view of Lemma 2.1 \( \int_P c dx = \int_P u_0(x) dx \). Hence, the constant
\[
c = |P|^{-1} \int_P u_0(x) dx
\]
and we claim that the decay property holds for every r.e.s. \( u(t,x) \).

Conversely, assume that condition (7) fails. Then we can find the segment \([a,b]\), \( a < b \), and a nonzero point \((\tau,\xi) \in \mathbb{R} \times L^r \) such that the function \( u \to \tau u + \xi \cdot \varphi(u) \) is constant on the segment \([a,b]\). Then, as is easy to verify, the function
\[
u(t,x) = \frac{a+b}{2} + \frac{b-a}{2} \sin(2\pi(\tau t + \xi \cdot x))
\]
is a periodic bounded e.s. of (1), which does not satisfy the decay property. The obtained contradiction shows that condition (7) is also necessary for the decay property. This completes the proof of our main Theorem 1.3.

The proof of Theorem 1.3 for bounded entropy solutions of equation (1) can be found in paper [20], see also preprint [19].

REFERENCES


E-mail address: Eugeny.Panov@novsu.ru
A MIXTURE-ENERGY-CONSISTENT NUMERICAL APPROXIMATION OF A TWO-PHASE FLOW MODEL FOR FLUIDS WITH INTERFACES AND CAVITATION

MARICA PELANTI
ENSTA ParisTech
Department of Mechanical Engineering
828, Boulevard des Maréchaux
91762 Palaiseau Cedex, France

KEH-MING SHYUE
National Taiwan University
Department of Mathematics
Taipei 106, Taiwan

Abstract. We model cavitating flows by a variant of the six-equation single-velocity two-phase model with stiff mechanical relaxation of Saurel–Petitpas–Berry [J. Comput. Phys. 228, 1678 (2009)]. In our approach we employ phasic total energy equations instead of the phasic internal energy equations of the classical system. This alternative formulation allows us to easily design a simple numerical method that ensures consistency with mixture total energy conservation at the discrete level and agreement of the relaxed pressure with the correct mixture equation of state. The two-phase system is solved in two dimensions by a fully-discretized high-resolution wave propagation scheme based on a HLLC/Roe Riemann solver. Numerical experiments show the ability of the numerical model to describe mechanical cavitation processes.

1. Introduction. The modelling of cavitating flows is relevant in numerous areas of engineering, from naval industry to aerospace technology. These flows involve complex thermo-hydrodynamic processes: liquid/vapor phase transition, dynamical creation of interfaces, shock formation. Our ultimate goal is to design new efficient methods for the simulation of cavitating fluids in the framework of diffuse-interface compressible multi-phase flow models [1, 5, 6, 7]. In this work we present the basic numerical model that we have developed, which allows the description of mechanical cavitation processes (with no phase transition). Extensions of this model to account for mass and heat transfer phenomena are reported in [3]. We begin by considering the hyperbolic six-equation single-velocity two-phase flow model with instantaneous pressure relaxation of Saurel–Petitpas–Berry [7], which we recall in Section 2. We propose a variant of this model, by employing phasic total energy equations instead of phasic internal energy equations in the mathematical formulation. This alternative formulation is presented in Section 3, and the numerical scheme used

2000 Mathematics Subject Classification. 65M08, 76T10.
Key words and phrases. Compressible two-phase flows, pressure relaxation, cavitation, finite volume schemes, Riemann solvers.
The first author (M. Pelanti) was partially supported by DGA (Direction Générale de l’Armement), Grant DGA 2009.60.035.00.470.75.01. The second author (K.-M. Shyue) was supported in part by the National Science Council of Taiwan Grant NSC 99-2115-M-002-005-MY2.
for the system solution is described in Section 4. The mixture-energy-consistency property of the numerical method is discussed in Section 4.1. Some numerical experiments are finally illustrated in Section 5.

2. The six-equation single-velocity two-phase flow model. The single-velocity two-phase flow model with stiff mechanical relaxation of Saurel et al. [7] (see also [8]) in one spatial dimension has the form

\[ \begin{align*}
\partial_t \alpha_1 + u \partial_x \alpha_1 &= \mu(p_1 - p_2), \\
\partial_t (\alpha_1 \rho_1) + \partial_x (\alpha_1 \rho_1 u) &= 0, \\
\partial_t (\alpha_2 \rho_2) + \partial_x (\alpha_2 \rho_2 u) &= 0, \\
\partial_t (p_1 u) + \partial_x (pu^2 + \alpha_1 \rho_1 + \alpha_2 \rho_2) &= 0, \\
\partial_t (\alpha_1 E_1) + \partial_x (\alpha_1 E_1 u) + \alpha_1 \rho_1 \partial_x u &= -p_1 \mu(p_1 - p_2), \\
\partial_t (\alpha_2 E_2) + \partial_x (\alpha_2 E_2 u) + \alpha_2 \rho_2 \partial_x u &= p_2 \mu(p_1 - p_2). \end{align*} \]

Here \( \alpha_k \) is the volume fraction of phase \( k, k = 1, 2 \), \( \rho_k \) is the phasic density, \( p_k \) the phasic pressure, and \( E_k \) the phasic internal energy, \( E_k = \rho_k \varepsilon_k \), where \( \varepsilon_k \) denotes the phasic specific internal energy. Moreover, \( \rho = \alpha_1 \rho_1 + \alpha_2 \rho_2 \) is the mixture density and \( u \) denotes the flow velocity. The source terms appearing in the volume fraction equation and in the phasic energy equations model mechanical relaxation. In these terms \( \mu \) is the pressure relaxation parameter and \( p_1 \) the interface pressure, \( p_1 = \frac{E_2 Z_1^2 + E_1 \rho_2}{Z_1} \), where \( Z_k = \rho_k c_k^2 \) is the acoustic impedance of phase \( k \), and \( c_k \) is the phasic sound speed. As in [7, 8], the pressure relaxation parameter \( \mu \) is assumed to be infinite, that is we consider instantaneous mechanical equilibrium. The closure of system (1) is obtained through the specification of equations of state for the two phases, which we express here in terms of \( E_k \) and \( \rho_k, p_k = p_k(E_k, \rho_k) \), \( k = 1, 2 \). The phasic sound speed can be written as \( c_k = \sqrt{\kappa_k \gamma_k - \chi_k} \), where \( h_k = \frac{E_k + \rho_k}{\rho_k} \) is the phasic specific enthalpy, \( \kappa_k = \frac{\partial p_k(E_k, \rho_k)}{\partial \rho_k} \) and \( \chi_k = \frac{\partial p_k(E_k, \rho_k)}{\partial E_k} \). The mixture sound speed is \( c = \sqrt{Y_1 \gamma_1 + Y_2 \gamma_2} \), where \( Y_k = \frac{2 \alpha_k \rho_k}{\rho} \) is the mass fraction of phase \( k \). Here we will restrict our study to the case of species governed by the stiffened gas law (SG EOS), \( p_k = (\gamma_k - 1)E_k - \gamma_k \pi_k - (\gamma_k - 1) \eta_k \rho_k \), where \( \gamma_k, \pi_k, \eta_k \) are constant parameters. The mixture specific internal energy is defined as \( \varepsilon = Y_1 \varepsilon_1 + Y_2 \varepsilon_2 \), and, equivalently, the mixture energy \( E = \rho \varepsilon \) is \( E = \alpha_1 E_1 + \alpha_2 E_2 \). This relation, by using \( E_k = E_k(p_k, \rho_k) \) with \( p_1 = p_2 = p \), determines the equation of state for the mixture pressure \( p = p(E, \rho_1, \rho_2, \alpha_1) \). In the particular case of the SG EOS we obtain

\[ p(E, \rho_1, \rho_2, \alpha_1) = \left( E - (\alpha_1 \rho_1 \eta_1 + \alpha_2 \rho_2 \eta_2) - \left( \frac{\alpha_1 \gamma_1 \pi_1}{\gamma_1 - 1} + \frac{\alpha_2 \gamma_2 \pi_2}{\gamma_2 - 1} \right) \right) \left( \frac{\alpha_1}{\gamma_1 - 1} + \frac{\alpha_2}{\gamma_2 - 1} \right). \]

Let us finally remark that the model (1) is hyperbolic and it has eigenvalues given by \( \lambda_1 = u - c, \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = u, \lambda_6 = u + c \).

3. Phasic-total-energy-based formulation. For numerical reasons that we will discuss in the following, we propose to consider a mathematically equivalent formulation of the model (1), obtained by replacing the two phasic internal energy equations with two phasic total energy equations. We denote with \( E_k \) the total
energy of phase $k$, $E_k = \mathcal{E}_k + \frac{1}{2}\rho_k u^2$. The two-phase model (1) then takes the form

\begin{align}
\partial_t \alpha_1 + u \partial_x \alpha_1 &= \mu(p_1 - p_2), \\
\partial_t (\alpha_1 \rho_1) + \partial_x (\alpha_1 \rho_1 u) &= 0, \\
\partial_t (\alpha_2 \rho_2) + \partial_x (\alpha_2 \rho_2 u) &= 0, \\
\partial_t (\rho u) + \partial_x (\rho u^2 + \alpha_1 p_1 + \alpha_2 p_2) &= 0, \\
\partial_t (\alpha_1 E_1) + \partial_x (\alpha u_1 (E_1 + p_1)) + \Sigma(q, \partial_x q) &= -p_1 \mu(p_1 - p_2), \\
\partial_t (\alpha_2 E_2) + \partial_x (\alpha u_2 (E_2 + p_2)) - \Sigma(q, \partial_x q) &= p_1 \mu(p_1 - p_2),
\end{align}

where the non-conservative contribution $\Sigma$ in the phasic total energy equations is

\begin{align}
\Sigma(q, \partial_x q) &= -u((Y_2 p_1 + Y_1 p_2)\partial_x \alpha_1 + \alpha_1 Y_2 \partial_x p_1 - \alpha_2 Y_1 \partial_x p_2) \\
&= -u(Y_2 \partial_x (\alpha_1 p_1) - Y_1 \partial_x (\alpha_2 p_2)),
\end{align}

and where we have denoted with $q$ the vector of the system unknowns. The sum of the phasic energy equations (3e) and (3f) recovers the equation expressing conservation of the mixture total energy $E = \mathcal{E} + \rho u^2 = \alpha_1 E_1 + \alpha_2 E_2$:

\begin{equation}
\partial_t E + \partial_x (u E + \alpha_1 p_1 + \alpha_2 p_2)) = 0.
\end{equation}

In compact form the system reads:

\begin{equation}
q = \begin{pmatrix}
\alpha_1 \\
\alpha_1 \rho_1 \\
\alpha_2 \rho_2 \\
\rho u \\
\alpha_1 E_1 \\
\alpha_2 E_2
\end{pmatrix},
\begin{pmatrix}
0 \\
\alpha_1 \rho_1 u \\
\alpha_2 \rho_2 u \\
\rho u^2 + \alpha_1 p_1 + \alpha_2 p_2 \\
\alpha_1 u_1 (E_1 + p_1) \\
\alpha_2 u_2 (E_2 + p_2)
\end{pmatrix},
\begin{pmatrix}
\partial_t \alpha_1 \\
\partial_x (\alpha_1 \rho_1 u) \\
\partial_x (\alpha_2 \rho_2 u) \\
\partial_x (\rho u^2 + \alpha_1 p_1 + \alpha_2 p_2) \\
\partial_x (\alpha_1 u_1 (E_1 + p_1)) \\
\partial_x (\alpha_2 u_2 (E_2 + p_2))
\end{pmatrix},
\begin{pmatrix}
\mu(p_1 - p_2) \\
0 \\
0 \\
0 \\
\Sigma \\
-\Sigma
\end{pmatrix},
\psi = \begin{pmatrix}
\psi(q) \\
\partial_x f(q)
\end{pmatrix},
\end{equation}

Above, we have put into evidence the conservative portion of the system $\partial_x f(q)$, where $f(q)$ is the flux function, and the non-conservative term $\sigma(q, \partial_x q)$. The source term $\psi(q)$ contains the pressure relaxation terms.

4. **Numerical method.** To numerically solve system (5) we use a fractional step technique, similar to [7, 8], where we alternate between the solution of the homogeneous hyperbolic system and the solution of a system of ordinary differential equations (ODEs) that takes into account pressure relaxation source terms. The second step drives the two-phase flow to mechanical equilibrium.

The steps of the algorithm are the following:

1. **Homogeneous hyperbolic system.** We solve over a time interval $\Delta t$ the homogeneous hyperbolic portion of (5):

\begin{equation}
\partial_t q + \partial_x f(q) + \sigma(q, \partial_x q) = 0.
\end{equation}

2. **Pressure relaxation.** We solve in the limit $\mu \to \infty$ the system of ordinary differential equations (ODEs)

\begin{equation}
\partial_t q = \psi(q).
\end{equation}

This step relaxes the phasic pressures to an equilibrium value $p_1 = p_2 = p$ (relaxed pressure). In this step the partial densities, the mixture momentum, the mixture total energy and the mixture internal energy remain constant.
The volume fraction $\alpha_1$, the mixture pressure $p$ and the phasic energies $\alpha_kE_k$, 
$E_k = E_k(p,(\alpha_k\rho_k)/\alpha_k)$ are updated before returning to Step 1.

4.1. **Mixture-energy-consistent discretization.** Let us denote with superscript 0 the quantities computed by solving the homogeneous system in Step 1 of the algorithm above, and with superscript * the quantities at mechanical equilibrium computed in Step 2. Let us also denote with $E^{0,C}$ discrete values of the mixture total energy that come from a conservative approximation of the conservation law for $E$ in (4).

**Definition 4.1.** We say that the numerical scheme based on the fractional step algorithm above is mixture-energy-consistent if the following two properties are satisfied

(i) **Mixture total energy conservation consistency:**

$$E^0 = E^{0,C}, \quad \text{where} \quad E^0 = (\alpha_1E_1)^0 + (\alpha_2E_2)^0. \quad (8a)$$

(ii) **Relaxed pressure consistency:**

$$E^{0,C} = \alpha_1E_1(p^*,(\alpha_1\rho_1)^0\gamma_1) + \alpha_2E_2(p^*,(\alpha_2\rho_2)^0\gamma_2), \quad \text{where} \quad E^{0,C} = E^{0,C} - \frac{(\rho u_2)^2}{\rho_2^2}. \quad (8b)$$

The first property (i) means that the sum of the discrete values of the phasic total energies given by the solution of the homogeneous system must recover discrete values of the mixture total energy that are consistent with a conservative discrete form of (4). The second property (ii) means that the value of the relaxed (equilibrium) pressure $p^*$ predicted in the relaxation step must be equal to the pressure as computed through the mixture equation of state $p(E^{0,C},\alpha_1^*,\rho_1^*,\rho_2^*)$. In the particular case of the SG EOS this consistency condition reads

$$p^* = \left(\frac{E^{0,C} - (\alpha_1\rho_1)^0\gamma_1 + (\alpha_2\rho_2)^0\gamma_2 - (\alpha_1^*\gamma_1\pi_1 + \alpha_2^*\gamma_2\pi_2)}{\gamma_1 + \gamma_2}\right) \bigg/ \left(\frac{\alpha_1^*}{\gamma_1 - 1} + \frac{\alpha_2^*}{\gamma_2 - 1}\right). \quad (9)$$

The mathematical formulation of the two-phase model with the phasic total energy equations (3) easily allows us to satisfy both properties (i) and (ii). To ensure property (i), it suffices in Step 1 to apply a standard conservative scheme to the conservative portion of the energy equations $\partial_t(\alpha_kE_k) + \partial_x(\alpha_ku_kE_k + \alpha_kp_k)$, $k = 1, 2$, and to discretize symmetrically the non-conservative contribution $\Sigma$ written in (3g). In such a way the sum of the discrete non-conservative energy equations recovers a conservative discrete form of the mixture energy equation, as a consequence of the cancellation of non-conservative discrete contributions. The fulfillment of mixture total energy conservation consistency then easily enables us to ensure also property (ii), the agreement of the relaxed equilibrium pressure with the correct mixture equation of state. See the simple pressure relaxation procedure for Step 2 in Section 4.3. On the other hand, it appears difficult to obtain a mixture-energy-consistent scheme if we apply an analogous fractional step algorithm to the classical 6-equation two-phase model (1). Although clearly both formulations (1) and (3) mathematically recover the conservation law for the mixture total energy, it is hard to discretize the phasic internal energies equations (1c) and (1f) in a way that recovers a conservative discrete form of (4). Indeed, numerical models such as [7, 8] built on the formulation (1) need to augment the 6-equation system with the equation for $E$. The additional conservation law for $E$ is solved through a standard conservative scheme to obtain consistent discrete values $E^{0,C}$. These values are then used to correct the thermodynamic state predicted by the non-conservative
consistent discrete values of the mixture total energy from the computed phasic
Since the wave propagation scheme that we employ in Step 1 recovers conservation-
\[ p^* = p(E^{0,C}, \alpha_1^*, \rho_1^*, \rho_2^*). \]

4.2. Solution of the homogeneous system. To solve the homogeneous system
the approximate solution of \( q \) at the \( i \)th cell and at time \( t^n \), the updating formula
\[ Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x} (A^+ \Delta Q_{i-1/2} + A^- \Delta Q_{i+1/2}) - \frac{\Delta t}{\Delta x} (F^h_{i+1/2} - F^h_{i-1/2}), \quad (10) \]
where \( A_{i+1/2}^\pm \Delta Q \) are the so-called fluctuations at interfaces \( x_{i+1/2} \) between cells
\( i \) and \( (i + 1) \), and \( F^h_{i+1/2} \) are correction fluxes for second order resolution.
The fluctuations \( A_{i+1/2}^\pm \Delta Q \) are computed by solving Riemann problems at cells
interfaces for each pair of data \( Q_i^n, Q_i^{n+1} \). A Riemann solver must be supplied to
perform this task. We have developed two approximate Riemann solvers for the
model system (5): a HLLC solver, similar to the solvers described in [7, 8], and a
new Roe solver [4]. Both solvers can be easily designed to ensure mixture-energy-
consistency (Def. 4.1), see [3] for details. The expression of the Roe eigenstructure
for the Roe method is reported in Appendix A. The HLLC method, as expected, is
more robust than the Roe method, and it is the method that we choose to employ
for the one-dimensional scheme and for the computation of normal waves in the
two-dimensional scheme. The Roe eigenstructure is useful in the two-dimensional
wave propagation algorithm to define transverse fluctuations [2]. This is done by
projecting the normal fluctuations obtained via the HLLC solver onto the basis of
Roe eigenvectors associated to the orthogonal direction.

4.3. Pressure relaxation step. We consider Step 2 of the fractional step
algorithm (solution of (7)), which drives the system to mechanical equilibrium.
We employ a procedure similar to the relaxation technique illustrated in [7]. We
will use here the superscript notation introduced in Section 4.1. We easily see that
the solution of the system of equations (7) gives \((\alpha_k \rho_k)^* = (\alpha_k \rho_k)^0, \quad k = 1, 2, \)
\((p u)^* = (p u)^0, \quad E^* = E^0, \quad E^* = E^0. \) By assuming a linear variation of the interface
pressure \( p_1 \) with \( \alpha_1 \), by integration we obtain the two equations:
\[ (\alpha_k E_k)^* - (\alpha_k E_k)^0 = (\alpha_k \mathcal{E}_k)^* - (\alpha_k \mathcal{E}_k)^0 = (-1)^k \frac{p^0_1 + p^0_2}{2} (\alpha^*_1 - \alpha^*_2), \quad k = 1, 2. \quad (11) \]
At the final time we impose mechanical equilibrium: \( p^1_1 = p^2_2 = p^*_2 = p^*. \) Then,
by using the phasic equations of state \( \mathcal{E}_k = \mathcal{E}(\rho_k, \rho_k) \), the two equations above
give the solution for the equilibrium values \( \alpha^*_1 \) and \( p^* \). For the particular case of
the SG EOS we obtain a simple quadratic equation for the relaxed pressure \( p^* \)
(which has always a physically admissible solution). Let us now remark that the
values \( \alpha^*_1 \) and \( p^* \) by construction satisfy \( E^0 = \alpha^*_1 \mathcal{E}_1(p^*, (\alpha_1 \rho_1)^0) + \alpha^*_2 \mathcal{E}_2(p^*, (\alpha_2 \rho_2)^0) \).
Since the wave propagation scheme that we employ in Step 1 recovers conservation-
consistent discrete values of the mixture total energy from the computed phasic
energies, \( E^0 = (\alpha_1 E_1)^0 + (\alpha_2 E_2)^0 = E^{0,C} \), then we deduce that the simple pressure
relaxation procedure described here computes the thermodynamically correct value
of the equilibrium pressure \( p^* \). Both properties (i) and (ii) of Definition 4.1 are
ensured, and therefore the numerical scheme is mixture-energy-consistent.
5. Numerical experiments.

5.1. 1D cavitation tube. We perform the one-dimensional cavitation tube test proposed in [7] (see also [5]). We consider a tube filled with liquid water at atmospheric pressure, \( p = 10^5 \) Pa. The liquid (\( \rho_l = 1000 \) kg/m\(^3\)) contains initially a uniformly distributed small amount of air (\( \alpha_g = 10^{-2}, \rho_g = 1 \) kg/m\(^3\)). The SG EOS parameters are \( \gamma_l = 4.4, \pi_l = 6 \times 10^8 \) Pa, \( \eta_l = 0 \) J/kg for the liquid, and \( \gamma_g = 1.4, \pi_g = 0 \) Pa, \( \eta_g = 0 \) J/kg for air. A velocity discontinuity is set at the middle of the tube at initial time, with \( u = -100 \) m/s on the left and \( u = 100 \) m/s on the right. Two strong rarefactions symmetrically propagating in opposite directions are generated, which induce a decrease of the liquid pressure in the middle region of the tube. In this region the gas volume fraction increases, generating a cavitation pocket (mechanical cavitation). Second-order results at time \( t = 1.85 \) ms obtained with the HLLC solver (and MC limiter) with 1000 grid cells are shown in Figure 1 (CFL = 0.5). They are in good agreement with those reported in [7]. We have also tested the Roe solver for this experiment and results are not distinguishable from those obtained with the HLLC solver.

5.2. Cavitating Richtmyer–Meshkov instability. We consider the cavitating water-gas Richtmyer–Meshkov instability test problem proposed in [7]. A curved interface (with left-facing convexity) at initial time separates a region filled with nearly pure water (\( \alpha_g = 10^{-6} \)) and a region of nearly pure gas (\( \alpha_l = 10^{-6} \)). The initial pressure is \( p = 10^5 \) Pa, and the densities \( \rho_l = 1000 \) kg/m\(^3\) and \( \rho_g = 100 \) kg/m\(^3\). Both water and gas have an initial velocity of \(-200\) m/s. The SG EOS parameters are those used in [7]. Top, bottom and left boundaries are solid walls, whereas the right side is a free flow boundary. When the flow impinges against the left wall a right-going shock wave propagates through the curved water-gas interface generating a Richtmyer–Meshkov instability. The pressure decrease in the expansion zones close to the left wall generates cavitation pockets in this region. Second-order results obtained with the HLLC solver and Roe transverse splitting are displayed in Figure 2 (grid with \( 600 \times 200 \) cells, CFL = 0.5), and they qualitatively agree with those reported in [7].

5.3. High-velocity underwater projectile. We perform the high-velocity underwater projectile numerical experiment presented in [6] (see also [5]). Liquid water at speed \( u = 600 \) m/s flows over an immersed obstacle with hexagonal section. Initially the liquid is at atmospheric pressure, \( p = 10^5 \) Pa, and it has density \( \rho_l = 1500 \) kg/m\(^3\). A small amount of vapor is present in the liquid at initial time, \( \alpha_g = 10^{-3} \). The SG EOS for the two phases are those used in [6]. Due to the symmetry of the problem, we carry the computation only on the portion of the physical domain above the symmetry axis. We use a uniform grid with \( 600 \times 200 \) cells...
over a rectangular computational domain that is mapped to a curvilinear grid in the physical domain. The flow interaction with the edges of the obstacle generates strong rarefaction waves, which determine a pressure decrease and consequently the growth of a cavitation wake. After a time sufficiently large the flow reaches a configuration with a stationary cavitation zone. Second-order numerical results obtained with the HLLC solver are shown in Figure 3 (CFL = 0.5).

Figure 2. Cavitating Richtmyer–Meshkov instability experiment. Density at $t = 1.72$ and $t = 8.6$ ms.

Figure 3. High-speed underwater projectile experiment. Density and gas volume fraction at $t = 0.9$ s (approximately stationary conditions).

6. Conclusions. We have developed a new numerical model for two-phase compressible flows based on the single-velocity 6-equation two-phase flow model with stiff pressure relaxation of Saurel et al. [7]. We have proposed a variant of the standard model system [7, 8] that easily allows us to design a mixture-energy-consistent numerical scheme, in contrast with the classical formulation. A fully-discretized two-dimensional high-resolution wave propagation scheme has been developed for the system solution. We have already developed an extended numerical model that includes thermal and chemical potential relaxation source terms to describe heat and mass transfer processes. This is illustrated in [3].

Appendix A. Roe matrix and eigenstructure. We consider the quasi-linear form of system (6), $\partial_t q + A(q)\partial_x q = 0$, and following the classical Roe approach [4], we define an approximate solver by using the exact solution to the Riemann problem for a linearized system $\partial_t q + \tilde{A}(q_L, q_R)\partial_x q = 0$ with initial left and right data $q_L$ and $q_R$. The definition of the Roe matrix $\tilde{A}(q_L, q_R)$ must guarantee conservation of the partial densities $\alpha_k \rho_k$, $k = 1, 2$, the mixture momentum $\rho u$, and the mixture total energy $E$. That is, denoting with $q^{(l)}$ the $l$th component of $q$, and with $f^{(l)}$ the associated flux function ($l$th component of $f(q)$), the Roe matrix $\tilde{A}$ must be defined such that $\sum_{j=1}^6 \tilde{A}_{lj}(q_r - q_l) = f^{(l)}(q_r) - f^{(l)}(q_l)$, $l = 2, 3, 4$, and $\sum_{j=1}^6 (\tilde{A}_{5j} + \tilde{A}_{6j})(q_r - q_l) = f_E(q_r) - f_E(q_L)$, where $\tilde{A}_{lj}$ is the entry $(l, j)$ of the matrix $\tilde{A}$, and $f_E = u(E + \alpha_1 p_1 + \alpha_2 p_2)$ is the flux function for the mixture total...
where the averaged quantities \( \hat{\mathcal{A}} \) is defined as the matrix \( \hat{\mathcal{A}}(q) \) of the original system evaluated in an average state \( \hat{q} = \hat{q}(\hat{q}_t, \hat{q}_r) \). By imposing the conservation conditions above, a Roe matrix for the model system (6) closed with the SG EOS can be determined as \( \hat{\mathcal{A}} \).

Above \( (6) \) closed with the SG EOS can be determined as \( \hat{\mathcal{A}} = A(\hat{q}_t, \hat{q}_r) \). Following again \( [4] \), the Roe matrix \( \hat{\mathcal{A}} \) is defined as 

\[
\hat{\mathcal{A}} = \begin{pmatrix}
\hat{u}_1 \\
\hat{u}_2 \\
\hat{u}_3 \\
\hat{u}_4 \\
\hat{u}_5 \\
\hat{u}_6
\end{pmatrix} = \begin{pmatrix}
\hat{u}_1 \\
\hat{u}_2 \\
\hat{u}_3 \\
\hat{u}_4 \\
\hat{u}_5 \\
\hat{u}_6
\end{pmatrix}, \quad \hat{u}_i = \frac{u_i \sqrt{\rho \kappa} + \sqrt{\rho \kappa} u_r}{\sqrt{\rho \kappa} + \sqrt{\rho \kappa}}, \quad \hat{Y}_k = \frac{Y_{k_t} \sqrt{\rho \kappa} + \sqrt{\rho \kappa} Y_{k_r}}{\sqrt{\rho \kappa} + \sqrt{\rho \kappa}}, \quad k = 1, 2.
\]

Note that \( \hat{Y}_1 + \hat{Y}_2 = 1 \) and \( \hat{u}Y_1 + \hat{u}Y_2 = \hat{u} \). The Roe eigenvalues are given by

\[
\hat{\lambda}_1 = \hat{u} - \hat{c}, \quad \hat{\lambda}_2 = \hat{\lambda}_3 = \hat{\lambda}_4 = \hat{\lambda}_5 = \hat{u}, \quad \hat{\lambda}_6 = \hat{u} + \hat{c},
\]

where \( \hat{c} = \sqrt{\hat{Y}_1 \hat{c}_1^2 + \hat{Y}_2 \hat{c}_2^2} \) with \( \hat{Y}_k \hat{c}_k^2 = \kappa_k (\hat{Y}_k H_k - \frac{\hat{u}^2}{2} \hat{Y}_k) + \chi_k \hat{Y}_k \).

The corresponding matrix of the Roe right eigenvectors, \( \hat{\mathcal{R}} = [\hat{r}_1, \cdots, \hat{r}_6] \), is

\[
\hat{\mathcal{R}} = \begin{pmatrix}
0 & 0 & 0 & 0 & 1 & 0 \\
\hat{Y}_1 & 0 & 0 & 1 & 0 & \hat{Y}_1 \\
\hat{Y}_2 & 0 & 1 & 0 & 0 & \hat{Y}_2 \\
\hat{u} - \hat{c} & 0 & \hat{u} & \hat{u} & 0 & \hat{u} + \hat{c} \\
\hat{Y}_1 \hat{H}_1 - \hat{u} \hat{Y}_1 \hat{c} & -\frac{\hat{u}^2}{2} \hat{Y}_1 \hat{c}_1 + \frac{\hat{u}^2}{2} \hat{Y}_1 \hat{c}_1 & -\frac{\hat{u}^2}{2} \hat{Y}_1 \hat{c}_1 & -\frac{\hat{u}^2}{2} \hat{Y}_1 \hat{c}_1 & \hat{H}_1 - \hat{H}_2 & \hat{Y}_1 \hat{H}_1 + \hat{u} \hat{Y}_1 \hat{c} \\
\hat{Y}_2 \hat{H}_2 - \hat{u} \hat{Y}_2 \hat{c} & 1 & 0 & 0 & 0 & \hat{Y}_2 \hat{H}_2 + \hat{u} \hat{Y}_2 \hat{c}
\end{pmatrix}
\]

Above \( H_k = \hat{h}_k + \frac{\hat{u}^2}{2} \), \( \kappa_k = (\gamma_k - 1) \), \( \chi_k = - (\gamma_k - 1) \eta_k \) and \( \hat{H}_k = \gamma_k \pi_k, k = 1, 2 \).

REFERENCES


E-mail address: marica.pelanti@ensta-paristech.fr
E-mail address: shyue@ntu.edu.tw
A POSTERIORI ESTIMATES FOR EULER AND NAVIER-STOKES EQUATIONS

CARLO MOROSI
Dipartimento di Matematica, Politecnico di Milano
P.za Leonardo da Vinci 32, I-20133 Milano, Italy

MARIO PERNICI
Istituto Nazionale di Fisica Nucleare, Sezione di Milano
Via Celoria 16, I-20133 Milano, Italy

LIVIO PIZZOCCHERO
Dipartimento di Matematica, Università di Milano
Via Cesare Saldini 50, I-20133 Milano, Italy
and
Istituto Nazionale di Fisica Nucleare
Sezione di Milano, Italy

Abstract. The first two sections of this work review the framework of [6] for approximate solutions of the incompressible Euler or Navier-Stokes (NS) equations on a torus $T^d$, in a Sobolev setting. This approach starts from an approximate solution $u_a$ of the Euler/NS Cauchy problem and, analyzing it a posteriori, produces estimates on the interval of existence of the exact solution $u$ and on the distance between $u$ and $u_a$. The next two sections present an application to the Euler Cauchy problem, where $u_a$ is a Taylor polynomial in the time variable $t$; a special attention is devoted to the case $d = 3$, with an initial datum for which Behr, Nečas and Wu have conjectured a finite time blowup [1]. These sections combine the general approach of [6] with the computer algebra methods developed in [9]; choosing the Behr-Nečas-Wu datum, and using for $u_a$ a Taylor polynomial of order 52, a rigorous lower bound is derived on the interval of existence of the exact solution $u$, and an estimate is obtained for the $H^3$ Sobolev distance between $u(t)$ and $u_a(t)$.

1. Preliminaries. Throughout this work we fix a space dimension $d \in \{2, 3, \ldots\}$; in the application of section 4 we will put $d = 3$. For $a, b \in \mathbb{R}^d$ or $\mathbb{C}^d$ we put $a \cdot b := \sum_{r=1}^d a_r b_r$ and $|a| := \sqrt{a^\ast a}$, with $^\ast$ indicating the complex conjugate.

Let us consider the $d$-dimensional torus $T^d := (\mathbb{R}/2\pi\mathbb{Z})^d$; we denote with $(e_k)_{k \in \mathbb{Z}^d}$ the Fourier basis made of the functions $e_k : T^d \to \mathbb{C}$, $e_k(x) := (2\pi)^{-d/2} e^{ik \cdot x}$. Here and in the sequel, “a vector field on $T^d$ means “an $\mathbb{R}^d$-valued distribution on $T^d$” (see, e.g., [5]); we write $D'(T^d) \equiv \mathcal{D}'$ for the space of such distributions. Any $v \in \mathcal{D}'$ has a weakly convergent Fourier expansion $v = \sum_{k \in \mathbb{Z}^d} v_k e_k$, with coefficients $v_k \in \mathbb{C}^d$ such that $\overline{v_k} = v_{-k}$.

2000 Mathematics Subject Classification. Primary: 35Q31, 35Q30; Secondary: 76B03, 76D03, 68W30.

Key words and phrases. Euler and Navier-Stokes equations, existence and regularity theory, theoretical approximation, symbolic computation.
Furthermore, we choose a “forcing” (in the above, $C.$ MOROSI, M. PERNICI AND L. PIZZOCCHERO

\[ \nu \]

all $m$ and $n$

\[ v, w \]

Consider two vector fields The bilinear map for the Euler or Navier-Stokes (NS) equations.

1.1. The bilinear map for the Euler or Navier-Stokes (NS) equations. Consider two vector fields $v, w$ on $\mathbb{T}^d$ such that $v \in L^2$ and $\partial_r w \in L^2$ for $r = 1, \ldots, d$; then we have a well defined vector field $v \partial_r w \in L^1$ of components $(v \partial_r w)_r := \sum_{r=1}^d v_r \partial_r w_r$; we can apply to this the Leray projection $\mathcal{L}$, sending $\mathbb{D}'$ onto the space of divergence free vector fields, and form the vector field

\[ P(v, w) := \mathcal{L}(v \partial_r w) . \]  

The bilinear map $P : (v, w) \mapsto P(v, w)$, which is a main character of the incompressible Euler/NS equations, is known to possess the following properties:

(i) For each $n > d/2$, $P$ is continuous from $\mathbb{H}^n_\mathcal{L} \times \mathbb{H}^{n+1}_\mathcal{L}$ to $\mathbb{H}^n_\mathcal{L}$, so there is a constant $K_n = K_n$ such that

\[ \| P(v, w) \|_n \leq K_n \| v \|_n \| w \|_{n+1} \quad \text{for } v \in H_\mathcal{L}^n, w \in H_\mathcal{L}^{n+1} . \]  

(ii) For each $n > d/2 + 1$, there is a constant $G_n = G_n$ such that

\[ \| (P(v, w)|w) \|_n \leq G_n \| v \|_n \| w \|_n \quad \text{for } v \in H_\mathcal{L}^n, w \in H_\mathcal{L}^{n+1} . \]  

The result (ii) is due to Kato, see [3]. In papers [7] [8], (1.4) and (1.5) are called the “basic inequality” and the “Kato inequality”, respectively; in these papers, computable upper and lower bounds are given for the sharp constants appearing therein. From here to the end of this work, $K_n$ and $G_n$ are constants fulfilling the previous inequalities (and not necessarily sharp). From [7] [8] we know that we can take

\[ K_3 = 0.323 , \quad G_3 = 0.438 \quad \text{if } d = 3 ; \]  

these values will be useful in the sequel.

1.2. The Euler/NS Cauchy problem. Let us fix a Sobolev order

\[ n \in \left\{ \left\lceil \frac{d}{2} \right\rceil + 1, +\infty \right\} . \]  

We choose a “viscosity coefficient” $\nu \in [0, +\infty)$, and put

\[ \mathcal{P} := \left\{ \begin{array}{ll} 1 & \text{if } \nu = 0, \\ 2 & \text{if } \nu > 0. \end{array} \right. \]  

Furthermore, we choose a “forcing”

\[ f \in C([0, +\infty), H_\mathcal{L}^n) \]  

and an initial datum
\[ u_0 \in H^{n+\sigma}_{20}. \]  

**Definition 1.1.** The Cauchy problem for the (incompressible) fluid with viscosity \( \nu \), initial datum \( u_0 \) and forcing \( f \) is the following:

Find \( u \in C(0, T, H^{n+\sigma}_{20}) \cap C^1([0, T), H^n_{20}) \) such that
\[
\frac{du}{dt} = \nu \Delta u + P(u, u) + f, \quad u(0) = u_0
\]  
(with \( T \in (0, +\infty) \), depending on \( u \)). As usually, we speak of the “Euler Cauchy problem” if \( \nu = 0 \), and of the “NS Cauchy problem” if \( \nu > 0 \).

It is known [4] that the above Cauchy problem has a unique maximal (i.e., nonextendable) solution; any solution is a restriction of the maximal one.

2. **Approximate solutions of the Euler/NS Cauchy problem.** We consider again the Cauchy problem (1.11), for given \( n, \nu, f, u_0 \) as in the previous section. The definitions and the theorem that follow are taken from [6].

**Definition 2.1.** An approximate solution of problem (1.11) is any map \( u_a \in C([0, T_a), H^{n+\sigma}_{20}) \cap C^1([0, T_a), H^n_{20}) \) (with \( T_a \in (0, +\infty] \)). Given such a function, we stipulate (i) (ii).

(i) The differential error of \( u_a \) is
\[
\frac{du_a}{dt} - \nu \Delta u_a - P(u_a, u_a) - f \in C([0, T_a), H^n_{20}) ;
\]
the datum error is
\[
u u_a(0) - u_0 \in H^{n+\sigma}_{20}.
\]

(ii) Let \( m \in R, m \leq n \). A differential error estimator of order \( m \) for \( u_a \) is a function
\[
\epsilon_m \in C([0, T_a), [0, +\infty)) \text{ such that }
\|
\frac{du_a}{dt} - \nu \Delta u_a - P(u_a, u_a) - f\|_m \leq \epsilon_m(t) \text{ for } t \in [0, T_a).
\]

Let \( m \in R, m \leq n + \sigma \). A datum error estimator of order \( m \) for \( u_a \) is a real number
\[
\delta_m \in [0, +\infty) \text{ such that } \|u_a(0) - u_0\|_m \leq \delta_m ;
\]
a growth estimator of order \( m \) for \( u_a \) is a function
\[
D_m \in C([0, T_a), [0, +\infty)) \text{ such that } \|u_a(t)\|_m \leq D_m(t) \text{ for } t \in [0, T_a).\]

In particular \( \epsilon_m(t) := \|\frac{du_a}{dt} - \nu \Delta u_a - P(u_a, u_a) - f\|_m, \delta_m := \|u_a(0) - u_0\|_m \) and \( D_m(t) := \|u_a(t)\|_m\) will be called the tautilogical estiamtes of order \( m \) for the differential error, the datum error and the growth of \( u_a \).

From here to the end of the section we consider an approximate solution \( u_a \) of problem (1.11) of domain \([0, T_a]\); this is assumed to possess differential, datum error and growth estimators of orders \( n \) or \( n + 1 \), indicated with \( \epsilon_n, \delta_n, D_n, D_{n+1} \).

**Definition 2.2.** Let \( R_n \in C([0, T_c], [0, +\infty)) \), with \( T_c \in (0, T_a] \). This function is said to fulfil the control inequalities if
\[
\frac{d^+ R_n}{dt} \geq -\nu R_n + (G_n D_n + K_n D_{n+1}) R_n + G_n R_n^2 + \epsilon_n \text{ everywhere on } [0, T_c),
\]
\[
R_n(0) \geq \delta_n.
\]

In the above \( d^+/dt \) indicates the right, upper Dini derivative: so, for all \( t \in [0, T_c) \),
\[
(d^+ R_n/dt)(t) := \lim \sup_{h \to 0^+} (R_n(t + h) - R_n(t))/h.
\]
Proposition 2.1. Assume there is a function $R_n \in C([0,T_\nu],[0, +\infty))$ fulfilling the control inequalities; consider the maximal solution $u$ of the Euler/NS Cauchy problem (1.11), and denote its domain with $[0,T)$. Then

$$T \geq T_\nu,$$

$$\|u(t) - u_k(t)\|_n \leq R_n(t) \quad \text{for } t \in [0, T_\nu).$$

Proof. (Sketch) One introduces the function $\|u - u_k\|_n : t \in [0, T) \cap [0, T_\nu) \mapsto \|u(t) - u_k(t)\|_n$ and shows that $\|u(t) - u_k(t)\|_n / dt \leq -\nu \|u - u_k\|_n + (G_n D_n + K_n D_{n+1}) \|u - u_k\|_n + G_n \|u - u_k\|_n^2 + \epsilon_n$, (see Lemma 4.2 of [6], greatly indebted to [2]); moreover, $\|u(0) - u_k(0)\|_n \leq \delta_n$. From here, from the control inequalities (2.6) (2.7) and from the Caplygin comparison lemma one infers that $\|u(t) - u_k(t)\|_n \leq R_n(t)$ for $t \in [0, T) \cap [0, T_\nu)$. Finally, it is $T \geq T_\nu$; in fact, it were $T < T_\nu$, the previous inequality about $u, u_k$ and $R_n$ would imply $\limsup_{t \to T} \|u(t)\|_n < +\infty$, a fact contradicting the maximality assumption for $u$. See [6] for more details. \qed

Paper [6] presents some applications of the previous proposition, dealing with both the Euler case $\nu = 0$ and the NS case $\nu > 0$; a special attention is devoted therein to the approximate solutions $u_k$ provided by the Galerkin method.

In this work we present an application of Proposition 2.1 to the Euler case $\nu = 0$, choosing for $u_k$ a polynomial in the time variable $t$. In the next section we present this procedure in general, giving the error estimators for approximate solutions of this kind; in the last section we apply the procedure choosing for $u_0$ the so-called Behr-Necs-Wu initial datum.

3. Polynomial approximate solutions for the Euler equations. Let us recall that $n \in (d/2 + 1, +\infty)$, and consider the Euler Cauchy problem with a datum $u_0 \in H^{n+1}_\Sigma_0$ and zero external forcing:

Find $u \in C([0,T), H^{n+1}_\Sigma_0) \cap C^1([0,T), H^n_\Sigma_0)$ such that

$$\frac{du}{dt} = P(u,u), \quad u(0) = u_0. \quad (3.1)$$

Let us choose an order $N \in \{0, 1, 2, \ldots\}$ and consider as an approximate solution for (3.1) a polynomial of degree $N$ in time, of the form

$$u^N : [0, +\infty) \to H^{n+1}_\Sigma_0, \quad t \mapsto u^N(t) := \sum_{j=0}^N u_j t^j \quad (u_j \in H^{n+1}_\Sigma_0 \text{ for all } j). \quad (3.2)$$

Here $u_0$ is the initial datum, and $u_j$ is to be determined for $j = 1, \ldots, N$.

Proposition 3.1. (i) Let $u^N$ be as in (3.2). The datum and differential errors of $u^N$ are

$$u^N(0) - u_0 = 0; \quad (3.3)$$

$$\frac{du^N}{dt}(t) - P(u^N, u^N)(t) = \sum_{j=0}^{N-1} \left[(j+1)u_{j+1} - \sum_{\ell=0}^j P(u_{\ell}, u_{j-\ell})\right]t^j - \sum_{\ell=0}^{2N} \sum_{j=0}^N \left[\sum_{\ell=j-N}^N P(u_{\ell}, u_{j-\ell})\right]t^j. \quad (3.4)$$

(ii) In particular, assume

$$u_{j+1} = \frac{1}{j+1} \sum_{\ell=0}^j P(u_{\ell}, u_{j-\ell}) \quad \text{for } j = 0, \ldots, N-1; \quad (3.5)$$
\[
\frac{du^N}{dt}(t) - \mathcal{P}(u^N, u^N)(t) = - \sum_{j=N}^{2N} \left[ \sum_{\ell=j-N}^{N} \mathcal{P}(u_\ell, u_{j-\ell}) \right] t^j = O(t^N) \quad \text{for } t \to 0 . \tag{3.6}
\]

(iii) If (3.5) is used to define recursively \( u_1, \ldots, u_N \), it produces a sequence of elements of \( \mathbb{H}^{n+1} \) under the condition \( u_0 = H^{n+1} \). More precisely, from \( u_0 \in \mathbb{H}^{n+1+N} \) it follows \( u_j \in \mathbb{H}^{n+1+N-j} \subset \mathbb{H}^{n+1} \) for \( j = 1, \ldots, N \).

(iv) Let \( u_0 \in \mathbb{H}^{n+1+N} \) and use (3.5) to define \( u_j \) for \( j = 1, \ldots, N \). Then

\[
\| \frac{du^N}{dt}(t) - \mathcal{P}(u^N, u^N)(t) \|_n \leq \epsilon_n(t) \quad \text{for } t \in [0, +\infty) , \tag{3.7}
\]

\[
\epsilon_n(t) := K_n \sum_{j=N}^{2N} \left[ \sum_{\ell=1}^{N} \| u_\ell \|_n \| u_{j-\ell} \|_{n+1} \right] t^j \quad \text{for } t \in [0, +\infty) . \tag{3.8}
\]

Proof. (i) (3.3) is obvious; let us prove (3.4). To this purpose, we note that

\[
\frac{du^N}{dt} - \mathcal{P}(u^N, u^N) = \frac{d}{dt} \left( \sum_{\ell=0}^{N} u_\ell t^\ell \right) - \mathcal{P}(\sum_{\ell=0}^{N} u_\ell t^\ell, \sum_{h=0}^{N} u_h t^h) = \sum_{\ell=1}^{N} \ell u_\ell t^{\ell-1} - \sum_{\ell=0}^{N} \mathcal{P}(u_\ell, u_h) t^\ell = \sum_{j=0}^{N-1} (j+1) u_{j+1} t^j - \sum_{j=0}^{N} \left[ \sum_{(\ell, h) \in I_{Nj}} \mathcal{P}(u_\ell, u_h) \right] t^j ,
\]

\[I_{Nj} := \{ (\ell, h) \in \{0, \ldots, N\}^2 \mid \ell + h = j \} .\]

One easily checks that

\[j \in \{0, \ldots, N-1\} \Rightarrow I_{Nj} = \{ (\ell, j-\ell) \mid \ell \in \{0, \ldots, j\} \} , \]

\[j \in \{N, \ldots, 2N\} \Rightarrow I_{Nj} = \{ (\ell, j-\ell) \mid \ell \in \{j-N, \ldots, N\} \} ;\]

this readily yields the thesis (3.4).

(ii) Obvious.

(iii) Let \( u_0 \in \mathbb{H}^{n+1+N} \) and define \( u_1, \ldots, u_N \) via the recursion relation (3.5). Then \( u_1 = \mathcal{P}(u_0, u_0) \in \mathbb{H}^{n+1+N} \), \( u_2 = (1/2)\mathcal{P}(u_0, u_1) + (1/2)\mathcal{P}(u_1, u_0) \in \mathbb{H}^{n+1+N-1} \), etc. .

(iv) Eq. (3.6) implies \( \| (du^N/dt)(t) - \mathcal{P}(u^N, u^N)(t) \|_n \leq \sum_{j=N}^{2N} \left[ \sum_{\ell=j-N}^{N} \| \mathcal{P}(u_\ell, u_{j-\ell}) \|_n \right] t^j .\) On the other hand Eq. (1.4) gives \( \| \mathcal{P}(u_\ell, u_{j-\ell}) \|_n \leq K_n \| u_\ell \|_n \| u_{j-\ell} \|_{n+1} ,\) whence the thesis (3.7) (3.8).

4. A special case of the previous framework: The Euler equations on \( T^3 \), with the Behr-Necas-Wu initial datum. In this section we consider the Euler Cauchy problem (3.1) with space dimension and Sobolev order

\[d = 3 , \quad n = 3 ; \tag{4.1}\]

the initial datum is

\[u_0 := \sum_{k=\pm a, \pm b, \pm c} u_{0k} e_k , \tag{4.2}\]

\[a := (1, 1, 0) , \quad b := (1, 0, 1) , \quad c := (0, 1, 1) ; \]

\[u_{0,\pm a} := (2\pi)^{3/2}(1, -1, 0) , \quad u_{0,\pm b} := (2\pi)^{3/2}(1, 0, -1) , \quad u_{0,\pm c} := (2\pi)^{3/2}(0, 1, -1) \]

(of course, being a Fourier polynomial, \( u_0 \) belongs to \( \mathbb{H}_m^{n+1+N} \) for each \( m \in \mathbb{R} \)). The above initial datum is considered by Behr, Necas and Wu in [1]; it is analyzed with a similar attitude in [9] (and, from a different viewpoint, in [6]). In both papers [1] [9], attention is fixed on the function \( u^N(t) = \sum_{j=0}^{N} u_j t^j \) for a rather large
value of $N$, where the $u_j$’s are determined for $j = 1, \ldots, N$ by the recursion relation (3.5). The $u_j$’s are Fourier polynomials and can be calculated exactly by computer algebra methods; such computations are performed in [1] for $N = 35$, and in [9] up to $N = 52$ (using, respectively, the $C^{++}$ and the Python languages).

The Python program of [9] gives exact expressions for the $u_j$’s, whose Fourier components are rational (up to factors $(2\pi)^{3/2}$); for large $j$, these expressions are terribly complicated. Here, to give a partial illustration of such Python computations we consider the Fourier components $u_k^{52}(t)$ for $k = (1,1,0)$ and $k = (0,0,2)$, and report the graphs of the functions $t \mapsto |u_k^{52}(t)|$ for these wave vectors: see Figures 1 and 2.

In both papers [1] [9], computations are used to get hints about $\lim_{N \rightarrow +\infty} u^N$, giving the exact solution of the Euler Cauchy problem on the time interval where the limit exists; however the statements of [1] [9] rely on the assumption that certain facts on the $N \rightarrow +\infty$ limit can be extrapolated from $u^{35}$ or $u^{52}$. In particular [1] makes the conjecture, disputed in [9], that the solution of the Euler Cauchy problem blows up for $t \rightarrow \tau^-$, with $\tau \approx 0.32$.

In the present work we make no conjecture or extrapolation about the $N \rightarrow +\infty$ limit and just consider the function $u^{52}$ of [9] according to the general framework of approximate solutions and control inequalities. This approach produces:

(i) a rigorous lower bound on the interval of existence of the exact solution $u$ of the $(d = 3, n = 3)$ Cauchy problem (3.1);

(ii) a bound on $\|u(t) - u^{52}(t)\|_3$.

To get these results we regard $u^{52}$ as an approximate solution of (3.1), using the tautological datum error and growth estimators

$$\delta_3 := 0 ; \quad D_3(t) := \|u^{52}(t)\|_3, \quad D_4(t) := \|u^{52}(t)\|_4 \quad \text{for} \ t \in [0, +\infty) \quad (4.3)$$

(concerning $\delta_3$, we recall that $u^{52}(0) - u_0 = 0$). For $m = 3, 4$ one has $D_m(t) = (2\pi)^{3/2} \sum_{j=0}^{52} d_{mj} t^{2j}/2$ where the $d_{mj}$ are rational coefficients; the Python program employed for our work [9] computes exactly these coefficients. For $m = 3$ these coefficients are reported in [9], in a 16-digits decimal representation (see Eq. (5.12) of [9], not containing the factor $(2\pi)^{3/2}$ due to a different normalization of the norm $\| \cdot \|_3$); we have no room to report here the coefficients of the $m = 4$ case. Figures 3 and 4 contain the graphs of the functions $t \mapsto D_3(t), D_4(t)$.

Let us pass to the differential error estimator for $u^{52}$; we use for it the function $\epsilon_3$ defined by (3.8) with $n = 3$ and $K_3 = 0.323$, see (1.6). $\epsilon_3$ is computed exactly by our Python program; again, the explicit expression is too complicated to be reported. (The tautological error estimator $\epsilon_3(t) := \| (du^{52}/dt)(t) - P(u^{52}, u^{52})(t) \|_3 = \| \sum_{j=52}^{104} \sum_{l=j-52}^{52} P(u_t, u_{j-l}) t^l \|_3$ is more accurate, but it has an even more complicated expression; its calculation by computer algebra is too expensive.)

For the graph of $\epsilon_3$ and some information on its numerical values, see Figure 5 and its caption. With the previous ingredients, we build the following “control Cauchy problem”: find $\mathcal{R}_3$ such that

$$\mathcal{R}_3 \in C^3([0, T_2], \mathbb{R}), \quad \frac{d\mathcal{R}_3}{dt} = (G_3 D_3 + K_3 D_4) \mathcal{R}_3 + G_3 \mathcal{R}_3^2 + \epsilon_3, \quad \mathcal{R}_3(0) = 0 \quad (4.4)$$

($G_3 = 0.438$, see again (1.6)). This control problem has a unique maximal solution $\mathcal{R}_3$, which is strictly increasing and thus positive for $t \in (0, T_2)$. Of course, this $\mathcal{R}_3$ fulfils as equalities Eqs. (2.6) (2.7) (with $\nu = 0$).
Once we have $R_3 : [0, T_c) \rightarrow [0, +\infty)$, due to Proposition 2.1 we can grant that:

(i) The maximal solution $u$ of the ($n = 3$) Euler Cauchy problem (3.1) is defined on an interval $[0, T)$ with $T \geq T_c$;

(ii) It is

$$
\|u(t) - u^{52}(t)\|_3 \leq R_3(t) \quad \text{for} \quad t \in [0, T_c).
$$

The function $R_3$ can be determined numerically by a cheap computation using any package for ODEs, e.g. Mathematica (the result is reliable, since (4.4) is the Cauchy
Problem for a simple ODE in one dimension). This numerical computation indicates that the (maximal) domain of $\mathcal{R}_3$ is $[0, T_c)$, with

$$T_c = 0.242\ldots; \quad (4.6)$$

After having been extremely small for most of the time between 0 and $T_c$, $\mathcal{R}_3(t)$ diverges abruptly for $t \to T_c^-$; for the graph of this function and some information on its numerical values, see Figure 6 and its caption. Due to (4.6), we can grant that the solution $u$ of the Euler Cauchy problem (1.11) exists on a time interval of length $T \geq 0.242$ (this is four times larger than the lower bound on $T$ obtained in [6] using a Galerkin approximate solution).

Eq. (4.5) and the previously described behavior of $\mathcal{R}_3$ ensure that $u^{52}(t)$ approximates with extreme precision $u(t)$ on most of the time interval $[0, T_c)$. We remark
Figure 6. Plot of $R_3(t)$ for $t \in [0.20, 0.24]$. One has: $R_3(t) < 2 \times 10^{-6}$ for $t \in [0, 0.20]$; $R_3(t) < 1.2 \times 10^{-4}$ for $t \in (0.20, 0.21]$; $R_3(t) < 0.013$ for $t \in (0.21, 0.22]$; $R_3(t) < 2$ for $t \in (0.22, 0.23]$; $R_3(t) < 610$ for $t \in (0.23, 0.24]$.

that (4.5) can be used to infer other interesting estimates about $u - u^{52}$, e.g.,

$$\left| u_k(t) - u^{52}_k(t) \right| \leq \frac{R_3(t)}{|k|^3} \quad \text{for } k \in \mathbb{Z}^3 \setminus \{0\}, \ t \in [0, T_c) \ ;$$

(4.7)

this follows from (4.5) and from the elementary inequality $|v_k| \leq \|v\|_3/|k|^3$, holding for all $v \in H^3_{\Sigma_0}$ and $k \in \mathbb{Z}^3 \setminus \{0\}$ (recall that $\|v\|_3^2 = \sum_{k \in \mathbb{Z}^3 \setminus \{0\}} |k|^6 |v_k|^2$).

Acknowledgments. This work has been partially supported by by INdAM, INFN and by MIUR, PRIN 2010 Research Project “Geometrical methods in the theory of nonlinear waves and applications”.

REFERENCES


E-mail address: carlo.morosi@polimi.it
E-mail address: mario.pernici@infn.it
E-mail address: livio.pizzocchero@unimi.it
ON THE STABILITY OF DEGENERATE VISCOUS SHOCK PROFILES

Ramón G. Plaza
Departamento de Matemáticas y Mecánica
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
Apdo. Postal 20-726 C.P. 01000 México D.F. Mexico

Abstract. In [16], we showed how to obtain $L^p$-decay rates for zero-mass perturbations of degenerate scalar viscous shock waves using energy methods. The proof is based upon previous work by Matsumura and Nishihara [12], by extending their weighted energy estimates to $L^p$-norms, and by obtaining sharp decay rates with the aid of basic interpolation inequalities. This contribution summarizes the results in [16] and illustrates the use of interpolation inequalities to obtain decay rates by analyzing the $L^2$ case as an example.

1. Introduction. Consider a scalar viscous conservation law,
   \[ u_t + f(u)_x = (b(u)u_x)_x, \tag{1} \]
where $f, b \in C^2$, $b > 0$, and $(x, t) \in \mathbb{R} \times [0, +\infty)$. This paper pertains to the stability of viscous shock wave solutions to equation (1) of form $u(x, t) = \bar{u}(x - st)$, where $\bar{u}$ satisfies
   \[ (b(\bar{u})\bar{u}')' = f(\bar{u}') - s\bar{u}', \]
   \[ \bar{u}(\xi) \to u_\pm, \quad \text{as} \quad \xi \to \pm\infty. \]
Here, $' = d/d\xi$ denotes differentiation with respect to $\xi := x - st$, and $s$ is the shock speed. We assume that the triple $(u_+, u_-, s) \in \mathbb{R}^3$, with $u_- \neq u_+$ (say $u_+ < u_-$), is a generalized shock front [10] satisfying the classical Rankine-Hugoniot jump condition and the generalized entropy condition (cf. [11, 17]). We pay particular attention to the case where the flux function $u \mapsto f(u)$ changes its convexity in $u \in (u_+, u_-)$. Such a hypothesis allows us to consider sonic (or degenerate) shocks, namely, waves whose speed matches one of the characteristic speeds with $s = f'(u_+)$ or $s = f'(u_-)$.

The stability of degenerate viscous shocks has been addressed by many authors. One of the first approaches consisted on obtaining a priori energy estimates for the perturbations (cf. [13, 14, 15, 12]). This approach is usually called the energy method. In particular, the seminal paper of Matsumura and Nishihara [12] introduced a suitable, asymmetric weight function, not bounded on the sonic side, which accomodates properly on the compressive side, yielding the right sign in the weighted energy estimates. Another approach, developed a few years later by Howard [8, 7] (and which generalizes the method introduced by Howard [6], and by Howard and Zumbrun [18] for classical shocks), follows pointwise bounds on the
Green’s function for the linearized operator around the wave. These techniques offered more information on the asymptotic behavior of the solutions than energy methods, yielding sharper decay rates in all $L^p$ spaces (see [8] for details). The application of the pointwise Green’s function method to degenerate shocks is, however, more complicated than in the classical case, due to the fact that degenerate profiles decay algebraically, in contrast to the exponential decay of Laxian shocks. This makes the Evans function non-analytic near profiles decay algebraically, in contrast to the exponential decay of Laxian shocks.

One arguable drawback of the energy approach has been its inability to provide sharp estimates on the rate at which perturbations decay. In [16], however, we show that a sharp rate of decay can be obtained by energy methods very similar to those applied by Mei [13, 14], Matsumura-Nishihara [12], and Nishikawa [15]. Here we summarize the results and illustrate the use of interpolation inequalities to get sharp rates of decay by examining the $L^2$ case (see §3). For further details the reader is referred to [16].

2. Preliminaries and main result. By translation invariance, we normalize the flux function $f$ such that the profile is stationary (i.e. $s = 0$), obeying the equation

$$b(\bar{u})u_x = f(\bar{u}).$$

(2)

Therefore we summarize our assumptions as follows:

$$f, b \in C^2, b > 0,$$ (regularity and positive diffusion), \hspace{0.5cm} (A1)

$$f(u_-) = f(u_+) = 0,$$ (Rankine-Hugoniot condition), \hspace{0.5cm} (A2)

$$f(u) < 0 \text{ for all } u \in (u_+, u_-),$$ (generalized entropy condition). \hspace{0.5cm} (A3)

The generalized entropy condition (A3) implies the non-strict condition $f'(u_+) \leq 0 \leq f'(u_-)$, which allows sonic waves with $f'(u_+) = 0$ or $f'(u_-) = 0$. For concreteness, we assume that the shock is sonic on the positive side, namely, that $f'(u_+) = 0$. Whence, we rewrite assumption (A3) as

$$f(u) < 0, \text{ for all } u \in (u_+, u_-), \text{ and, } f'(u_+) = 0 < f'(u_-).$$ (A3')

2.1. Existence and structure of profiles. The following existence result is well-known (cf. [17, 12, 2]; for a proof, see Howard [8], Section 2).

Proposition 2.1. Under hypotheses (A1) - (A3'), let us define $\theta := \min\{k \in \mathbb{Z}^+ : (d^k f/du^k)(u_+) \neq 0\} \geq 1$, as the degree of degeneracy of the shock. Then there exists a traveling wave solution $\bar{u}$ of (2) with $\bar{u}(\pm \infty) = u_\pm$, unique up to translations. Moreover, $\bar{u}$ is monotone decreasing, $\bar{u}_x < 0$, and $\bar{u}$ and its derivatives decay like

$$|\partial_x^j(\bar{u}(x) - u_-)| \leq C e^{-\theta|x|}, \quad \text{as } x \to -\infty,$$

$$|\partial_x^j(\bar{u}(x) - u_+)| \leq C|x|^{-1/\theta}, \quad \text{as } x \to +\infty,$$

for $j = 0, 1, 2$, and some uniform $C > 0$.

2.2. Perturbation equations. We restrict our analysis to the class of perturbations with zero-mass, and choose $\delta$ such that $\int_\mathbb{R}(u_0(x) - \bar{u}(x - \delta)) \, dx = 0$. This allows us to write the perturbation as $u(x,t) - \bar{u}(x - \delta) = v_\delta(x, t)$, for some function $v_\delta(\cdot, t)$ in $L^2$, i.e., we can integrate the equation [3, 5]. Suppose, without loss of generality, that $\delta = 0$; this yields

$$\int_\mathbb{R}(u_0(x) - \bar{u}(x)) \, dx = 0.$$ (3)
In view of last observations, substitute now \( u(x, t) = v_x(x, t) + \bar{u}(x) \) into (1), integrate in \((\infty, x)\) and use the profile equation (2) to obtain the perturbation equation
\[
v_t = b(\bar{u})v_{xx} - a(x)v_x + F,
\]
where \( a(x) = f'(\bar{u}) - b(\bar{u})x \), and
\[
F = -(f(v_x + \bar{u}) - f'(\bar{u})v_x) + (b(v_x + \bar{u}) - b(\bar{u}) - b'(\bar{u})v_x)(\bar{u}_x + v_{xx}) + b'(\bar{u})v_x v_{xx},
\]
comprises the nonlinear terms. Therefore, after these reformulations, the Cauchy problem for the perturbation \( v \) is written as follows,
\[
v_t = b(\bar{u})v_{xx} - a(x)v_x + F, \quad \text{for } (x, t) \in \mathbb{R} \times (0, +\infty), \quad (4)
\]
\[
v(x, 0) = v_0(x) = \int_{-\infty}^{x} (u_0(y) - \bar{u}(y)) \, dy, \quad \text{for } x \in \mathbb{R}. \quad (5)
\]

### 2.3. The Matsumura-Nishihara weight function.
Matsumura and Nishihara [12] introduced the following weight function,
\[
\eta(x) = \bar{\eta}(\bar{u}(x)), \quad \bar{\eta}(u) = \frac{(u - u_+)(u - u_-)}{f(u)} > 0, \quad u \in (u_+, u_-). \quad (6)
\]
which behaves like
\[
\eta \sim \langle x \rangle_+ = \begin{cases} 
(1 + x^2)^{1/2}, & x \geq 0, \\
1, & x < 0.
\end{cases}
\]
(See [12] for details.) In particular we have that \( \eta \) is bounded below, \( \eta \geq C^{-1} > 0 \), for all \( x \), and with some uniform \( C > 0 \). Note, however, that it is not bounded above as it blows up on the sonic side when \( x \to +\infty \). The remarkable property of the weight function (6) is that it leads to the right sign of the term
\[
\Phi(x) = ((b(\bar{u})\eta)_x + a(x)\eta)_x = ((d/du)(f(u)\eta(u))|_{u=\bar{u}})_x = -2|\bar{u}_x| < 0,
\]
for all \( x \in \mathbb{R} \) (see [12]). It can also be shown that \( |\eta_x| \leq C\eta \) for some uniform \( C > 0 \).

Finally, let us specify some notation. \( W^{m,p} \) denotes the standard Sobolev spaces in \( \mathbb{R} \). In terms of the weight function, for each \( 1 \leq p < +\infty \), \( L^p_\eta \) denotes the space of measurable functions \( u \) such that \( \eta^{1/p}u \in L^p \), or that \( ||u||_{L^p_\eta}^p := \int_{\mathbb{R}} \eta |u|^p \, dx < +\infty \).

### 2.4. Main result.
We are ready to state the main theorem in [16].

**Theorem 2.2** ([16]). Under assumptions (A1) - (A3'), with \( u_+ < u_- \), let \( \bar{u} \) be the traveling wave solution to (1) of Proposition 2.1. Suppose that the zero-mass condition (3) holds, and that
\[
v_0 := \int_{-\infty}^{x} (u_0(x) - \bar{u}(x)) \, dx \in Z_{\eta,p},
\]
where \( Z_{\eta,p} = L^1_\eta \cap L^2_\eta \cap L^p_\eta \cap W^{2,p} \), for some \( 2 \leq p < +\infty \), and \( \eta \) denotes the Matsumura-Nishihara weight function (6). Then there exists a positive constant \( \hat{\epsilon} > 0 \) such that if \( ||v_0||_{Z_{\eta,p}} < \hat{\epsilon} \), then the Cauchy problem for equation (1) with initial condition \( u(0) = u_0 \) has a unique global solution \( u - \bar{u} \in C([0, +\infty]; W^{1, p}) \) satisfying
\[
||u - \bar{u}||_{L^p} \leq CME_0t^{-1/2}(1 + t)^{-\frac{1}{2}(1-1/p)}, \quad (7)
\]
\[
||u - \bar{u}||_{L^\infty} \leq CME_0t^{-1/2-1/2p}(1 + t)^{-\frac{1}{2}(1-1/p)}, \quad (8)
\]
for all \( 0 < t < +\infty \), where \( E_0 = ||v_0||_{L^1_\eta} + ||v_0||_{L^2_\eta} + ||v_0||_{L^p_\eta}^2 \), and with uniform constant \( M > 0 \).
Like that of Matsumura and Nishihara [12], this result applies to zero-mass perturbations only and require very rapidly decaying data, as \( u_0 \) must belong to the weighted space \( Z_{\eta,p} \). Observe, however, that the decay rates (7) are sharp. It is also to be noted that, unlike the Green’s function method, the analysis works for all degrees of degeneracy.

3. Decay rates. In this section we show how to use interpolation inequalities to obtain sharp rates of decay. The following inequality is the weighted norm version of the interpolation inequality by Escobedo and Zuazua [1] (Lemma 1, pg. 129). The proof in [16] follows [1] closely, with the appropriate adaptations to the weighted spaces under consideration. Notably, the original inequality remains valid in weighted spaces, even though the function \( \eta \) is not bounded above on the sonic side.

**Lemma 3.1** (Weighted interpolation inequality [1, 16]). For each \( 2 \leq p < +\infty \) there exists some constant \( C = C(p) > 0 \) such that
\[
\|u\|_{L^2_\eta}^{p/(p-1)} \leq C\|u\|_{L^2_\eta}^{2p/(p-1)}\|(|u|^{p/2})_x\|_{L^2_\eta}^{2},\tag{9}
\]
for every \( u \in W^{2,p} \cap L^1_\eta, \) with \( u_x \in W^{2,p} \cap L^1_\eta \), where \( \eta \) denotes the Matsumura-Nishihara weight function.

We illustrate how to obtain this inequality in the \( L^2 \) (non-weighted) case (for the general proof, see [16]). By the Sobolev inequality \( \|u\|_{L^2}^2 \leq 2\|u\|_{L^2} \|u_x\|_{L^2} \), we readily get
\[
\|u\|_{L^2}^2 = \left( \int u^2 dx \right)^{\frac{2}{p}} \leq \|u\|_{L^\infty}^4 \|u\|_{L^2} \leq 4\|u\|_{L^2} \|u_x\|_{L^2} \|u\|_{L^4}^2.
\]
This immediately yields \( \|u\|_{L^2}^6 \leq C\|u_x\|_{L^2}^2 \|u\|_{L^4}^4 \) which is the inequality (9) for \( p = 2 \).

For simplicity, let us explain to obtain sharp decay rates in \( L^2 \). We begin with an observation. Let \( \rho(t) \geq 0 \) be of class \( C^1 \) in \( t > 0 \), such that \( \frac{d\rho}{dt} \leq C \rho^\beta \) for some \( C > 0, \beta > 1 \), and with \( \rho(0) = \rho_0 > 0 \). Then clearly \( \rho(t) \leq \xi(t), \) a.e. in \( t > 0 \), where \( \xi = \xi(t) \) is the solution to \( \frac{d\xi}{dt} = -C\xi^\beta \), and \( \xi(0) = \rho_0 \). Now suppose that for \( 0 < t < T \leq +\infty \), and some \( C_0, C_1 > 0 \), there hold
\[
\frac{d}{dt}\|u(t)\|_{L^2}^2 \leq -C_1\|u_x\|_{L^2}^2, \quad \text{and} \quad \|u(t)\|_{L^2} \leq C_0.
\]
Use the interpolation inequality to get
\[
\frac{d}{dt}\|u\|_{L^2}^2 \leq -C_1\|u_x\|_{L^2}^2 \leq -C_1\|u\|_{L^2}^2 \|u\|_{L^2}^{-2} \leq -C_2\|u\|_{L^2}^6,
\]
where \( C_2 = C_1/(CC_0^3) \). It is easy to verify that \( \xi = (2C_2 t + \|u(0)\|_{L^2}^{-1})^{-1/2} \) is the solution to \( \frac{d\xi}{dt} = -C_2\xi^3 \), \( \xi(0) = \|u(0)\|_{L^2}^2 \) (here \( \beta = 3 \)). Hence, we clearly have \( \xi(t)^{-2} \geq C_3(1 + t), \) with \( C_3 = \min\{2C_2,\|u(0)\|_{L^2}^{-4}\} \), and by the observation above we obtain
\[
\|u(t)\|_{L^2}^2 \leq \xi(t)^2 \leq C_3^{-1/2}(1 + t)^{-1/2} \leq \tilde{C}(C_0 + \|u(0)\|_{L^2})^2(1 + t)^{-1/2}.
\]
Therefore we have proved the following decay rate:
\[
\|u(t)\|_{L^2} \leq C(1 + t)^{-1/4},
\]
where the constant is $C = O(C_0 + \|u(0)\|_{L^2})$. In a similar fashion, and using the weighted interpolation inequality (9), it is possible to prove the following result (see [16] for details):

**Lemma 3.2.** Let $2 \leq p < +\infty$, and suppose that $u$ is the solution to a certain evolution (linear or nonlinear) equation, which satisfies the bounds

\[
\frac{d}{dt} \|u(t)\|_{L^p}^p \leq -C_1 \|(|u|^{p/2})_x(t)\|_{L^2}^2, \quad (10)
\]

\[
\|u(t)\|_{L^1} \leq C_0, \quad (11)
\]

for all $0 < t < T \leq +\infty$, and uniform constants $C_1, C_0 > 0$. Then, there exists a positive constant $\tilde{C} > 0$ such that

\[
\|u(t)\|_{L^p} \leq \tilde{C}(1 + t)^{-\frac{1}{2}(1-1/p)},
\]

for all $0 < t < T$. Moreover, the constant $\tilde{C}$ is of order $\tilde{C} = O(C_0 + \|u(0)\|_{L^p})$.

4. **Energy estimates.** According to custom, the global existence and the stability are proved by a continuation argument based on a local existence result combined with the corresponding a priori energy estimates. Assume $2 \leq p < +\infty$ is fixed. Let us define the suitable space for solutions as $Z_{\eta,p} = W^{2,p} \cap L^1 \cap L^2$, and $X_{\eta,p}(0,T) = \{ v \in C([0,T];Z_{\eta,p}), v_x \in L^2([0,T];Z_{\eta,p}) \}$ with $0 < T \leq +\infty$. Using the variation of constants formula and by a standard contraction mapping argument it is possible to prove the following short-time existence result.

**Proposition 4.1** (Local existence). For any $\epsilon_0 > 0$ there exists a positive constant $T_0$ depending on $\epsilon_0$ such that if $v_0 \in Z_{\eta,p}$ and $\|v_0\|_{Z_{\eta,p}} \leq \epsilon_0$, then the Cauchy problem (4) and (5) has a unique solution $v \in X_{\eta,p}(0,T_0)$ satisfying $\|v(t)\|_{Z_{\eta,p}} < 2\epsilon_0$ for each $0 \leq t \leq T_0$.

4.1. **The basic energy estimate.** Here we describe the energy estimates for the perturbations obtained in [16]. The proofs follow standard energy estimates in $L^p$ spaces.

**Lemma 4.2** (Basic energy estimate [16]). Let $2 \leq p < +\infty$, and let $v(t) \in X_{\eta,p}(0,T)$ be a solution to (4) for some $T > 0$. Then, for each $0 \leq t \leq T$ there holds

\[
\frac{1}{p} \|v(t)\|_{L^p}^p + \frac{4(p-1)}{p^2} \int_0^t \|b(\bar{u})^{1/2}(|v(\tau)|^{p/2})_x\|_{L^2}^2 \, d\tau + \frac{2}{p} \int_\mathbb{R} |\bar{u}_x| |v(\tau)|^p \, dx \, d\tau =
\]

\[
= \frac{1}{p} \|v(0)\|_{L^p}^p + \int_0^t \int_\mathbb{R} \eta F v(\tau) |v(\tau)|^{p-2} \, dx \, d\tau.
\]

Let us now define $R(t) := \sup_{\tau \in [0,t]} \|v(\tau)\|_{Z_{\eta,p}}$ for each $t \in [0,T]$, with $T > 0$ fixed.

**Lemma 4.3** ([16]). There exists $\epsilon_1 > 0$ sufficiently small such that if $R(t) < \epsilon_1$ for $t \in [0,T]$ then we have the estimate

\[
\|v(t)\|_{L^p}^p + \hat{C}_1 \int_0^t \|(|v(\tau)|^{p/2})_x\|_{L^2}^2 \, d\tau \leq \|v(0)\|_{L^p}^p, \quad (12)
\]

for some $\hat{C}_1 > 0$ depending on $p$ and $\epsilon_1$, and for all $0 \leq t \leq T$.

Two immediate corollaries follow.
Corollary 4.4. Specializing (12) to the case $p = 2$ we have that if $R(t) < \epsilon_1$ for all $0 \leq t \leq T$ then

$$\|v(t)\|_{L^2}^2 + \dot{C}_1 \int_0^t \|v_x(\tau)\|_{L^2}^2 \, d\tau \leq \|v(0)\|_{L^2}^2.$$  

Corollary 4.5. If $v \in X_{\eta,p}(0,T)$ is a solution with $R(t) \leq \epsilon_1$ for $0 \leq t \leq T$, then

$$\frac{d}{dt}\|v(t)\|_{L^p}^p + C_1 \|v\|_{L^p}^2 \leq 0,$$

for some $C_1 > 0$.

4.2. $L^1_\eta$-bound and decay rates. The boundedness of $\|v(t)\|_{L^1_\eta}$, plays a key role in Lemma 3.2. This a remarkable property of the solutions to (4), and of the Matsumura-Nishihara weight function. (See [16] for details.)

Lemma 4.6. Assuming $R(t) < \epsilon_1$, with $\epsilon_1 > 0$ just as in Lemma 4.3, the following estimate holds

$$\|v(t)\|_{L^1_\eta} \leq C(\|v(0)\|_{L^1_\eta} + \|v(0)\|_{L^2}^2),$$

for some $C > 0$, all $0 \leq t \leq T$.

The previous observations and the decay rates of the previous section readily imply the following

Corollary 4.7. Let $v \in X_{\eta,p}(0,T)$ be a solution to (4), for some $T > 0$. If $R(t) < \epsilon_1$ for all $0 \leq t \leq T$, then $v$ satisfies the decay rate

$$\|v(t)\|_{L^p}^p \leq CE^p_0(1 + t)^{-(p-1)/2},$$

with $E_0 := \|v(0)\|_{L^1_\eta} + \|v(0)\|_{L^2}^2$.

Proof. Since $R(t) < \epsilon_1$, we may apply Lemma 4.6 and Corollary 4.5 to conclude that properties (10) and (11) hold with $C_0 = O(\|v(0)\|_{L^1_\eta} + \|v(0)\|_{L^2}^2)$. Then, by Lemma 3.2 we obtain the desired decay rate (13).

4.3. Higher order estimates. The $L^p$ estimates for the derivatives cannot be controlled as in the $L^2$ case (where $\|\|v\|^{p/2}\|_r$ is equivalent to $\|u_x\|$ and there is a natural way to construct a decreasing norm). Thus, we follow the general method of [9] instead. (For details, the reader is referred to [16].)

Lemma 4.8 (Higher order estimates [16]). Suppose $v \in X_{\eta,p}(0,T)$, $2 \leq p < +\infty$, with $0 < T \leq 1$ solves (4). Then there exists $\epsilon_2 > 0$, sufficiently small, such that if $R(t) < \epsilon_2$ for $0 \leq t \leq T \leq 1$, then there hold the estimates

$$\frac{1}{p} t^\alpha (1 + t)^\beta \|v_x(t)\|_{L^p}^p + \dot{C}_2 \int_0^t \tau^\alpha (1 + \tau)^\beta \|v_x(\tau)\|_{L^2}^2 \, d\tau \leq CE_0^p t^{\alpha - p/2}(1 + t)^{\beta - \frac{1}{2}(p - 1)},$$

$$\frac{1}{p} t^\gamma (1 + t)^\delta \|v_{xx}(t)\|_{L^p}^p + \dot{C}_3 \int_0^t \tau^\gamma (1 + \tau)^\delta \|v_{xx}(\tau)\|_{L^2}^2 \, d\tau \leq CE_0^p t^{\gamma - p}(1 + t)^{\delta - \frac{1}{2}(p - 1)},$$
for $0 \leq t \leq T$, where $\alpha, \beta > 0$ satisfy $\alpha > p/2$, $\beta > \frac{1}{2}(p-1)$; $\gamma, \delta > 0$ satisfy $\gamma > p$, $\delta > \frac{1}{2}(p-1)$, and $C_2, C_3 > 0$ are constants depending on $p, \epsilon_2, \alpha, \beta, \gamma$ and $\delta$. Moreover, for $0 < t < T$ there holds the decay rates
\begin{align}
\|v_x(t)\|_{L^p}^p &\leq CE_0^p t^{-p/2} (1 + t)^{-\frac{1}{2}(p-1)}, \\
\|v_{xx}(t)\|_{L^p}^p &\leq CE_0^p t^{-p} (1 + t)^{-\frac{1}{2}(p-1)}.
\end{align}

The additional assumption $T \leq 1$ means no loss of generality, as the local existence time in Proposition 4.1 can be chosen as $\hat{T}_0 = \min \{1, T_0(\epsilon_0)\}$.

5. Stability and proof of Theorem 2.2. We apply the previous a priori estimates to show stability and to prove the main Theorem. This is achieved by a continuation argument to obtain global existence of solutions to the Cauchy problem, paying special attention to the fact that estimates for the derivatives (14) and (15) apply only within time intervals of measure one. Thus, the proof slightly deviates from the standard argument. Notably, the energy $E_0$ does not involve norms of the derivatives, and this allows to extend the decay rates globally in time.

**Theorem 5.1** ([16]). Suppose $v_0 \in Z_{n,p}$, with $2 \leq p < +\infty$. Then there exists a positive constant $\hat{\epsilon} > 0$ such that if $\|v_0\|_{Z_{n,p}} < \hat{\epsilon}$, then the Cauchy problem (4) and (5) has a unique global solution $v \in X_{n,p}(0, +\infty)$ which satisfies the following estimates
\begin{align}
\|v(t)\|_{L^1} &\leq ME_0, \\
\|v(t)\|_{L^p} &\leq ME_0(1 + t)^{-\frac{1}{2}(1-1/p)}, \\
\|v(t)\|_{L^2} &\leq ME_0(1 + t)^{-\frac{1}{2}}, \\
\|v_x(t)\|_{L^p} &\leq ME_0 t^{-\frac{1}{2}} (1 + t)^{-\frac{1}{2}(1-1/p)}, \\
\|v_{xx}(t)\|_{L^p} &\leq ME_0 t^{-1} (1 + t)^{-\frac{1}{2}(1-1/p)},
\end{align}
for all $0 < t < +\infty$, with some uniform $M > 0$, and where $E_0 = \|v_0\|_{L^1} + \|v_0\|_{L^p} + \|v_0\|_{L^2}^2$.

The main Theorem 2.2 is now a direct consequence of Theorem 5.1.

**REFERENCES**


\textit{E-mail address: plaza@mym.iimas.unam.mx}
ASYMPTOTICALLY IMPLICIT SCHEMES FOR THE HYPERBOLIC HEAT EQUATION

GIACOMO DIMARCO
Institut de Mathématiques de Toulouse
Université Paul Sabatier, Toulouse, France

LORENZO PARESCHI AND VITTORIO RISPOLI
Department of Mathematics and Computer Science
University of Ferrara, Italy

Abstract. The main concern of this paper is to derive numerical schemes for the solution of kinetic equations with diffusive scaling which works efficiently for a wide range of the scaling parameter $\varepsilon$. We will concentrate on the simple Goldstain-Taylor model from kinetic theory and propose a resolution method based on the reformulation first introduced in [8]. We show how this reformulation corresponds to the use of interpolated fluxes and then we adopt the penalized implicit-explicit Runge-Kutta approach recently introduced in [1] to overcome the parabolic time step restriction in the diffusive regime. The resulting schemes permit to choose a time step $\Delta t = O(\Delta x)$, independent from $\varepsilon$, in all regimes. Some numerical examples show the efficiency and accuracy of the proposed methods.

1. Introduction. Kinetic theory of rarefied gases has been used for studies in various and very important fields of research and nowadays it still represents one of the most powerful tools for simulations and applications for cutting edge sectors of academic and industrial research.

The aim of this paper is to derive a numerical scheme for the solution of kinetic equations with diffusive scaling which works efficiently for a wide range of the scaling parameter $\varepsilon$. The Goldstain-Taylor model, despite its simplicity, is a prototype kinetic equation which contains some of the major difficulties encountered when dealing with more sophisticated kinetic models in the diffusive scaling. For this reason it is often used for the derivation and the analysis of robust numerical schemes for the diffusion limit [2, 4, 10, 11]. In particular, for small values of the scaling parameter, the kinetic model is well approximated by a standard heat equation for the mass density. This is the rationale behind the use of the terminology hyperbolic heat equation when considering the Goldstain-Taylor model in the diffusive scaling.

More in general when dealing with kinetic equations in the diffusive scaling, it is necessary to perform numerical simulations in both resolved (when spatial grid size is smaller than the mean free path) and under-resolved (when spatial grid size is larger than the mean free path) regimes. In the former case, standard numerical methods for kinetic equations work effectively. In the latter case, one expects that the best possible numerical solution is the approximation of the diffusion equation.

2000 Mathematics Subject Classification. Primary: 58F15, 58F17; Secondary: 53C35.
Key words and phrases. Asymptotically implicit schemes, hyperbolic heat equation, IMEX schemes, Goldstain-Taylor model, diffusive scaling.
However, even to achieve this goal is nontrivial for standard numerical methods when applied to kinetic equations. Earlier studies on numerical methods for transport or kinetic equations indicate that, in order for the under-resolved numerical approximation to capture the correct diffusive behavior, the scheme should be asymptotic preserving (AP), in the sense that the asymptotic limit that leads from the transport or kinetic equations to the diffusion equations should be preserved at the discrete level [6]. Moreover, an efficient method should be able to numerically go along with the physics of the system which goes from an hyperbolic behavior underlying the kinetic regime to a parabolic behavior in the limiting regime.

Here we described a general approach to tackle this kind of problems. For the construction of the method first we rely on the problem reformulation presented in [8] and show that the method can be seen as a natural way to construct interpolated fluxes for the transport part. It is well-known that this reformulation in combination with a splitting method or an Implicit-Explicit (IMEX) time discretization provides a consistent way to approximate the kinetic model in the different regimes avoiding the time constraints $\Delta t = O(\varepsilon \Delta x)$ for small values of $\varepsilon$.

In fact, since the characteristic speeds of the hyperbolic part are of order $1/\varepsilon$, standard approaches developed for hyperbolic systems with stiff relaxation become useless in such parabolic scaling, because the CFL condition would require $\Delta t = O(\varepsilon \Delta x)$. Of course, in the diffusive regime where $\varepsilon \ll \Delta x$, this is too much restrictive since also for an explicit method a parabolic condition $\Delta t = O(\Delta x^2)$ would suffice.

Most previous works on asymptotic preserving schemes for hyperbolic systems and kinetic equations with diffusive relaxation focus on schemes which, in the limit of infinite stiffness, become consistent explicit schemes for the diffusive limit equation [2, 4, 5, 8, 11]. Such explicit (in the limit) schemes clearly suffer from the usual parabolic stability restriction $\Delta t = O(\Delta x^2)$. Here, following the strategy presented in [1], we construct schemes that work uniformly in $\varepsilon$ and that, in the diffusion limit, originate a fully implicit solver for the diffusion equation. Therefore the resulting schemes permit to choose a time step $\Delta t = O(\Delta x)$ in all regimes.

The paper is organized as follows: in the next section we present the prototype system we are going to solve together with the proposed techniques used to implement our strategy. Later section is devoted to the presentation of system’s discretization and is followed by the section of the numerical results. A concluding section ends the paper.

2. The Goldstain-Taylor model. A two-velocity model of the Boltzmann equation describes the behavior of a fictitious gas of two kind of particles that move parallel to the $x-$axis with constant and equal speed. We can consider at time $t$ the particles with a density $f(x,t)$, which move in the positive $x-$direction, and the particles which move in the negative $x-$direction with a density $g(x,t)$. The simplest two-velocity gas, which is in local equilibrium when $f = g$, is described by the following hyperbolic system

$$\begin{align*}
  f_t + cf_x &= k(g - f), \\
  g_t - cg_x &= k(f - g),
\end{align*}$$

(1)

where $c > 0$ and $k > 0$ characterize respectively the velocities and the interactions of particles.
In the diffusive scaling we consider the system of equations (1) in the form
\[\begin{align*}
    f_t + \frac{c}{\epsilon} f_x &= \frac{k}{\epsilon^2} (g - f), \\
    g_t - \frac{c}{\epsilon} g_x &= \frac{k}{\epsilon^2} (f - g).
\end{align*}\] (2)

In (2) \(\epsilon\) is called the relaxation time and the limit problem for \(\epsilon \to 0\) is called diffusive limit.

Introducing the macroscopic variables \(u = f + g\) and \(v = c(f - g)/\epsilon\), corresponding to the mass density and the flux, the model is rewritten in the form of a relaxation system
\[\begin{align*}
    u_t + v_x &= 0, \\
    v_t + \frac{c^2}{\epsilon^2} u_x &= -\frac{2k}{\epsilon^2} v.
\end{align*}\] (3)

In the limit \(\epsilon = 0\) formally we obtain the local equilibrium \(2kv = -c^2u_x\) and the system reduces to the heat equation \(u_t = \frac{c^2}{2k} u_{xx}\).

For notation simplicity in the sequel we will assume \(c = 1\) and \(k = 1/2\), so that the limiting heat equation has a constant diffusion coefficient equal to one.

We will introduce the spatial grid points \(x_{i+1/2}, i = \ldots, -1, 0, 1, \ldots\) with uniform mesh width \(\Delta x = x_{i+1/2} - x_{i-1/2}\). As usual we denote by \(U_{i+1/2}(t) = U(x_{i+1/2}, t)\) the nodal values and by \(U_i(t)\) the cell averages of \(U\) in the cell \([x_{i-1/2}, x_{i+1/2}]\) at time \(t\)
\[U_i(t) = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} U(x, t) \, dx.\]

Thus we consider the following semi-discrete system in conservative form
\[\begin{align*}
    (u_i)_t + \frac{v_{i+1/2} - v_{i-1/2}}{\Delta x} &= 0, \\
    (v_i)_t + \frac{1}{\epsilon^2} \frac{u_{i+1/2} - u_{i-1/2}}{\Delta x} &= -\frac{1}{\epsilon^2} v_i.
\end{align*}\] (4)

As observed in [11] a standard upwind selection of the numerical fluxes in (4) originates a numerical dissipation which may dominate the physical one. For example, the first order upwinding yields
\[\begin{align*}
    u_{i\pm 1/2} &= \frac{1}{2}(u_i + u_{i\pm 1}) \pm \frac{\epsilon}{2}(v_i - v_{i\pm 1}), \\
    v_{i\pm 1/2} &= \frac{1}{2}(v_i + v_{i\pm 1}) \pm \frac{1}{2\epsilon}(u_i - u_{i\pm 1}).
\end{align*}\] (5)

It is easy to verify that the use of (5) in the discrete equation (4) for small values of \(\epsilon\) leads to the semi-discrete approximation
\[\begin{align*}
    (u_i)_t - \frac{u_{i+2} + u_{i-2} - 2u_i}{(2\Delta x)^2} - \frac{\Delta x}{2\epsilon} \frac{u_{i+1} + u_{i-1} - 2u_i}{\Delta x^2} &= 0,
\end{align*}\] (6)

for which the numerical dissipation dominates the physical one unless \(\Delta x \ll \epsilon\).

Note also that we obtain a wider stencil for the second order derivative in the heat equation with respect to the classical one. There are several alternatives to deal with this problem, see for example [4, 7, 10, 11].
The upwind selection (5) can be modified in another way so as to capture the proper parabolic behavior. The idea now is to apply the upwind selection only in the hyperbolic regime and to use a standard central discretization in the parabolic one. This can be better understood by rewriting system (3) in the form [8]

\begin{align*}
  u_t + v_x &= 0 \\
  v_t + \phi^2 u_x &= -\frac{1}{\varepsilon^2} \left( v + (1 - \phi^2 \varepsilon^2) u_x \right),
\end{align*}

where \( \phi = \phi(\varepsilon) \) is a suitable function such that \( \phi = O(1) \) for small values of \( \varepsilon \). The above reformulation is equivalent to rewrite

\begin{align*}
  u_x &= \underbrace{\phi^2 \varepsilon^2 u_x}_{\text{hyperbolic regime}} + \underbrace{(1 - \phi^2 \varepsilon^2) u_x}_{\text{parabolic regime}},
\end{align*}

the key point now is to combine two different discretization for the different regimes. For example in [8] it was proposed to consider an upwind discretization on the variables \((u \pm v)/2\) for the first derivative combined with a central scheme for the second term. This interpolation is also relevant in term of the time discretization of the system, since the hyperbolic flux can be evaluated explicitly whereas for the parabolic flux it is desirable to have an implicit integrator.

In the simplest case, taking \( \phi \equiv 1 \) this idea leads to the modified fluxes

\begin{align*}
  u_{i\pm1/2} &= \frac{1}{2} (u_i + u_{i\pm1}) \pm \frac{\varepsilon^2}{2} (v_i - v_{i\pm1}), \\
  v_{i\pm1/2} &= \frac{1}{2} (v_i + v_{i\pm1}) \pm \frac{1}{2} (u_i - u_{i\pm1}).
\end{align*}

With these fluxes, for small values of \( \varepsilon \), the semi-discrete system (4) gives

\begin{align*}
  (u_i)_t - u_{i+2} + u_{i-2} - 2u_i - \frac{\Delta x}{(2\Delta x)^2} \frac{u_{i+1} + u_{i-1} - 2u_i}{\Delta x^2} = 0,
\end{align*}

which is an approximation of the equilibrium heat equation with an accuracy of \( O(\Delta x/2) \). Second order extension can be obtained easily simply increasing to second order the choice of the flux used for the hyperbolic derivative. By combining this with standard implicit-explicit discretizations one obtain a scheme that work uniformly in \( \varepsilon \) with a CFL condition of the type \( \Delta t = O(\Delta x^2) \) in the limit \( \varepsilon \to 0 \).

3. Asymptotically implicit IMEX schemes. In order to write a stable discretization to system (3) we should use implicit temporal integrators on the stiff terms. Because the stiffness of the convection term depends on the size of the two eigenvalues \( \pm 1/\varepsilon \) of the Riemann invariant form (2) of system (3), both convection terms have to be implicit. However a fully implicit scheme causes global data dependencies, a severe disadvantage if we consider second order nonlinear schemes using slope limiters. Furthermore, the gain of stability is partially offset by the loss of accuracy typical of implicit schemes in the context of wave-propagation phenomena. The approach just introduced permits to overcome this problem since we can apply an Implicit-Explicit discretization in the form

\begin{align*}
  u_t + \underbrace{v_x}_{\text{Explicit}} &= 0, \\
  v_t + \underbrace{\phi^2 u_x}_{\text{Explicit}} &= -\frac{1}{\varepsilon^2} (v + (1 - \phi^2 \varepsilon^2) u_x),
\end{align*}

\begin{align*}
  v_t + \underbrace{\phi^2 u_x}_{\text{Implicit}} &= -\frac{1}{\varepsilon^2} (v + (1 - \phi^2 \varepsilon^2) u_x).
\end{align*}
This approach was proposed in [8] and originates an explicit discretization of the heat equation in the limit. We refer to this standard approach as *asymptotically explicit* IMEX method.

Here we adopt a different technique which has been proposed recently in [1] which permits to overcome the parabolic time step limitation for small regimes of \( \varepsilon \). The method is based on a penalization technique consisting in adding and subtracting to the first equation in system (3) an appropriate term, which will establish the correct limiting diffusion equation. This lead to the modified system

\[
\begin{align*}
\frac{\partial u}{\partial t} + \left( v + \mu \frac{\partial u}{\partial x} \right)_{x} &= \mu \frac{\partial^{2} u}{\partial x^{2}}, \\
\frac{\partial v}{\partial t} + \frac{1}{\varepsilon^2} \frac{\partial u}{\partial x} &= -\frac{1}{\varepsilon^2} v,
\end{align*}
\]

(12)

where \( \mu = \mu(\varepsilon) \) is such that \( \mu(0) = 1 \). Finally, using the interpolated fluxes approach we end up with the following system

\[
\begin{align*}
\frac{\partial u}{\partial t} + \left( v + \mu \frac{\partial u}{\partial x} \right)_{x} &= \mu \frac{\partial^{2} u}{\partial x^{2}}, \\
\frac{\partial v}{\partial t} + \phi^2 \frac{\partial u}{\partial x} &= -\frac{1}{\varepsilon^2} \left( v + (1 - \phi^2 \varepsilon^2) \frac{\partial u}{\partial x} \right),
\end{align*}
\]

(13)

which is now tackled with an IMEX scheme in the form

\[
\begin{align*}
\text{Explicit} & & \frac{\partial u}{\partial t} + \left( v + \mu \frac{\partial u}{\partial x} \right)_{x} = \mu \frac{\partial^{2} u}{\partial x^{2}}, \\
\text{Implicit} & & \frac{\partial v}{\partial t} + \phi^2 \frac{\partial u}{\partial x} = -\frac{1}{\varepsilon^2} \left( v + (1 - \phi^2 \varepsilon^2) \frac{\partial u}{\partial x} \right)
\end{align*}
\]

(14)

In addition to the space discretizations of the first order derivatives described in the last section let us specify that in (14) the second order space derivative in the first equation is discretized using a standard second order central discretization. Note that in the limit \( \varepsilon \to 0 \) we obtain a fully implicit discretization of the heat equation with a standard compact stencil instead of the explicit discretization on a wider stencil described before. The only additional cost is due to the inversion of the linear system originated by the implicit discretization of the heat equation. We refer to this new approach as *asymptotically implicit* IMEX method.

Let us finally recall the general structure of the IMEX-RK scheme adopted. Here we omit the space derivatives for simplicity of notation. For the internal stages for \( k = 1, \ldots, \nu \) we have

\[
\begin{align*}
U^{(k)} &= u^n - \Delta t \sum_{j=0}^{k-1} \bar{a}_{kj} \left( V^{(j)} + \mu U_{x}^{(j)} \right)_{x} + \mu \Delta t \sum_{j=1}^{k} a_{kj} U_{x}^{(j)}, \\
V^{(k)} &= v^n - \Delta t \phi^2 \sum_{j=0}^{k-1} \bar{a}_{kj} U_{x}^{(j)} - \frac{\Delta t}{\varepsilon^2} \sum_{j=1}^{k} a_{kj} \left[ v^{(j)} + (1 - \phi^2 \varepsilon^2) U_{x}^{(j)} \right].
\end{align*}
\]

For the numerical solution we simply have \( u^{n+1} = U^\nu \) and \( v^{n+1} = V^\nu \) since we restrict to globally stiffly accurate IMEX schemes [1]. In all considered IMEX schemes, matrix \( \bar{A} \) is lower triangular with zero diagonal, while matrix \( A \) is lower triangular, i.e. the implicit scheme is a Diagonally Implicit Runge-Kutta (DIRK) scheme. This choice guarantees that implicit terms are, indeed, always explicitly
evaluated. For more details on properties and requirements of IMEX-RK schemes when applied to kinetic equations in the diffusion limit we refer to [1] and the references therein.

4. Numerical results. We present some numerical results in order to show the behavior of the proposed scheme. In the computed results second order accuracy is obtained using a WENO scheme for the hyperbolic derivative whereas the parabolic flux is always evaluated by central difference. First order Implicit-Explicit Euler scheme and second order ARS(2, 2, 2) scheme are adopted for the time discretization (see [1]). In all test cases the initial conditions are given by

\[
\begin{align*}
  u_0(x) &= 1 \text{ and } v_0(x) = 0 \text{ if } x \leq 0, \\
  u_0(x) &= 0 \text{ and } v_0(x) = 0 \text{ if } x > 0.
\end{align*}
\]

The space variable \(x\) ranges in the computational domain \([-1, 1]\) and we choose \(N_x\) grid points.

**Test 1.** In this case computations are performed in the hyperbolic regime for \(\epsilon = 1\) and consider system (13) with \(\phi \equiv 1\) and \(\mu \equiv 0\). For this test case, we assume \(\Delta t = \Delta x/2\) and \(N_x = 100\). In fig. 4 we show the results of the computation at time \(T_f = 0.3\) of the \(u\) component. In this non equilibrium regime we can see that the transport term, which dominant with respect to the source term, is properly approximated.

![Figure 1](image1.png)

**Figure 1.** Test 1: Hyperbolic regime. Solution for \(u(x)\) at time \(T_f = 0.3\) for \(\epsilon = 1\). Left: first order method. Right: second order method.

**Test 2.** Here we consider the parabolic regime for \(\epsilon = 10^{-3}\), and approximate system (13) with \(\phi \equiv 1\) and \(\mu \equiv 1\). For this test case, we can assume \(\Delta t = \Delta x\) and \(N_x = 100\). In fig. 4 we show the result of the computation at time \(T_f = 0.1\) of the \(u\) component. We stop the computation before the stationary state is reached, thus showing a plot which shows we catch the correct diffusive behavior; in this case, the contribution given by the transport term is less important then that of the relaxation term.
Test 3. In the last test case we consider a mixing regime where on the left of the computational domain for $x \leq 0$ we are in the hyperbolic regime with $\varepsilon = \varepsilon_L = 0.2$ while on the right for $x > 0$ we are in the parabolic regime with $\varepsilon = \varepsilon_R = 0.01$. For this last test case we consider $N_x = 60$ points as in [2]. We set $\mu = 1$ while we set $\phi = 1/\varepsilon_L$ for $x \leq 0$ and $\phi = 1$ for $x > 0$ (we consider the correct “physical” velocity in the kinetic regime). For this test case, we set $\Delta t = \varepsilon_L \Delta x/5$. In fig. 4 we show the result of the computation at time $T_f = 0.05$ of the $u$ component.

We observe that our scheme is able to efficiently compute the solution, with a good behavior also at the interface between the two regimes. We refer to [2] for comparisons.

5. Conclusions. We have presented a general approach to tackle kinetic equations in the diffusive scaling which leads to a fully implicit discretization of the limiting diffusion terms. The method is based on combining the strategies presented in [8]
for the space derivatives together with the time discretization presented in [1]. We emphasize that different choices of the space discretizations may originate schemes with different stability properties. In particular the optimal choice of the schemes parameters $\phi$ and $\mu$ which permits to switch between the hyperbolic and the parabolic flux and the explicit and implicit integrators deserves further investigations.

REFERENCES


E-mail address: giacomo.dimarco@math.univ-toulouse.fr
E-mail address: lorenzo.pareschi@unife.it
E-mail address: vittorio.rispoli@unife.it
STEADY SELF-SIMILAR INVISCID FLOW

Volker Elling and Joseph Roberts
Department of Mathematics
University of Michigan
Ann Arbor, MI 48105, USA

Abstract. We consider admissible \( L^\infty \) solutions to 2-dimensional \( m \times m \) systems of hyperbolic conservation laws that are steady in time, constant along rays emanating from the origin, and sufficiently small perturbations of a constant background state. We classify the possible structures of these solutions and prove that they must be \( BV \). As a special case, we obtain uniqueness in this class of \( L^\infty \) functions for admissible forward in time solutions to 1-dimensional Riemann problems and show that backward in time solutions also must be \( BV \).

1. Introduction. We consider a 2-dimensional system of conservation laws

\[
U_t + f^x(U)_x + f^y(U)_y = 0.
\] (1)

Here the unknown \( U \) and the flux functions \( f^x \) and \( f^y \) take values in \( \mathbb{R}^m \). The fluxes are assumed to be smooth and possess an entropy-entropy flux pair \( (\eta, \psi) = (\eta, \psi_x, \psi_y) \) with uniformly convex \( \eta \) on some open nonempty set in state space. The entropy inequality is then

\[
\eta(U)_t + \psi^x(U)_x + \psi^y(U)_y \leq 0.
\] (2)

In multidimensional inviscid Euler flow, there are well studied cases in which there exists a distinguished point around which the flow is, to first order, constant along rays starting at this point. These include regular reflection (four shock waves meeting at a point) (see [7], [6], [3]) and Mach reflection (three shocks and a contact meeting at a point) (see [1]). However, other configurations such as triple points (three shocks with no other waves in between) cannot occur for most commonly used equations of state (see [13],[4], [12]). Beyond these and some other special cases, the possible configurations of such waves meeting at a point have not been classified.

The flow is steady from the point of view of an observer moving with the distinguished point described above. Therefore, we consider solutions of the form

\[
U(t, x, y) = U(\phi), \text{ with } \phi = \angle(x, y) \in [0, 2\pi).
\] (3)

We are also motivated by a possible nonuniqueness result found in [5], in which an unsteady numerical solution was observed which took a steady self-similar solution as initial data. Perhaps analysis of the steady problem will lead to an analytical example of nonuniqueness of this form.

2000 Mathematics Subject Classification. 35L02, 35L67.
Key words and phrases. Conservation law, Riemann problem, bounded variation, self-similar, shocks.

This material is based on work supported by NSF grant DMS-0907074.
Though \( BV \) seems to be the ideal function space for the study of 1-dimensional conservation laws (see [2]), it is well known (see [10]) that it is not appropriate for multidimensional conservation laws. However, when considering steady self-similar solutions, (1) reduces to a simpler form that is similar enough to the 1-dimensional setting that suggests \( BV \) may be an appropriate function space for these solutions.

We assume \( \mathcal{U} \) is an admissible weak solution of (1) that is \( L^\infty \)-close to some constant supersonic background state \( \bar{\mathcal{U}} \); that is
\[
||\mathcal{U}(\cdot) - \bar{\mathcal{U}}||_{L^\infty} \leq \epsilon. \tag{4}
\]

In the remainder, \( \epsilon \) denotes a small positive constant which may be adjusted a finite number of times. This background state is supersonic in the sense that the polynomial
\[
P(x : y) = \det \left( x f_y(U) - y f_x(U) \right)
\]
(5)
(where \((x : y)\) are homogeneous coordinates) has exactly \( m \) distinct real roots. This is equivalent to requiring the steady form of the system of conservation laws to be strictly hyperbolic. The terminology \emph{supersonic} is borrowed from the case we have in mind — the steady isentropic Euler equations, which satisfy this assumption if and only if \( \mathcal{U} \) is supersonic.

We prove \( \mathcal{U} \) must be constant outside of \( 2m \) thin sectors near the roots of (5), and classify the possible configurations of shocks, simple waves, and contact discontinuities in each sector. We also prove that under these hypotheses \( \mathcal{U} \) must be \( BV \).

The reduced form of (1) is precisely of the form of a self-similar (function of \( \frac{x}{t} \)) 1-dimensional Riemann problem, and so our results prove that self-similar solutions that are \( L^\infty \)-close to a constant background state are \( BV \). Combined with known uniqueness results in \( BV \) for forward in time Riemann problem solutions, we obtain uniqueness in this class of \( L^\infty \) functions. This improves an earlier result [9] which required all characteristic fields to be genuinely nonlinear; we assume simple eigenvalues that are either genuinely nonlinear or linearly degenerate.

The details of our work appeared in [8].

2. Description of the problem. Suppose \( \mathcal{U} \in L^\infty \) satisfies (3), and denote \( \xi := \frac{x}{y} \). If we interpret (1) and (2) in the distributional sense for test functions supported away from \( x = 0 \), we obtain
\[
\begin{align*}
\left( f_y(U) - \xi f_x(U) \right)_{\xi} + f^x(U) &= 0, \\
\left( \psi_y(U) - \xi \psi_x(U) \right)_{\xi} + \psi^x(U) &\leq 0, \quad x > 0, \\
\left( \psi_y(U) - \xi \psi_x(U) \right)_{\xi} + \psi^x(U) &\geq 0, \quad x < 0,
\end{align*}
\]
(6)
(where \( U = U(\xi) \)).

Without loss of generality, we can rotate coordinates so that none of the \( m \) distinct roots of (5) lie on the \( y \)-axis. Then the roots of (5) correspond to the generalized eigenvalues of the matrix pair \((f^y_{\mathcal{U}}(\mathcal{U}), f^x_{\mathcal{U}}(\mathcal{U}))\) — that is, the roots of the polynomial
\[
p(\xi) := \det \left( f^y_{\mathcal{U}}(\mathcal{U}) - \xi f^x_{\mathcal{U}}(\mathcal{U}) \right).
\]
(7)

The fact that no roots of (5) lie on the \( y \)-axis also implies that \( f^y_{\mathcal{U}}(U) \) is non-degenerate for \( |U - \bar{\mathcal{U}}| < \epsilon \) with \( \epsilon > 0 \) sufficiently small. Therefore, since we are
only interested in small perturbations of $\mathbf{U}$, $f^x(U)$ is a diffeomorphism on the phase space under consideration and we can make the change of dependent variables
\[
V := f^x(U), \quad f(V) := f^y(U(V)),
\]
\[
e(V) := \psi^x(U(V)), \quad q(V) := \psi^y(U(V)).
\]

It can be verified that $(e, q)$ is an entropy-entropy flux pair for the flux $f$, although $e$ may not be convex. In any case, (6) becomes
\[
\begin{cases}
(f(V) - \xi V)_\xi + V = 0, \\
(q(V) - \xi e(V))_\xi + e(V) \leq 0, & x > 0, \\
(q(V) - \xi e(V))_\xi + e(V) \geq 0, & x < 0.
\end{cases}
\] (8)

Let $\mathbf{V} := f^x(\mathbf{U})$, and denote
\[
\mathcal{P}_\epsilon := \left\{ V \in \mathbb{R}^n \left| |V - \mathbf{V}| \leq \epsilon \right. \right\}
\]
for $\epsilon > 0$ sufficiently small. Since the roots of (7) are distinct, the eigenvalues and eigenvectors of $f_V(V)$, for $\epsilon > 0$ sufficiently small, will be distinct smooth functions of $V$ (note the eigenvalues of $f_V$ coincide with the generalized eigenvalues of the matrix pair $(f^x_V, f^y_V)$). Denote the eigenvalues by $\lambda^\alpha(V)$ and the right and left eigenvectors by $\rho^\alpha(V)$ and $l^\alpha(V)$, for $\alpha = 1, \ldots, m$. We then have
\[
\lambda^1(V) < \lambda^2(V) < \cdots < \lambda^m(V) \text{ for all } V \in \mathcal{P}_\epsilon.
\] (9)

We assume that each eigenvalue is either genuinely nonlinear or linearly degenerate. That is, for each $\alpha = 1, \ldots, m$,
\[
\lambda^\alpha_V(V(r^\alpha(V) \neq 0 \text{ for all } V \in \mathcal{P}_\epsilon) \text{ or } \lambda^\alpha_V(V(r^\alpha(V) \equiv 0 \text{ for all } V \in \mathcal{P}_\epsilon.
\] (10)

For $V^\pm \in \mathcal{P}_\epsilon$ we can define a diagonalizable averaged matrix $A(V^-, V^+)$. It satisfies $\hat{A}(V, V) = f_V(V)$ and is smooth in $V^\pm$ with smooth, distinct eigenvalues. Most importantly,
\[
\hat{A}(V^-, V^+)(V^+ - V^-) = f(V^+) - f(V^-).
\] (11)

Denote the eigenvalues by $\hat{\lambda}^\alpha$ and the left eigenvectors by $\hat{l}^\alpha$ for $\alpha = 1, \ldots, m$.

2.1. Intuition regarding the structure of solutions. Even though we cannot justify differentiating $V$, it is still instructive to consider the strong form of the first line of (8), which is
\[
(f_V(V) - \xi I)V_\xi = 0.
\] (12)

This suggests that if on some interval $\xi$ is not an eigenvalue of $f_V(V(\xi))$, then $V_\xi = 0$ and thus $V$ is constant. On the other hand, if $\xi$ is an eigenvalue then $V_\xi$ is parallel to the associated eigenvector and we have a simple wave.

The Rankine-Hugoniot conditions for a shock at $\xi$ connecting two states $V^\pm$ are equivalent to $\xi$ being an eigenvalue of $\hat{A}(V^-, V^+)$, which by smoothness will be near an eigenvalue of $f_V(V)$.

Therefore, we anticipate the interesting behavior to occur when $\xi$ is close to an eigenvalue of $f_V(V)$. However, careful analysis must be done to rigorously justify this since we are only assuming $V \in L^\infty$, and therefore $V$ could be nowhere differentiable. Even worse, the Rankine-Hugoniot and Lax conditions must be derived carefully since we are not necessarily assuming piecewise smooth solutions, or even solutions with well-defined left and right limits.
3. Main results. The intuition we developed above can be made rigorous, and we have the following theorem concerning the structure of such solutions.

**Theorem 3.1.** (See Figure 1.) Suppose a distributional solution to (1) and (2) (with fluxes so that (9) and (10) are satisfied) satisfies (3) and (4), and that the background state $U$ satisfies (5). Then the following must be true.

(a) $U$ must be constant outside of $2m$ thin sectors centered at the roots of (7), which we can group as $m$ forward and $m$ backward sectors.

(b) Linearly degenerate sectors each contain at most one contact discontinuity.

(c) Genuinely nonlinear forward sectors each contain at most one shock or simple wave.

(d) Genuinely nonlinear backward sectors can each contain infinitely many shocks and simple waves, but there cannot be consecutive simple waves.

(e) Each shock wave has a neighborhood on each side on which $U$ is constant, and the size of the neighborhood is lower bounded proportionally to shock strength.

(f) The width of the sectors and constant of proportionality depend only on the system and $||U(\cdot) - \overline{U}||_{L^\infty}$.

In addition, these solutions must possess certain regularity properties.

**Theorem 3.2.** Suppose $U$ satisfies the hypotheses of the previous theorem. Then $U$ is of bounded variation. More specifically, $U = U_S + U_L$ where $U_S$ is a saltus (jump) function of bounded variation, and $U_L$ is Lipschitz, and so $U$ is a special function of bounded variation. The total variation of $U_S$ and the Lipschitz constant of $U_L$ are independent of $U$ and only depend on the system and $||U(\cdot) - \overline{U}||_{L^\infty}$.

**Figure 1.** Theorems 3.1 and 3.2 for isentropic Euler and a background state with constant supersonic horizontal velocity to the right. There is one linearly degenerate eigenvalue (corresponding to shear waves), with associated sectors near the $x$-axis. The genuinely nonlinear sectors (corresponding to acoustic waves) occur near the angles $\pm \arcsin \left( \frac{c}{v} \right)$ from the positive and negative $x$-axis.
Remark 1. The terms forward and backward sectors for genuinely nonlinear fields need to be defined carefully and do not necessarily correspond to \( x > 0 \) and \( x < 0 \), respectively. It is related to whether a field is hyperbolic in the positive or negative \( x \)-direction, which is determined by the properties of the “entropy” \( e \).

Although \( e \) may not be a convex function of \( V \), it can still be used to derive Lax-type conditions for genuinely nonlinear fields. The important quantity in deriving the Lax condition from the entropy inequality is

\[
e_{VV} r^\alpha r^{\alpha}.
\]

(13)

More specifically, whether it is positive or negative (zero would pose major problems in the analysis) determines the direction of the inequalities in the Lax conditions. Fortunately, for convex \( \eta \) and nondegenerate \( f_U \) this quantity can be shown to be nonzero. For a given field, if (13) is positive, then the forward sectors described in the theorem are in fact the sectors with \( x > 0 \), and the backward sectors have \( x < 0 \). The situation is reversed for that field if (13) is negative. However, if the eigenvalues of \( f_U(U) \) are all positive (negative), \( e \) will be convex (concave). If \( f_U(U) \) has eigenvalues with mixed signs, then (13) needs to be checked for each field individually.

For the isentropic Euler equations, if the background state is constant supersonic horizontal velocity to the right, then all the forward sectors have \( x > 0 \), as in Figure 1 (that is, the steady system is hyperbolic in the positive \( x \) direction). If the background state is constant supersonic horizontal velocity to the left, then the forward sectors have \( x < 0 \) (corresponding to the system being hyperbolic in the negative \( x \) direction). It is not hard to construct systems and background states where the forward and backward sectors are intermixed between \( x > 0 \) and \( x < 0 \).

Remark 2. If we consider self-similar solutions to 1-dimensional Riemann problems, then we initially are considering a system of the form (8) with \( e \) convex. Then, the \( t > 0 \) sectors are forward while the \( t < 0 \) sectors are backward. Since forward in time solutions to the Riemann problem are unique in the class of \( BV \) with small total variation, Theorem 3.2 extends this uniqueness to the class of self-similar solutions \( L^\infty \)-close to a constant background state. Theorem 3.2 shows that the backward in time solutions also must be \( BV \), though they will not be unique.

4. Some ideas used in the proofs. The full details are found in [8], but some selected ideas from various steps used to prove the main results follow.

4.1. Pointwise information from the weak form. It is useful to derive pointwise information about \( V \) from the weak form (8). The first line shows that the distributional derivative of \( (f(V) - \xi V) \) is \( L^\infty \) — therefore \( (f(V) - \xi V) \) is Lipschitz and the fundamental theorem of calculus can be used. Using (11) we obtain, for almost every \( \xi_1, \xi_2 \), that

\[
(A(V(\xi_1), V(\xi_2)) - \xi_1 I)(V(\xi_2) - V(\xi_1)) = \int_{\xi_1}^{\xi_2} V(\eta) d\eta.
\]

(14)

It can be shown that there exists a version of \( V \) such that (14) holds for all \( \xi_1, \xi_2 \). Similarly, since the left side of the entropy inequality in the third line of (8) is a non-negative distribution (and hence a non-negative measure), \( (q(V) - \xi e(V)) \) is a
non-decreasing function of bounded variation, and so
\[
\left(q(V(\xi)) - \xi e(V(\xi))\right)\frac{\xi_2 - \xi_1}{\xi_2 - \xi_1} \geq - \int_{\xi_1}^{\xi_2} e(V(\eta)) \, d\eta, \text{ for } x < 0, \tag{15}
\]
for almost all \(\xi_1 < \xi_2\) (a version of \(V\) can be constructed satisfying (14) and (15) everywhere).

4.2. **Rankine-Hugoniot and Lax conditions for \(L^\infty\) functions.** Since we do not assume \(V \in BV\), \(V\) may not have well defined left or right limits at any point \(\xi\). Consider a pair of sequences \(\{\xi_k^-\}, \{\xi_k^+\}\) both converging to \(\xi\), with \(\xi_k^- < \xi_k^+\). Since \(V\) has values in the compact set \(\mathcal{P}\), we may choose subsequences \(\{\xi_k^-\}\) and \(\{\xi_k^+\}\) such that \(V(\xi_k^\pm) \rightarrow V^\pm\). Taking further subsequences if necessary, we can ensure that \(\xi_k^- < \xi_k^+\) for all \(k\). Assuming there is no ambiguity in which sequences are used, in this context we define for any function \(g\)
\[
[g(V)] := g(V^+) - g(V^-).
\]
Applying (14) with \(\xi_1 = \xi_k^-\) and \(\xi_2 = \xi_k^+\) and taking the limit \(k \rightarrow \infty\) we obtain
\[
(A(V^+) - \xi I)[V] = 0.
\]
If \(V\) is discontinuous at \(\xi\), there exists at least one pair of sequences on which \(V \rightarrow V^\pm\) with \(V^+ \neq V^-\). Then, \(\xi\) is one of the \(m\) distinct eigenvalues of \(A(V^\pm)\) and \(|V| \neq 0\) is parallel to the associated eigenvector, which is the usual Rankine-Hugoniot condition for shocks. Note these \(V^\pm\) are not necessarily the same for different choices of sequences \(\{\xi_k^\pm\}\), and not all choices of sequences lead to \(|V| \neq 0\). Applying (15) with \(\xi_1 = \xi_k^-\) and \(\xi_2 = \xi_k^+\) (the use of (15) is why we require \(\xi_k^- < \xi_k^+\) for all \(k\)) and taking \(k \rightarrow \infty\) we obtain
\[
[g(V)] - \xi[e(V)] \geq 0, \quad x < 0. \tag{16}
\]
Suppose that (13) is positive for the \(\alpha\)-characteristic field, that the \(\alpha\)-field is genuinely nonlinear, and that there is a discontinuity at some \(\xi\) near the eigenvalue \(\lambda^\alpha(V^-)\). Since \(V^\pm \in \mathcal{P}\), with \(\epsilon\) sufficiently small, the implicit function theorem ensures that \(V^+\) lies on the \(\alpha\)-shock curve of \(V^-\), and (16) is satisfied if and only if
\[
\lambda^\alpha(V^-) < \xi < \lambda^\alpha(V^+). \tag{17}
\]
The direction of these inequalities is reversed if either \(x > 0\) or (13) is negative, but is the same if both of these are true.

4.3. **\(V\) must be constant away from eigenvalues of \(f_V(V)\).** We now sketch an argument that shows how some of the intuition in Section 2.1 can be made rigorous.

Suppose that \(V\) is continuous on some open interval \(I\), and that \(\xi\) is not equal to any eigenvalue of \(f_V(V(\xi))\) for all \(\xi \in I\). We claim that \(V\) is in fact Lipschitz continuous on this interval. Suppose it is not at some \(\xi \in I\). Then we can choose a sequence \(\{h_n\} \rightarrow 0\) (with \(h_n \neq 0\)) such that
\[
\left|\frac{V(\xi + h_n) - V(\xi)}{h_n}\right| \nearrow \infty. \tag{18}
\]
Divide both sides of (14) (with \(\xi_1 = \xi, \xi_2 = \xi + h_n\)) by \(|V(\xi + h_n) - V(\xi)|\) to obtain
\[
\left(\frac{A(V(\xi), V(\xi + h_n)) - \xi I}{|V(\xi + h_n) - V(\xi)|}\right) \frac{V(\xi + h_n) - V(\xi)}{|V(\xi + h_n) - V(\xi)|} = \frac{1}{|V(\xi + h_n) - V(\xi)|} O(|h_n|) = o(1) \quad \text{by (18)}.
\]
However, since $\xi$ is not an eigenvalue of $f_V(V(\xi))$, for $n$ sufficiently large $(\hat{A} - \xi I)$ is uniformly regular (by continuity of $V$) and the left side is thus bounded away from zero, a contradiction. Therefore, $V$ is Lipschitz and we can invoke the strong form (12) on a set of full measure, obtain $V_\xi = 0$ almost everywhere in $I$, and conclude $V$ is constant on $I$.

Assuming $V$ is continuous is too strong of an assumption for general $L^\infty$ solutions. However, as discussed above, if $V$ is discontinuous at $\xi$, then $\xi$ is an eigenvalue of $\hat{A}(V)^\pm$ for some $V^\pm \in P_\xi$. By smoothness, if $\epsilon$ is sufficiently small, we can find a $\delta > 0$ such that if $\xi$ is separated from the eigenvalues of $f_V(V)$ by at least $\delta$, then $V$ must be continuous. Applying the previous argument shows it is constant. This $\delta$ is then the thickness of the $2m$ sectors, outside of which $V$, and therefore $U$, will be constant.

4.4. $V$ must be locally constant on either side of a shock. Suppose there is a discontinuity at $\xi_0$ in a backward, genuinely nonlinear sector. Pick $\{\xi_k^+\}$ with $\xi_k^+ \searrow \xi_0$. Suppose there is no $\delta_0^+ > 0$ such that $\lambda^+(V(\xi)) > \xi$ for all $\xi \in (\xi_0, \xi_0 + \delta_0^+)$. Pick $\{\xi_k^{++}\}$ converging to $\xi_0$ with $\lambda^+(V(\xi_k^{++})) \leq \xi_k^{++}$ and $\xi_k^+ < \xi_k^{++}$ (take a subsequence of $\{\xi_k^+\}$ if necessary). Then the backward Lax condition (17), for the two pairs of sequences $\{\xi_k^+\}$ and $\{\xi_k^{++}\}$, implies $\lambda^+(V^{++}) > \xi_0$, contradicting the construction of $\{\xi_k^{++}\}$.

Therefore, there exists $\delta_0^+ > 0$ such that $\lambda^+(V(\xi)) > \xi$ for all $\xi \in (\xi_0, \xi_0 + \delta_0^+)$. Analogously, there exists $\delta_0^- > 0$ such that $\lambda^-(V(\xi)) < \xi$ for all $\xi \in (\xi_0 - \delta_0^-, \xi_0)$.

Suppose there were another discontinuity at $\xi_1 \in (\xi_0, \xi_0 + \delta_0^+)$. Then, perform the same argument for the shock at $\xi_1$ to find an $\eta \in (\xi_0 + \delta_0^+, \xi_0)$. Then, $\lambda^+(V(\eta)) > \xi$, which is a contradiction. Therefore $V$ is continuous on $(\xi_0, \xi_0 + \delta_0^+)$. By definition of $\delta_0^+$ and the argument in the Section 4.3, $V$ must be constant on $(\xi_0, \xi_0 + \delta_0^+)$. A similar argument works for a neighborhood on the other side of $\xi_0$, and also for a shock in a forward sector. Then, at least for discontinuities in sectors corresponding to genuinely nonlinear fields, the left and right limits are well defined and the Rankine-Hugoniot and Lax conditions hold as in the familiar case of piecewise-smooth solutions.

4.5. Degenerate sectors contain at most one contact. If $\lambda^\alpha$ is linearly degenerate, then $\xi \mapsto \lambda^\alpha(V(\xi))$ is continuous in the $\alpha$-sector — the only possible jumps in $V$ at $\xi$ are between two states on a contact curve (by the implicit function theorem), on which $\lambda^\alpha(V^\pm) = \xi$. This implies the set $R := \{\xi \in I^\alpha \mid \lambda^\alpha(V(\xi)) = \xi\}$ is closed in the sector, and so its complement in the sector is a countable union of open intervals. We can argue its complement could be at most two open intervals, so $R$ is at most a single closed interval.

If the strong form (12) could be considered, we would have a contradiction, since it would imply $V_\xi$ is parallel to $r^\alpha(V(\xi))$. This contradicts the differentiated form of $\lambda^\alpha(V(\xi)) = \xi$, which is

$$\lambda^\alpha_0(V(\xi))V_\xi = 1,$$

(19)
since linear degeneracy would make the left side zero. Differentiating $V$ cannot be justified, but we only need a single sequence in $R$ converging to $\xi_0 \in R$ on which the difference quotients for $V$ converge to obtain the contradiction. We can use a
theorem due to Saks [11], which states that for any finite (real valued) function \( F \), the set of points \( x \) at which
\[
\lim_{h \to 0^+} \frac{|F(x + h) - F(x)|}{h} = +\infty
\]
is of measure zero. We apply the theorem to the real valued function
\[
\xi \mapsto l^\alpha(\nabla)(V(\xi) - V(\xi_0)). \tag{20}
\]
In addition, for \( \beta \neq \alpha \),
\[
\xi \mapsto \hat{l}^\beta(V(\xi_0), V(\xi))(V(\xi) - V(\xi_0)) \tag{21}
\]
is Lipschitz at \( \xi_0 \) due to the smoothness of the eigenvalues of the averaged matrix. This can be seen by using (14) with \( \xi_1 = \xi_0 \) and \( \xi_2 = \xi \) and multiplying on the left by \( \hat{l}^\beta(V(\xi_0), V(\xi)) \) to obtain
\[
(\hat{\lambda}^\beta(V(\xi_0), V(\xi)) - \xi_0)\hat{\lambda}^\beta(V(\xi_0), V(\xi))(V(\xi) - V(\xi_0)) = O(|\xi - \xi_0|),
\]
and using the fact that \( \xi_0 \) is in the sector for \( \lambda^\alpha \) implies that it is well separated from \( \hat{\lambda}^\beta \) for \( \beta \neq \alpha \).

Together these \( m \) functions defined in (20) and (21) can be used to define a diffeomorphism on \( \mathcal{P}_e \), and so by taking further and further subsequences we obtain a sequence of points in \( \mathcal{R} \) converging to \( \xi_0 \in \mathcal{R} \) on which \( V \) has a convergent difference quotient, which yields the contradiction in (19) described above. Thus \( \lambda^\alpha(V(\xi)) = \xi \) at at most one point, and so there can be at most one contact in each degenerate sector (after using the continuity on either side of the single contact and the argument in Section 4.3).

REFERENCES


E-mail address: velling@umich.edu
E-mail address: joeprob@umich.edu
LOSS OF STRICT HYPERBOLICITY FOR VERTICAL THREE-PHASE FLOW IN POROUS MEDIA

PANTERS RODRIGUEZ-BERMUDEZ AND DAN MARCHESIN

Universidade Federal Fluminense (UFF)
Ave. dos Trabalhadores 420
Vila Santa Cecília, Volta Redonda, RJ, 27255-125, Brasil
and
Instituto Nacional de Matemática Pura e Aplicada (IMPA)
Estrada Doña Castorina 110
Rio de Janeiro, 22460-320, Brasil

Abstract. We study the hyperbolicity for systems of two conservation laws, which model vertical three-phase flow, where both gravity and convection due to pressure-gradient are considered. Besides the well known umbilic points found in the horizontal problem (without gravity), here we find new types of points where characteristic values coincide, located at the boundary of the saturation triangle. We characterize these points using Schaeffer-Shearer cones.

1. Introduction. Systems of conservation laws modeling immiscible three-phase flow in porous media typically fail to be strictly hyperbolic. This is the case for horizontal one-dimensional convective flow, where this failure occurs at four umbilic points: the three vertices of the saturation triangle and an interior point [1]. When gravity is considered, the situation can be even more complex. In this paper we study the hyperbolicity for systems of two conservation laws, which model vertical three-phase flow, where both gravity and convection due to pressure gradient are considered. We will split our analysis in two different cases. In the first one the convection/gravity ratio parameter \( \alpha \) (defined in next section) is equal to zero: we call this case “the pure gravitational problem”, because the motion of the flow occurs only as a consequence of gravitational buoyancy effects due to fluid density differences. In the second case \( \alpha \) is non zero. For high values of \( \alpha \) the problem would be similar to the case without gravity studied by Marchesin et al. in [1], while for small values of \( \alpha \) (i.e., for predominant gravitational effects) the problem presents features from both the pure gravitational problem and the problem without gravity.

The coincidence points found in this work play a relevant role in the Riemann solutions for vertical three-phase flow in porous media, see [3], [4], [5].

2000 Mathematics Subject Classification. Primary: 35L65, 70S05; Secondary: 35L80.

Key words and phrases. Conservation laws, Loss of Hyperbolicity, Three-Phase flow, Riemann problems, Buoyancy driven flow.

The first author was supported in part by CAPES/PNPD grant 03005/09-7, CNPq-TWAS Doctoral Fellowship program, PCI grant 680018/2008-5 and by IMPA institution. The second author was supported in part by PRH32 (ANP731948/2010, Petrobras 6000.0061847.10.4), FAPERJ-E-26/102.965/2011, FAPERJ-E-26/111.416/2010, CNPq-472923/2010-2 and CNPq-301564/2009-4.

The $2 \times 2$ system of conservation laws that models vertical three-phase flow for immiscible fluids in porous media has the form [3]

$$\frac{\partial s_i}{\partial t} + \frac{\partial}{\partial x} F_i = 0, \quad i = 1, 2,$$

where the components of the flux function $F = (F_1, F_2)^T$ are

$$F_i = \alpha f_i(s_1, s_2) + G_i(s_1, s_2).$$

The fractional flow functions $f_i$ corresponding to each fluid $i$ are defined as

$$f_i = \Lambda_i / \Lambda = \frac{s_i}{\mu_i},$$

where $\Lambda_i(s_i) = s_i^2 / \mu_i$ and $\Lambda = \Sigma_{i=1,2,3} \Lambda_i$, (3)

The gravitational terms $G_1$ and $G_2$ in (2) are given by

$$G_1 = K \Lambda_1((1 - f_1) \rho_{13} + f_2 \rho_{21}), \quad G_2 = K \Lambda_2((1 - f_2) \rho_{21} + f_3 \rho_{13}).$$

Notice that we restrict our analysis to Corey model with quadratic mobilities. In Eqs. (1)-(3) $s_i$ denotes the saturation of fluid $i$, while $\mu_i$ and $\rho_i$ are the corresponding viscosities and densities, respectively. We also denote by $K$ the absolute permeability of the rock. The non-dimensional parameter $\alpha$ has roughly the meaning of the ratio between pressure gradient-induced and gravity-induced flows, see [3]. In (4) we are denoting by $\rho_{ij}$ the density difference $\rho_i - \rho_j$ between the fluids $i$ and $j$. The system (1) was obtained after considering equations for conservation of mass together with Darcy’s law for velocities, taking into account that the porous medium is completely saturated ($s_1 + s_2 + s_3 = 1$). We are assuming that the flow occurs uniformly in the vertical direction $x$. In our convention, the gravitational force points to the positive $x$-direction.

3. General theory of coincidence points.

3.1. Types of coincidence points. A system of conservation law (1) is called strictly hyperbolic if the Jacobian matrix $dF(S)$ has real and distinct eigenvalues for all $S$ in state space. The hyperbolic character of the system can be lost in different ways, in particular we say that $S$ is a coincidence point for the flux $F$ if the eigenvalues of the Jacobian matrix $dF$ coincide. A coincidence point $S^*$ is called an umbilic point of the PDE system (with flow function given by $F$), if it satisfies the following conditions: (H1) $dF(S^*)$ is diagonalizable, and (H2) there is a neighborhood $V$ of $S^*$ such that $dF(S)$ has distinct eigenvalues for all $S \in V - S^*$. The following definitions were stated in [3], [4]:

**Definition 3.1.** (i) We say that a coincidence point $S$ is a quasi-umbilic point if condition (H2) holds but (H1) fails at $S$. (ii) We say that a coincidence point $S$ belongs to a coincidence diagonalization curve if there exists a curve of coincidence points through $S$ along which condition (H1) holds. If such curve is a straight line, we call it a coincidence diagonalization line.

3.2. Hyperbolicity analysis. Following Schaeffer and Shearer [6], for a real $2 \times 2$ matrix $M$ let us define

$$\text{dev}M = M - (1/2)(trM)I$$

as the projection of $M$ onto the space of trace-free matrices. In (5), $\text{dev}$ is known as the deviator operator. The deviator of a matrix $M$ retains all information about
multiplicity of eigenvalues and eigenvectors. We introduce coordinates $X, Y, Z$ in the (three dimensional) space of trace-free matrices by the formula

$$devM = \begin{pmatrix} X & Y + Z \\ Y - Z & -X \end{pmatrix}.$$  

(6)

The following important result for the deviator of a matrix $M$ was proved in [6]:

**Proposition 1.** We have the following: (i) $M$ has equal eigenvalues and is diagonalizable if and only if $devM = 0$. (ii) $M$ has distinct real eigenvalues, coincident real eigenvalues, or complex conjugate eigenvalues according to whether $(X, Y, Z)$ lies outside, on the surface of, or inside the cone $X^2 + Y^2 \leq Z^2$, respectively.

**Application of the deviator operator to the Jacobian matrix.** The mapping $dev(dF)$ from the saturation two-dimensional $S$-domain into the space of trace-free matrices is given by $S = (s_1, s_2) \mapsto (X, Y, Z)$; geometrically it defines a surface in $\mathbb{R}^3$. The following proposition was also proved by Schaeffer and Shearer [6].

**Proposition 2.** Let $S$ be a coincidence point, condition $(H2)$ is satisfied if and only if the surface that is the image of a subset of $\mathbb{R}^2$ by $dev(dF)$ is nonsingular at $S$ and, the punctured tangent surface at $S$ lies in the open region $\{X^2 + Y^2 > Z^2\}$.

Condition $(H1)$ implies that the umbilic points are mapped into the origin, the vertex of the cone $X^2 + Y^2 \leq Z^2$. Let $S^*$ be an umbilic point; because of condition $(H2)$ and Prop. 2 the image by $dev(dF)$ of a punctured neighborhood $V - S^*$ must lie outside this cone. Furthermore this surface is nonsingular at $S^*$ and the punctured tangent plane lies in the region $X^2 + Y^2 > Z^2$, see Fig. 1(a). A quasi-umbilic point $Q$ will be mapped on the surface of the cone $X^2 + Y^2 = Z^2$ away of the vertex. Since $(H2)$ holds at $Q$, the surface image by $dev(dF)$ of a neighborhood of $Q$ is not singular in $Q$ and lies outside the cone, see Fig. 1(b).

**Figure 1.** a) $dev \ dF$ surface in a neighborhood of an umbilic point. b) $dev \ dF$ surface in a neighborhood of a quasi-umbilic point. c) $dev \ dF$ surface in a neighborhood of any point along a coincidence diagonalization curve, all the points of this curve are mapped into the vertex of the cone.

From Propositions 1 and 2 we see that all the points of a coincidence diagonalization curve are mapped into the vertex of the cone; furthermore the surface defined by $dev \ dF$ in a neighborhood of each point $S$ of the coincidence curve has a singularity at the origin as shown in Fig. 1(c).
4. Hyperbolicity analysis for vertical three-phase flow. Along this section we will denote by \( V \) the vertex of the saturation triangle \( T \) that represents the pure state \( s_i = 1 \) for the fluid \( i \). On the other hand we will denote by \( \partial T \) the edge of \( T \) opposite to \( V \), which represents a mixture of fluids without fluid \( i \).

Applying the deviator operator in Eq. (5) to the Jacobian matrix \( dF \) (see the formulas for it coefficients \( J_{ij} \) in Chap. 4 of [3]) we obtain a parametrization for the surface \( \text{dev}(dF) \)

\[
\text{dev}(dF) = (X(s_1, s_2), Y(s_1, s_2), Z(s_1, s_2)),
\]

where \( X, Y, Z \) are given below, remembering that \( s_3 = 1 - s_1 - s_2 \):

\[
X = \left( \frac{\alpha s_1^2 s_3}{\mu_1 \mu_2} + \frac{\alpha s_1^2 s_3}{\mu_1 \mu_3} - \frac{\alpha s_2^2 s_3}{\mu_2 \mu_3} - \frac{\alpha s_1 s_3^2}{\mu_1 \mu_2} + \frac{\alpha s_3^2}{\mu_1 \mu_3} + \frac{s_1 s_3^2}{\mu_1 \mu_2} + \frac{s_1 s_2}{\mu_1 \mu_3} + \frac{s_3^2}{\mu_2 \mu_3} \right) / \Lambda^2,
\]

\[
Y = \left( \frac{\alpha s_1^2 s_3}{\mu_1 \mu_2} + \frac{\alpha s_1^2 s_3}{\mu_1 \mu_3} - \frac{\alpha s_2^2 s_3}{\mu_2 \mu_3} - \frac{\alpha s_1 s_3^2}{\mu_1 \mu_2} + \frac{\alpha s_3^2}{\mu_1 \mu_3} + \frac{s_1 s_3^2}{\mu_1 \mu_2} + \frac{s_3^2}{\mu_2 \mu_3} \right) / \Lambda^2,
\]

\[
Z = \left( \frac{\alpha s_1^2 s_3}{\mu_1 \mu_2} + \frac{\alpha s_1^2 s_3}{\mu_1 \mu_3} - \frac{\alpha s_2^2 s_3}{\mu_2 \mu_3} - \frac{\alpha s_1 s_3^2}{\mu_1 \mu_2} + \frac{\alpha s_3^2}{\mu_1 \mu_3} + \frac{s_1 s_3^2}{\mu_1 \mu_2} + \frac{s_3^2}{\mu_2 \mu_3} \right) / \Lambda^2.
\]

4.1. The purely gravitational problem. Throughout this subsection we set \( \alpha = 0 \). We first analyze a simplified case in which two fluids (1 and 2) have equal densities while the third fluid (3) has a different density.

Proposition 3. (Simplified purely gravitational problem) Assume that \( \alpha = 0 \), \( \rho_1 = \rho_2 \neq \rho_3 \), and denote \( \rho = \rho_{13} = \rho_{23} \). Then the system (1) is hyperbolic on the saturation triangle and all the coincidence points lie on its boundary. The vertex \( V_3 \) is an umbilic point. The entire edge \( \partial T \) is a coincidence diagonalization line; also there exist two quasi-umbilic points \( Q_1 \in \partial Q, Q_2 \in \partial Q \) given in (17), (18). The system is strictly hyperbolic in the rest of the closed saturation triangle.

Proof. Substituting \( \alpha = 0 \) and \( \rho = \rho_{13} = \rho_{23} \) in equations (8)-(10) we obtain

\[
X = \rho \left( \frac{s_1 s_2}{\mu_1 \mu_2 \mu_3} + \frac{s_3^2}{\mu_1 \mu_3} + \frac{s_2^2}{\mu_2 \mu_3} - \frac{s_1^2}{\mu_1 \mu_2} - \frac{s_3^2}{\mu_1 \mu_3} - \frac{s_2^2}{\mu_2 \mu_3} \right) / \Lambda^2,
\]

\[
Y = \rho \left( - \frac{s_3^2}{\mu_1 \mu_2 \mu_3} + \frac{s_2^2}{\mu_2 \mu_3} - \frac{s_1^2}{\mu_1 \mu_2} - \frac{s_3^2}{\mu_1 \mu_3} - \frac{s_2^2}{\mu_2 \mu_3} - 2 \frac{s_2^2}{\mu_1 \mu_2 \mu_3} \right) / \Lambda^2,
\]

\[
Z = \rho \left( \frac{s_3^2}{\mu_1 \mu_3} + \frac{s_2^2}{\mu_2 \mu_3} - \frac{s_1^2}{\mu_1 \mu_2} - \frac{s_3^2}{\mu_1 \mu_3} - \frac{s_2^2}{\mu_2 \mu_3} \right) / \Lambda^2.
\]
From (11) we have

\[ Y + Z = \rho \frac{s_1^2 s_3}{\mu_1 \mu_3 \lambda^2} \left( -2 \frac{s_1^2}{\mu_1} - 2 \frac{s_2^2}{\mu_2} - 2 \frac{s_2 s_3}{\mu_2} \right), \]

\[ Y - Z = \rho \frac{s_2^2 s_3}{\mu_2 \mu_3 \lambda^2} \left( -2 \frac{s_2^2}{\mu_2} - 2 \frac{s_1^2}{\mu_1} - 2 \frac{s_1 s_3}{\mu_1} \right). \]  

\hfill (12)

We notice that \((Y + Z)/\rho \leq 0\) and \((Y - Z)/\rho \leq 0\), therefore

\[ Y^2 - Z^2 \geq 0 \quad \forall S \in T. \]  

\hfill (13)

Thus \(X^2 + Y^2 \geq Z^2\) in \(T\), showing that there does not exist an elliptic region (i.e., with complex conjugate eigenvalues), in other words, system (1) is hyperbolic in the saturation triangle.

We recall from Sec. 3.2 that \(S_0\) is a coincidence point for the characteristic speeds if and only if

\[ X^2(S_0) + Y^2(S_0) - Z^2(S_0) = 0. \]  

\hfill (14)

From (13)-(14) we obtain the following necessary and sufficient conditions for \(S_0\) to be a coincidence point:

(i) \(X(S_0) = 0\) together with either (ii) \(Y(S_0) = -Z(S_0)\) or \(Y(S_0) = Z(S_0)\).

Now for \(S_0 = (s_1^0, s_2^0, s_3^0)\), using (12) we see that

\[ Y(s_1^0, s_2^0, s_3^0) = -Z(s_1^0, s_2^0, s_3^0) \iff s_1^0 = 0 \quad \text{or} \quad s_3^0 = 0, \]

\[ Y(s_1^0, s_2^0, s_3^0) = Z(s_1^0, s_2^0, s_3^0) \iff s_2^0 = 0 \quad \text{or} \quad s_3^0 = 0. \]

\hfill (15)

We notice that the edge \(\partial 3\) is mapped by \(\text{dev}(dF)\) into the vertex of the cone, therefore the entire edge \(\partial 3\), including the vertices \(V_1\) and \(V_2\) of the saturation triangle, is a coincidence diagonalization line according to Prop. 1 and Eqs. (11). For \(S_0\) out of the edge \(\partial 3\) we have two cases. If \(s_3^0 > 0\), \(s_1^0 = 0\) we obtain two other coincidence points on the edge \(\partial 1\) by requiring that \(X(s_1^0, s_2^0, s_3^0) = 0\). One of them is the vertex \(V_3 = (0, 0, 1)\), which is an umbilic point (notice that \(X(0, 0, 1) = Y(0, 0, 1) = Z(0, 0, 1) = 0\), the fact that \(V_3\) is an isolated coincidence point can be obtained in a similar way to the one used for \(Q_1\) below). Now we define

\[ a(\mu_i, \mu_j) = \left( \sqrt[3]{\mu_i/\mu_j} \right) \left( 1 + \sqrt[3]{\mu_i/\mu_j} \right)^{-1}, \quad \forall i \neq j. \]

\hfill (16)

The other coincidence point

\[ Q_1 = (0, a(\mu_2, \mu_3), 1 - a(\mu_2, \mu_3)) \]

is mapped by \(\text{dev}(dF)\) onto the cone surface out of the vertex, so according to Prop. 1, the Jacobian matrix \(dF(Q_1)\) is non diagonalizable. Of course \(Q_1\) is a boundary coincidence point, which is isolated if we restrict our analysis to the saturation triangle. Nevertheless, in order to classify \(Q_1\) as a quasi-umbilic point, we prove that \(Q_1\) is an isolated coincidence point in a whole neighborhood of \(Q_1\), so we extend our hyperbolicity analysis for points with a negative first component, i.e., to allow infinitesimal “negative saturations” for fluid 1. To do that, we extend analytically the flux function (2)-(4) to negative saturations and saturations slightly greater than 1. From (12), we see that if a point \(S \neq Q_1\) belongs to a sufficiently small neighborhood of \(Q_1\) (and therefore its first component \(s_1^0\) could be negative but close to zero) we have that \((Y(S_e) + Z(S_e))/\rho < 0\) and \((Y(S_e) - Z(S_e))/\rho < 0\). Thus we obtain that \(X^2(S_e) + Y^2(S_e) + Z^2(S_e) > 0\), concluding that \(Q_1\) is an isolated coincidence point. Thus \(Q_1\) is a quasi-umbilic point.
For the case $s_3^0 > 0$, $s_2^0 = 0$ and $s_1^0 > 0$ we obtain in a similar way the point on the edge $\partial 2$

\[ Q_2 = \left( a(\mu_1, \mu_3), 0, 1 - a(\mu_1, \mu_3) \right), \] (18)

which is quasi-umbilic point.

For the case in which the three fluids have distinct densities we have

**Proposition 4.** Assume $\alpha = 0$, $\rho_1 \neq \rho_2 \neq \rho_3 \neq \rho_1$. The system (1) is hyperbolic on the saturation triangle; all the coincidence points are isolated and lie on its boundary. The vertices $V_1$, $V_2$ and $V_3$ are umbilic points. There exist three quasi-umbilic points $Q_i \in \partial 1$, $i = 1, 2, 3$, given in (17), (18) and (25). The system is strictly hyperbolic in the rest of the closed saturation triangle.

**Proof.** Strict hyperbolicity in the interior of the saturation triangle was proved by Medeiros [2]. Now we will analyze hyperbolicity at the boundary of the triangle.

Setting $\alpha = 0$ in equations (8)-(10) we obtain the following relations

\[ Y + Z = 2 \left( \frac{\hat{s}_1^2 \rho_{12}}{\mu_1 \mu_2 \mu_3} - \frac{s_1^2 \hat{s}_2 \rho_{13}}{\mu_1 \mu_2 \mu_3} - \frac{s_2^2 \hat{s}_3 \rho_{23}}{\mu_1 \mu_2 \mu_3} - \frac{s_3^2 \hat{s}_1 \rho_{13}}{\mu_1 \mu_2 \mu_3} \right)/\lambda^2, \] (19)

\[ Y - Z = -2 \left( \frac{\hat{s}_1^2 \hat{s}_2 \rho_{13}}{\mu_1 \mu_2 \mu_3} + \frac{s_1^2 \hat{s}_3 \rho_{12}}{\mu_1 \mu_2 \mu_3} + \frac{s_2^2 \hat{s}_3 \rho_{23}}{\mu_1 \mu_2 \mu_3} + \frac{s_3^2 \hat{s}_1 \rho_{12}}{\mu_1 \mu_2 \mu_3} \right)/\lambda^2, \] (20)

\[ X - Z = \left( \frac{s_1^2 \hat{s}_2 \rho_{13}}{\mu_1 \mu_2 \mu_3} + \frac{s_1 \hat{s}_3 \rho_{12}}{\mu_1 \mu_2 \mu_3} + \frac{s_2 \hat{s}_3 \rho_{23}}{\mu_1 \mu_2 \mu_3} - \frac{s_3 \hat{s}_1 \rho_{12}}{\mu_1 \mu_2 \mu_3} \right)/\lambda^2. \] (21)

As $\hat{S} = (0, \hat{s}_2, 1 - \hat{s}_2) \in \partial 1$, from (19) we have $Y(\hat{S}) + Z(\hat{S}) = 0$. Thus $Y^2(\hat{S}) = Z^2(\hat{S})$ and from (14) $\hat{S}$ is a coincidence point if and only if $X(\hat{S}) = 0$. From (8) we have

\[ X(\hat{S}) = -\frac{\hat{s}_2 (1 - \hat{s}_2) \rho_{23}}{\mu_2 \mu_3 \Lambda(\hat{S})^2} \left( \frac{(1 - \hat{s}_2)^3}{\mu_3} - \frac{\hat{s}_2^3}{\mu_2} \right); \] (22)

setting $X(\hat{S}) = 0$ we obtain three coincidence points on the edge $\partial 1$, which are $V_2$, $V_3$ and $Q_1$ given in (17). The points $V_2$, $V_3$ are umbilic because they are isolated coincidence points that are mapped into the vertex of the cone by $\text{dev}(dF)$, see Prop. 1. The point $Q_1$ is again quasi-umbilic because it is mapped by $\text{dev}(dF)$ onto the cone surface out of the vertex. The fact that $Q_1$ is an isolated coincidence point follows from (19)-(20), since $(Y(S_c) + Z(S_c))/\rho_{23} < 0$ and $(Y(S_c) - Z(S_c))/\rho_{23} < 0$ for all point $S_c$ in a sufficiently small neighborhood of $Q_1$, which imply $X^2(S_c) + Y^2(S_c) + Z^2(S_c) > 0$.

For $\hat{S} = (\hat{s}_1, 0, 1 - \hat{s}_1) \in \partial 2$, from (20) we have $Y(\hat{S}) - Z(\hat{S}) = 0$ and therefore $\hat{S}$ is again a coincidence point if and only if $X(\hat{S}) = 0$, where

\[ X(\hat{S}) = \frac{\hat{s}_1 (1 - \hat{s}_1) \rho_{13}}{\mu_1 \mu_3 \Lambda(\hat{S})^2} \left( \frac{(1 - \hat{s}_1)^3}{\mu_3} - \frac{\hat{s}_1^3}{\mu_1} \right); \] (23)

as in the previous case three coincidence points are obtained: the umbilic points $V_1$, $V_3$ and the quasi-umbilic point $Q_2$ in (18) on the edge $\partial 2$.

If $\hat{S} = (\hat{s}_1, 1 - \hat{s}_1, 0) \in \partial 3$, from (21) we have $X(\hat{S}) - Z(\hat{S}) = 0$, and $X^2(\hat{S}) = Z^2(\hat{S})$. Therefore $\hat{S}$ is a coincidence point if and only if $Y(\hat{S}) = 0$. From (9) we have

\[ Y(\hat{S}) = -\frac{\hat{s}_1 (1 - \hat{s}_1) \rho_{12}}{\mu_1 \mu_2 \Lambda(\hat{S})^2} \left( \frac{(1 - \hat{s}_1)^3}{\mu_2} - \frac{\hat{s}_1^3}{\mu_1} \right); \] (24)
in this case, we obtain the umbilics $V_1, V_2$ and the quasi-umbilic point $Q_3$ on $\partial 3$

$$Q_3 = (a(\mu_1, \mu_2), 1 - a(\mu_1, \mu_2), 0), \quad (25)$$

where $a(\mu_1, \mu_2)$ is given in (16).


4.2. Hyperbolicity analysis for the general gravitational problem. We analyze hyperbolicity when the convection/gravity ratio $\alpha$ is non-zero. As we will see for values of $|\alpha|$ so high that the convection dominates gravity, there are only four coincidence points in the saturation triangle, the three vertices and an additional interior point $S_\alpha^*$. As in the horizontal problem [1], these points are umbilic [2]. For small non-zero values of $|\alpha|$, besides the four umbilic points there exists at most one quasi-umbilic points at each edge of $T$. In summary:

**Proposition 5.** Consider the system (1) with $\alpha \neq 0$, in the saturation triangle. Then the following assertions are true:

(i) The vertices of the saturation triangle are umbilic points. In the interior of the triangle the system is strictly hyperbolic except at the umbilic point $S_\alpha^*$.

(ii) Without loss of generality, consider the two-phase edge $\partial 1$ where fluids 2 and 3 coexist. Assume also $\rho_2 > \rho_3$, then there exists a quasi-umbilic point $Q_1 \in \partial 1$ if and only if $\alpha$ lies in the interval $(-\frac{\rho_2}{\mu_3} < \alpha < \frac{\rho_3}{\mu_2})$. The system is strictly hyperbolic in the rest of the edge $\partial 1$.

(iii) In the case $\rho_2 = \rho_3$ all points on the edge $\partial 1$ are strictly hyperbolic, except for the vertices $V_2, V_3$.

**Remark 1.** Prop. 5(ii),(iii) holds for any permutation of indices $\{1, 2, 3\}$.

**Remark 2.** When $\rho_2 = \rho_3$, in the case $\alpha \neq 0$ we have strict hyperbolicity on the edge $\partial 1$ except for the vertices $V_2, V_3$. This behavior differs from the case $\alpha = 0$ where $\partial 1$ is a diagonalization line (see Prop. 3 with the indices 1 and 3 interchanged).

**Proof.** (i)- It is obvious that the vertices $V_1, V_2$ and $V_3$ are umbilic points since

$$X(V_i) = Y(V_i) = Z(V_i) = 0, \quad i = 1, 2, 3,$$

they are isolated coincidence points which can be obtained in analogous way as we proceeded in the previous theorems. The existence of the interior umbilic point $S_\alpha^*$ and the strict hyperbolicity of the remaining interior points for the case $\alpha \neq 0$ was proved by Medeiros [2].

(ii)- Taking into account the formula for the Jacobian matrix in [3], the characteristic values at a state $S \in \partial 1$ are

$$\lambda(S) = 0 \quad \text{and} \quad \lambda^*(S) = 2 \frac{s_2(1 - s_2)}{\mu_2 \mu_3 \Lambda^2} \left( - \frac{s_3^3}{\mu_2} + \left( \frac{1 - s_2}{\mu_3} \right)^3 \rho_{23} + \alpha \right). \quad (26)$$

Here the eigenvalues $\lambda$ and $\lambda^*$ are not associated to specific families since $\lambda^*$ can be positive or negative. The important fact is that $S \in \partial 1$ will be a coincidence point if and only if $\lambda^*(S) = 0$; this occurs on this edge at the vertices $V_2, V_3$ and at the intermediate point $Q_1^\alpha = (0, \rho_a, 1 - \rho_a)$ where $\rho_a$ solves

$$\left( \frac{\rho_a^3}{\mu_2} - (1 - \rho_a)^3 / \mu_3 \right) \rho_{23} = \alpha. \quad (27)$$

From (27) we have that $Q_1^\alpha \in T$ if and only if

$$- \rho_{23} / \mu_3 < \alpha < \rho_{23} / \mu_2. \quad (28)$$
Now we will prove under condition (28) that $Q_1^0$ is a quasi-umbilic point. As in Prop. 4 one can show that $Q_1^0 \in \partial \Omega$ is an isolated coincidence point. Next we will prove that it is not an umbilic point, by contradiction. Assume that $Q_1^0$ is an umbilic point, then necessarily $X(Q_1^0) = Y(Q_1^0) = Z(Q_1^0) = 0$. From (9) we have

$$Y(Q_1^0) = q_0^2(1 - q_0)(\alpha - q_0^2 \rho_{23}/\mu_2)/\mu_2 \mu_3 (\Lambda(Q_1^0))^2.$$  \tag{29}

Equations (27) and (28) imply $0 < q_0 < 1$, so we have from (29) that $\alpha = q_0^2 \rho_{23}/\mu_2$.

Substituting this value into the definition (27) for $q_0$, we obtain

$$(1/\mu_2)q_0^2(1 - q_0) = -(1/\mu_3)(1 - q_0)^3.$$ \tag{30}

The fact $0 < q_0 < 1$ contradicts (30). We conclude that $Q_1^0$ is a quasi-umbilic point because $Q_1^0$ is an isolated coincidence point that is not mapped to the cone vertex.

(iii)- For the case $\rho_2 = \rho_3$ we can see from (8)-(10) that for $S$ in $\partial \Omega$

$$X(S) = -\frac{\alpha s^2}{\mu_3}/\Lambda^2, \quad Y(S) = \frac{\alpha s^2}{\mu_3}/\Lambda^2, \quad Z(S) = -\frac{\alpha s^2}{\mu_3}/\Lambda^2,$$ \tag{31}

where $\Lambda = \Lambda(S)$. For $S$ distinct from $V_2$, $V_3$ we have from (31) that $X(S)$ is non-zero. We also have $Y(S) + Z(S) = 0$, therefore $X^2(S) + (Y^2(S) - Z^2(S)) > 0$. Hence the strict hyperbolicity along $\partial \Omega$ except at the vertices $V_2$ and $V_3$.

We give an implicit formula for the umbilic point $S_\alpha^*$ in Prop. 5(i). Requiring that $X(S_\alpha^*) = Y(S_\alpha^*) = Z(S_\alpha^*) = 0$ at $S_\alpha^*$, we obtain even for non-quadratic mobilities

$$\alpha(\Lambda_2' - \Lambda_3') + [\Lambda_1 \Lambda_2' \rho_{21} + (\Lambda_2 \Lambda_3' + \Lambda_1 \Lambda_3') \rho_{23} + \Lambda_1 \Lambda_2 \rho_{13} - \Lambda_1 \Lambda_3 \rho_{13}] = 0,$$ \tag{32}

$$\alpha(\Lambda_1' - \Lambda_3') + [\Lambda_2 \Lambda_3' \rho_{12} + (\Lambda_1 \Lambda_3' + \Lambda_2 \Lambda_3') \rho_{13} + \Lambda_2 \Lambda_1 \rho_{23}] = 0,$$ \tag{33}

$$\alpha(\Lambda_1' - \Lambda_2') + [\Lambda_1 \Lambda_3 \rho_{13} + (\Lambda_2 \Lambda_1' + \Lambda_3 \Lambda_2') \rho_{12} + \Lambda_3 \Lambda_2 \rho_{23}] = 0.$$ \tag{34}

**Remark 3.** Equations (32)-(34) represent curves crossing the umbilic point $S_\alpha^*$. Only two of these equations are independent.

**Acknowledgments.** We would like to thank the professors Dr. Vitor Matos from Universidade do Porto and Dr. Frederico Furtado from University of Wyoming for valuable discussions.

**REFERENCES**


E-mail address: pantersrb@id.uff.br, panters@impa.br

E-mail address: marchesi@impa.br
ON THE MANAGEMENT OF 
VEHICULAR AND PEDESTRIAN FLOWS

RINALDO M. COLOMBO
University of Brescia
via Branze 38, 25123
Brescia, Italy

PAOLA GOATIN
INRIA Sophia Antipolis-Méditerranée
route des Lucioles-BP 93
Sophia Antipolis Cedex, France

MASSIMILIANO D. ROSINI
ICM, University of Warsaw
ul. Prosta 69, 00-838
Warsaw, Poland

Abstract. In this paper optimization problems arising from vehicular traffic and crowd dynamics are considered from the analytical point of view.

1. Introduction. Several realistic situations in vehicular traffic and crowd dynamics can be modeled through conservation laws with boundary and unilateral constraints on the flux \[29\]. This paper provides an overview of rigorous analytical frameworks for these descriptions to rigorously state optimal management problems and prove the existence of the corresponding optimal controls. Specific cases are dealt with in detail through ad hoc numerical integrations. These are here obtained implementing the wave-front tracking algorithm, which appears to be very precise in computing, for instance, the exit times.

2. Vehicular traffic.

2.1. Introduction. The LWR model \[26, 27\] represents the starting point for the modeling of vehicular flows. Its basic assumptions are that there is only one class of vehicles moving along a unique homogeneous lane, overtaking is not allowed, the vehicles do not enter or exit the road and the speed of the vehicles depends on the density alone. The resulting model is represented by the scalar conservation law

\[ \rho_t + f(\rho)_x = 0 \]

where \( \rho \) represents the (mean) density, \( f(\rho) = \rho \ v(\rho) \) the flow and \( v = v(\rho) \) the (mean) velocity. The map \([\rho \mapsto v(\rho)]\) is a decreasing non-negative Lipschitz function

2000 Mathematics Subject Classification. Primary: 35L65; Secondary: 49J20, 90B20.

Key words and phrases. Conservation laws, vehicular traffic, crowd dynamics, optimal management strategy.

The third author is partially supported by the Organizing Committee of HYP2012; ICM (Interdisciplinary Centre for Mathematical and Computational Modelling), University of Warsaw; Narodowe Centrum Nauki, grant 4140; and Polonium 2011 (French-Polish cooperation program) under the project “CROWd Motion Modeling and Management”.

889
for $\rho$ between 0 and some positive maximal density $R$, which corresponds to a total traffic jam, so that $1/R$ is the (mean) vehicle length. More precisely, we assume

(F): $f \in C^{0,1}([0, R]; [0, \bar{f}])$ satisfies $f(0) = f(R) = 0$ and there exists a $\bar{\rho} \in ]0, R[$ such that $f'(\rho)(\bar{\rho} - \rho) > 0$ for a.e. $\rho \in [0, R[$.

The LWR model can be easily generalized as follows. The evolution of traffic flowing along a highway with a given entry, say sited at $x = 0$, can be described by (1) supplemented with an initial datum $\rho_0$ and by an inflow $q_0$ at the entry. Furthermore, often traffic flow is subject to various restrictions, such as traffic lights, toll gates, construction sites, or the effects of accidents. All these situations amount to limit the flow at specific locations along the road and, hence, can be described by adding suitable unilateral constraints on the flow. The resulting system is then

$$\begin{align*}
\partial_t \rho + \partial_x f(\rho) &= 0 \quad \text{for all } \phi \in C^1_c((0,R] \times \mathbb{R}_+^d) \\
\rho(0,x) &= \rho_0(x) \quad x \in \mathbb{R}_+ \quad (2a) \\
f(\rho(t,0)) &= q_0(t) \quad t \in \mathbb{R}_+ \quad (2b) \\
f(\rho(t,x_c)) &\leq q_c(t) \quad t \in \mathbb{R}_+ . \quad (2d)
\end{align*}$$

In (2d), $x = x_c$ is the location of the constraint and $q_c = q_c(t)$ the possibly time dependent maximal flow through $x = x_c$ allowed by the constraint. A rigorous definition of solution to (2) is obtained by suitably merging the definition of solution to an initial–boundary value problem with that of a solution to a scalar conservation law with a unilateral constraint law, see [1, 2, 4, 6, 7, 9, 14, 30, 11].

**Definition 2.1 ([15]).** A map $\rho \in C^0([0,R]; L^\infty([0,R]))$ is a weak entropy solution to (2) if the following conditions hold:

1. for every test function $\phi \in C^1_c([0,R] \times \mathbb{R}_+)\setminus[0, R]$\] and for every $k \in [0, R]$

$$\begin{align*}
&\int_0^{+\infty} \int_0^{+\infty} \left[ |\rho - k| \, \partial_t \phi + \text{sgn}(\rho - k) \, (f(\rho) - f(k)) \, \partial_x \phi \right] \, dx \, dt \\
&\quad + \int_0^{+\infty} |\rho_0 - k| \, \phi(0,x) \, dx \\
&\quad + \int_0^{+\infty} \text{sgn} \left( f^{-1}_x (q_b(t)) - k \right) \, |f(\rho(t,0+)) - f(k)| \, \phi(t,0) \, dt \\
&\quad + 2 \int_0^{+\infty} \left[ 1 - \frac{q_c(t)}{f(\bar{\rho})} \right] \, f(k) \, \phi(t,x_c) \, dt \geq 0 ;
\end{align*}$$

2. for a.e. $t \in \mathbb{R}_+$, $f(\rho(t,x_c-)) = f(\rho(t,x_c)) \leq q_c(t)$.

Above $f_*$ denotes the restriction of $f$ to $[0, \bar{\rho}]$, i.e. $f_* = f|_{[0,\bar{\rho}]}$, and $\rho(t,x_c \pm)$ are the measure theoretic traces, implicitly defined by

$$\begin{align*}
\lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \int_0^{+\infty} \int_{x_c + \epsilon}^{x_c + \epsilon} |\rho(t,x) - \rho(t,x_c +)| \, \phi(t,x) \, dx \, dt &= 0 \\
\lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \int_0^{+\infty} \int_{x_c - \epsilon}^{x_c} |\rho(t,x) - \rho(t,x_c -)| \, \phi(t,x) \, dx \, dt &= 0
\end{align*}$$

for all $\phi \in C^1_c([0,R] \times \mathbb{R})$. Remark that both traces at $x_c$ in 2. exist and are finite, by [6, Theorem 2.2]. The first two lines in 1. originate from the classical Kružkov definition [25, Definition 1] in the case of the Cauchy problem, i.e. with no boundary and no constraints. The third line is motivated by the boundary, following [7, (17)] or [30, (15.14)]. The latter line accounts for the constraint, as in [14, definitions 3.1}
and 3.2], see also [6, Definition 2.1]. For other equivalent formulations in the case of unilateral constraints, we refer to [6, Proposition 2.6].

As in [14, 31], introduce the map \( \Psi(\rho) = \text{sgn}(\rho - \bar{\rho}) [\bar{f} - f(\rho)] \) and the domain

\[
D = \{ \rho \in L^1(\mathbb{R}^+; [0, R]) : \Psi(\rho) \in BV(\mathbb{R}^+; \mathbb{R}) \}.
\]

(3)

**Theorem 2.2** ([15]). Let (F) hold. Assume moreover that

\[(R0): \rho_o \in D, \quad (Q): q_b, q_c \in BV(\mathbb{R}^+; [0, f]) \.
\]

Then, there exists a unique solution \( \rho = \rho(t; x; \rho_o, q_b, q_c) \) to (2) in the sense of Definition 2.1 and, for all \( t \in \mathbb{R}^+ \), \( \rho(t; \rho_o, q_b, q_c) \in D \). Moreover, if \( \rho_o, \rho_o' \) both satisfy (R0), \( q_b, q_b' \) and \( q_c, q_c' \) satisfy (Q), then the following Lipschitz estimate on the corresponding solutions \( \rho, \rho' \) holds, for every \( t > 0 \):

\[
\|\rho(t) - \rho'(t)\|_{L^1(\mathbb{R}^+; \mathbb{R})} \leq \|\rho_0 - \rho_0'\|_{L^1(\mathbb{R}^+; \mathbb{R})} + 2 \|q_b - q_b'\|_{L^1([0, t]; \mathbb{R})} + 2 \|q_c - q_c'\|_{L^1([0, t]; \mathbb{R})}.
\]

(4)

Moreover, if \( \rho_o, \rho_o' \equiv 0 \), there exist \( \tau_o, \tau_c \) such that \( \text{spt}(q_b) \cup \text{spt}(q_b') \subseteq [0, \tau_o] \) and

\[
\int_0^{\tau_c} f(\rho(t, x_c)) \, dt = \int_0^{\tau_c} q_b(t) \, dt \quad \text{and} \quad \int_0^{\tau_c} f(\rho'(t, x_c)) \, dt = \int_0^{\tau_c} q_b'(t) \, dt,
\]

(5)

then, for all \( x > x_c \) and \( t \geq 0 \), the following Lipschitz estimate holds:

\[
\|f(\rho(\cdot, x_c)) - f(\rho'(\cdot, x_c))\|_{L^1([0, \tau_o]; \mathbb{R})} \leq \|q_b - q_b'\|_{L^1([0, \min(\tau_o, t)]; \mathbb{R})} + 2 \|q_c - q_c'\|_{L^1([0, \min(\tau_o, t)]; \mathbb{R})}.
\]

(6)

**2.2. Cost functionals.** We now consider some reasonable cost functionals. We prove that their regularity, when computed along the solutions to (2), is sufficient to ensure the existence of an optimal management strategy.

**Stop & Go Waves:** A relevant criterion in the management of traffic dynamics is the minimization of start/stop phenomena since a more regular traffic flow reasonably reduces both the probability of accidents as well as pollution. Analytically, this leads to consider the functional

\[
J(\rho) = \int_0^T \int_{\mathbb{R}^+} p(x) \, d|\partial_x v(\rho)| \, dt
\]

measuring on the time interval \([0, T]\) the total variation \(|\partial_x v(\rho)|\) of the traffic speed \(v(\rho)\) weighted by a weight \(p(x) \in [0, 1]\), higher in more dangerous road segments. By (F), for any \( t \), as soon as \( \rho(t) \in BV \), also the function \([x \mapsto v(\rho(t, x))]\) is \(BV\) and its derivative \(\partial_x v(\rho)\) is a Radon measure on \(\mathbb{R}\). The inner integral in (7) is thus the integral of \(p\) with respect to the time dependent total variation measure \(|\partial_x v(\rho)|\).

The lower semicontinuity of the functional J in (7) follows from [13, Lemma 2.1].

**Travel Times:** From the point of view of drivers, key quantity is the time necessary to reach the destination \(x = x > 0\). Consider (2) with \( \rho_o \equiv 0 \) and \( \text{spt}(q_b) \subseteq [0, \tau_o] \). The total quantity of vehicles entering the road is thus \(Q_{in} = \int_0^{\tau_c} q_b(t) \, dt\), which is assumed to be finite and strictly positive. Then, the mean arrival time \([4, (5.3)]\) is

\[
T_a(\bar{x}) = \frac{1}{Q_{in}} \int_0^{+\infty} t f(\rho(t, \bar{x})) \, dt.
\]

(8)

The mean travel time between the points 0 and \(\bar{x}\) can then be easily computed:

\[
T_t(\bar{x}) = \frac{1}{Q_{in}} \int_0^{+\infty} t \left| f(\rho(t, \bar{x})) - f(\rho(t, 0)) \right| \, dt.
\]

(9)
Remark that evaluating the exit time \( T_a \). Consider an integral functional of the density

\[
\int_a^b \rho(t,x) w(t,x) \, dx \, dt
\]

for a time \( T > 0 \) and points \( b > a > 0 \). Examples of choices of the function \( \phi \) are:

- To have all vehicles travel at a speed as near as possible to an optimal speed \( \hat{v} \) along a given road segment \([a, b]\), for instance \( \phi(x) = (v(x) - \hat{v})^2 \).
- To maximize the traffic flow along \([a, b]\), choose \( \phi(x) = f(x) \).

As soon as the weight \( w \) and the cost function \( \phi \) are continuous, it is immediate to prove that \( F \) is also continuous. This, together with Theorem 2.2, through an application of Weierstraß Theorem, allows to prove the existence of choices of the initial/boundary data and/or of the constraint that optimize \( F \).

3. Crowd dynamics.

3.1. The CR model. The CR model [16, 17, 28] is the unique macroscopic model for pedestrian flows able to reproduce the fall in the efficiency of the exit, the rise of panic and the Braess’ paradox. From the physical point of view, its main assumption is recently experimentally confirmed in [21]. From the analytical point of view, this model is an example of a conservation law in which non–entropy solutions have a physical motivation and a global existence result for the Cauchy problem with large datum is available [17]. Furthermore, the availability of efficient numerical schemes [10, 11] developed ad hoc to approximate the solutions of the CR model makes it practically usable for simulations of real world starting from real data.

The CR model is expressed by a conservation law of the form

\[
\partial_t \rho + \partial_x f(\rho) = 0 ,
\]

where \( \rho \) is the crowd density and \( f \) the flow that satisfies the following properties:

- \( f \in W^{1,\infty}([0, R^*]; \mathbb{R}_+^*) \).
- \( f(\rho) = 0 \) if and only if \( \rho \in [0, R^*] \).
- The restrictions \( f|_{[0,R]} \) and \( f|_{[R,R^*]} \) are strictly concave down.
- \( \max \{ f(\rho) : \rho \in [0,R] \} > \max \{ f(\rho) : \rho \in [R,R^*] \} \).
- \( f \) has a local minimum at \( \rho = R \).
- \( f(R) < \min \{ f'(R^+) \, R, \, -f'(R^-) \, (R^* - R) \} \).
In the following, for all $\rho_1, \rho_2 \in [0, R^*]$ with $\rho_1 \neq \rho_2$, we denote the point $(\rho_1, f(\rho_1))$ on the graph $f = f(\rho)$ as $P_1(\rho_1)$ and segment with endpoints $P_1(\rho_1)$ and $P_1(\rho_2)$ as

$$s(\rho_1, \rho_2) = \left\{ (\rho, f) \in \mathbb{R}^2 : f = \frac{f(\rho_1) - f(\rho_2)}{\rho_1 - \rho_2} (\rho - \rho_1) + f(\rho_1) \right\}.$$

By (F.3)–(F.5), there exist a unique $R_M \in [0, R]$ and $R_M^* \in [R, R^*[ such that $f(R_M) = \max \{ f(\rho) : \rho \in [0, R] \}$ and $f(R_M^*) = \max \{ f(\rho) : \rho \in [R, R^*[$. Furthermore, by (F.6), there exists a unique $R_4 \in [R, R^*$ (respectively, $R_3 \in [0, R]$) such that the segment $s(0, R_4)$ (respectively, $s(R_1, R^*)$) intersects $f = f(\rho)$ at the point $P_1(R)$. Let us further introduce the auxiliary functions $\psi, \phi_\psi$ and $\phi_R$. First, let $\psi(R) = R$ and, for $\phi \neq R$, let $\psi(\rho)$ be such that the segment $s(\rho, \psi(\rho))$ is tangent to the graph of $f$ at $P_1(\psi(\rho))$. Assume that there exists only one couple $(R_T, R_T^*) \in [R_M, R] \times [R_M^*, R^*[ such that $\psi(R_T) = R_T^*$ and $\psi(R_T^*) = R_T$, i.e.:

$$\text{F.7} \quad f(\psi(R_T^*)) = -f(\psi(R_T)) (R^* - \psi(R_T)).$$

We now introduce $\phi_\psi : [0, R^*] \to [0, R^*]$. First, we define $\phi_\psi(R_T) = R_T$ and $\phi_\psi(R_T^*) = R_T^*$. If $\rho \in [0, R] \cup [R_T, \infty)$, then by (F.6) the segment $s(\rho, \psi(\rho))$ intersects the graph of $f$, say in $P_1(\phi_\psi(\rho))$. Let $R_2 = \phi_\psi(0)$ and $R_3 = \phi_\psi(R^*)$. If $\rho \in [R_T, R_2] \cup [R_3, R_T^*[$, then there exists a segment $s(\rho, \phi_\psi(\rho))$ that intersects $f = f(\rho)$ in $P_1(\rho)$. Finally, we define $\phi_\psi(\rho) = 0$ for all $\rho \in [R_2, R]$ and $\phi_\psi(\rho) = R^*$ for all $\rho \in [R, R_3]$. Whether it exists, let $s(\rho, \phi_R(\rho))$ be such that its intersection with $f = f(\rho)$ is $P_1(R)$. Let $R_1 = \phi_R(R^*)$ and $R_4 = \phi_R(0)$, then $\phi_R : [0, R] \cup [R_1, R^*] \to [0, R_1] \cup [R_4, R^*]$ is a continuous increasing function.

Finally, we concentrate our attention on the cases in which:

$$\text{F.8} \quad 0 < R_3 < R_T, 2R_2 > \psi(0) \text{ and } R_2^* < R_4 < R^*.$$  

$$\text{F.9} \quad [\rho \mapsto (\psi(\rho) - \rho)] \text{ is strictly decreasing in } [0, R].$$

We now define the Riemann solver $\mathcal{R}_{CR}$ associated to Riemann problems

$$\partial_t \rho + \partial_x f(\rho) = 0 \quad \rho(0, x) = \begin{cases} \rho^l \quad & \text{if } x < 0 \\ \rho^r \quad & \text{if } x \geq 0 \end{cases} \quad (10).$$

Introduce two thresholds $s$ and $\Delta s$ such that

$$0 < s < R_M, \quad \Delta s > s \quad \text{and} \quad R > s + \Delta s \geq \phi_\psi(s) > R_T > R - \Delta s. \quad (11).$$

The solution to (10) with $\rho^l, \rho^r \in [0, R^*]$ is selected through the following conditions:

$$\text{(R.1)}: \text{If } \rho^l, \rho^r \in [0, R], \text{ then } \mathcal{R}_{CR}[\rho^l, \rho^r] = \mathcal{R}_e[\rho^l, \rho^r], \text{ unless } \rho^l > \Delta s \text{ and } \rho^r > \Delta s. \text{ In this case, } \mathcal{R}_{CR}[\rho^l, \rho^r] \text{ consists of a non–entropic shock between } \rho^l \text{ and } \psi(\rho^r), \text{ followed by the entropy solution between } \psi(\rho^l) \text{ and } \rho^r.$$

$$\text{(R.2)}: \text{If } \rho^r \leq \rho^l, \text{ then } \mathcal{R}_{CR}[\rho^l, \rho^r] = \mathcal{R}_e[\rho^l, \rho^r].$$

$$\text{(R.3)}: \text{If } R \leq \rho^l < \rho^r \text{ or } \rho^l < R < \rho^r \text{ and the segment } s(\rho^l, \rho^r) \text{ does not intersect } f = f(\rho) \text{ for } \rho \neq R, \text{ then the solution is a shock between } \rho^l \text{ and } \rho^r.$$

$$\text{(R.4)}: \text{If } \rho^l < R < \rho^r \text{ and the segment } s(\rho^l, \rho^r) \text{ intersects } f = f(\rho) \text{ for } \rho \neq R, \text{ then } \mathcal{R}_{CR}[\rho^l, \rho^r] \text{ consists of a non–entropic shock between } \rho^l \text{ and a panic state, followed by a possibly null entropic wave. More precisely:}$$

$$\text{(a)}: \text{If } \rho^r \in [R, \psi(\rho^l)], \text{ then } \mathcal{R}_{CR}[\rho^l, \rho^r] \text{ consists of a non–entropic shock between } \rho^l \text{ and } \psi(\rho^r), \text{ followed by a decreasing rarefaction between } \psi(\rho^l) \text{ and } \rho^r.$$

$$\text{(b)}: \text{If } \rho^r \in [\psi(\rho^l), R^*], \text{ then } \mathcal{R}_{CR}[\rho^l, \rho^r] \text{ consists of a single non–entropic shock.}$$

Here $\mathcal{R}_e$ is the Riemann solver that selects the self similar entropy solutions $[9, 19]$.\]
namely, letting varying
compute the evacuation time. In particular, changing the position of the first door,
umber of Riemann problems, it is possible to construct a solution of (13) and
have any effect on the evacuation time.

It is also clear that the presence of a first door too close to the second one does not
model properly describe the Braess’ paradox for pedestrian flows. Furthermore, it
represented in Figure 1, left. Remarkably, there is an interval of values of

The Hughes model. Consider the one dimensional version of the Hughes model [23, 24] for pedestrians evacuating a corridor \( \Omega = ]-1,1[ \) with two exits

\[
\rho_t - \left[ \rho v(\rho) \frac{\phi_x}{|\phi_x|} \right]_x = 0 , \quad |\phi_x| = c(\rho) , \tag{14}
\]

together with homogeneous Dirichlet boundary conditions and initial datum

\[
\rho(t,-1) = \rho(t,1) = 0 , \quad \phi(t,-1) = \phi(t,1) = 0 , \quad \rho(0,x) = \rho_o(x) . \tag{15}
\]

Above \( \rho = \rho(t,x) \in [0,1] \) is the (normalized) crowd density, \( v(\rho) = 1 - \rho \) the (mean) velocity, \( f(\rho) = \rho v(\rho) \) the flow and \( c = c(\rho) \) the running cost. We assume that \( c : [0,1] \to [1,+\infty] \) is a smooth function such that \( c(0) = 1 \) and \( c'(\rho) \geq 0 \). As
already observed in \[3, 20\], the system (14) can be rewritten as
\[
\begin{align*}
\rho_t + F(t, x, \rho)_x &= 0, \quad (16a) \\
\int_{-1}^{\xi(t)} c(\rho(t, x)) \, dx &= \int_{\xi(t)}^{1} c(\rho(t, x)) \, dx, \quad (16b)
\end{align*}
\]
where \(F(t, x, \rho) = \text{sgn}(x - \xi(t)) \, f(\rho)\) and \(x = \xi(t)\) is the so called turning curve implicitly defined by (16b).

It is possible to consider the problem of how to optimize the evacuation time by choosing a proper running cost \(c\). In fact, \(c\) can be seen as the strategy pedestrians choose or are forced to use to evacuate the corridor, and therefore can be selected depending on the specific situation. In \[20\] is proved that does not exist a cost function that optimizes the evacuation time for any initial datum. However, see Figure 1, right, a cost function which optimizes the evacuation time in the case of “high” densities is
\[
c_o(\rho) = \begin{cases} 
1 & \text{if } \rho < 1/2, \\
2\rho & \text{if } \rho \geq 1/2.
\end{cases}
\quad (17)
\]

Acknowledgments. The third author is partially supported by ICM, University of Warsaw, Narodowe Centrum Nauki, grant 4140, Polonium 2011 (French?Polish cooperation program) under the project “CROWd Motion Modeling and Management” and the Organizing Committee of HYP2012.

REFERENCES


C. Chalons, P. Goatin and N. Seguin General constrained conservation laws. application to pedestrian flow modeling, Networks and Heterogeneous Media, 8 (2013), 433–463.

E. M. Cepolina Phased evacuation: An optimisation model which takes into account the capacity drop phenomenon in pedestrian flows, Fire Safety Journal, 44 (2009), 532–544.


S. N. Kružhkov, First order quasilinear equations with several independent variables, Mat. Sb. (N.S.), 81 (1970), 228–255.


*E-mail address*: rinaldo@ing.unibs.it

*E-mail address*: paola.goatin@inria.fr

*E-mail address*: mrosini@icm.edu.pl
EXACT SOLUTIONS WITH SINGULARITIES TO IDEAL HYDRODYNAMICS OF INELASTIC GASES

OLGA ROZANOVA
Mechanics and Mathematics Faculty, Moscow State University
Moscow, 119992, Russia

Abstract. We construct a large family of exact solutions to the hyperbolic system of 3 equations of ideal granular hydrodynamics in several dimensions for arbitrary adiabatic index $\gamma$. In dependence of initial conditions these solutions can keep smoothness for all times or develop singularity. In particular, in the 2D case the singularity can be formed either in a point or along a line. For $\gamma = -1$ the problem is reduced to the system of two equations, related to a special case of the Chaplygin gas. In the 1D case this system can be written in the Riemann invariant and can be treated in a standard way. The solution to the Riemann problem in this case demonstrate an unusual and complicated behavior.

1. Introduction. The motion of the dilute gas where the characteristic hydrodynamic length scale of the flow is sufficiently large and the viscous and heat conduction terms can be neglected is governed by the systems of equations of ideal granular hydrodynamics [3].

This system is given in $\mathbb{R} \times \mathbb{R}^n$, $n \geq 1$, and has the following form:

\begin{align*}
\partial_t \rho + \text{div}_x (\rho v) &= 0, \\
\partial_t (\rho v) + \text{Div}_x (\rho v \otimes v) &= -\nabla_x p,
\end{align*}

\begin{equation}
\partial_t T + (v, \nabla_x T) + (\gamma - 1)T \text{div}_x v = -\Lambda \rho T^{3/2},
\end{equation}

where $\rho$ is the gas density, $v = (v_1, \ldots, v_n)$ is the velocity, $T$ is the temperature, $p = R\rho T$ is the pressure (the constant $R$ is a adiabatic invariant, for the sake of simplicity we set $R = 1$), and $\gamma$ is the adiabatic index, $\Lambda = \text{const} > 0$. We denote $\text{Div}_x$ and $\text{div}_x$ the divergence of tensor and vector with respect to the space variables. The only difference between equations (1)–(3) and the standard ideal gas dynamic equations (where the elastic colliding of particles is supposed) is the presence of the inelastic energy loss term $-\Lambda \rho T^{3/2}$ in (3).

The granular gases are now popular subject of experimental, numerical and theoretical investigation (e.g. [3], [7], [8] and references therein). The Navier-Stokes granular hydrodynamics is the natural language for a theoretical description of granular macroscopic flows. A characteristic feature of time-dependent solutions of the
continuum equations is a formation of finite-time singularities: the density blowup signals the formation of close-packed clusters.

System (1) – (3) can be written in a hyperbolic symmetric form in variables \( \rho, v, K = \rho \rho^{-1}\gamma \) and the Cauchy problem

\[
(\rho, v, T \rho^{1-\gamma})|_{t=0} = (\rho_0, v_0, T_0 \rho_0^{1-\gamma})
\]
is locally solvable in the class of smooth functions.

System (1) – (3) has no constant solution except the trivial one \((p \equiv 0)\). Another trivial solution is \(v = p = T \equiv 0, \rho(t, x) = \rho_0(x)\). At the same time there exists a solution \(\rho, v, p = \text{const}, T = T(t) = (\Lambda \rho_0 t^2 + T_0^{-1/2})^{-2}, \) (4)

with initial constant value of temperature \(T_0\) (the Haff’s law). This solution is called the homogeneous cooling state.

Here we are going to construct new exact solutions to the ideal granular hydrodynamics with a concentration property and to compare them with the known family of solution obtained earlier in [8].

2. Family of exact solutions in 1D [8]. The authors employ Lagrangian coordinates and derive a broad family of exact non-stationary non-self-similar solutions. These solutions exhibit a singularity, where the density blowups in a finite time when starting from smooth initial conditions. Moreover, the velocity gradient also blowups while the velocity itself and develop a cusp discontinuity (rather then a shock) at the point of singularity.

System (1) - (3) in the Lagrangian coordinates takes the form

\[
\frac{\partial}{\partial t} \left( \frac{1}{\rho} \right) = \frac{\partial v}{\partial m}, \quad \frac{\partial v}{\partial t} = -\frac{\partial p}{\partial m}, \quad \frac{\partial p}{\partial t} = -\gamma p \frac{\partial v}{\partial m} - \Lambda p^{3/2} \rho^{1/2},
\]

and under certain assumptions on the pressure field can be reduced to

\[
\frac{\partial^2 p}{\partial m^2} = -\mu^2 p, \quad \mu = \frac{\Lambda}{\gamma \sqrt{2}}.
\]

Here \(m(x, t) = \int_0^x \rho(\xi, t) d\xi\) is the Lagrangian mass coordinate. The solution is the following:

\[
p = 2A \cos(\mu m), \quad A = \text{const}, \quad \rho(m, t) = \frac{\rho(m, 0)}{(1 - \mu t \sqrt{A \rho(m, 0) \cos \mu m})^2}, \] (5)

The rate of concentration at the maximum point of density as \(t\) tends to a critical time \(t_\ast\) is equivalent to \(\text{Const} \cdot (t_\ast - t)^{-2}\), the behavior of solution at different moments of time and formation of the singularity is presented at Fig.1.

3. Solutions with a constraint. Let us introduce a new dependent variable \(z(t, x)\) as follows: \(z = \rho - \phi(t) T^{-1/2}, \phi(t)\) is an arbitrary differentiable function. Thus, in the variables \(z, T, v\) the system (1)–(3) takes the form

\[
\partial_t z + \text{div}_x(zv) - \frac{\Lambda}{2} \phi(t) z + (\gamma + 1) \phi(t) T^{-1/2} \text{div}_x v + (2\phi'(t) + \Lambda \phi^2(t)) T^{-1/2} = 0, \]

\[
\partial_t v + (v, \nabla_x) v = -\frac{1}{z + \phi(t) T^{-1/2}} \nabla_x(zT + \phi(t) T^{1/2}), \quad (7)
\]

\[
\partial_t T + (v, \nabla_x T) + (\gamma - 1) T \text{div}_x v = -\Lambda z T^{3/2} - \Lambda \phi(t) T. \quad (8)
\]
We consider a particular class of solutions characterized by property \( z = 0 \). There are two possibilities:

- \( \gamma = -1 \) and \( \phi(t) \) is a solution to ODE that can be immediately solved:
  \[
  \phi'(t) = -\frac{\Lambda}{2} \phi^2(t), \quad \phi(t) = \left( \frac{\Lambda}{2} t + \phi(0) \right)^{-1}, \quad \phi(0) = \rho_0^{-1}(x) T_0^{-\frac{z}{2}}(x)
  \]  
  \( (9) \)

- \( v(t, x) = \alpha(t)x + \beta(t) \) and \( \phi(t) \) is a solution to ODE
  \[
  \phi'(t) = -\gamma + 1 - \frac{1}{2} \phi(t) \text{tr} \alpha(t) - \frac{\Lambda}{2} \phi^2(t). \quad (10)
  \]

The first possibility is the case of Chaplygin-like [4] gas, where the state equation is chosen as
  \[
  p = P_0 - \rho^{-1}, \quad P_0 = \text{const} > 0.
  \]  
  \( (11) \)

System (1)–(3),(11) with the constraint \( z = 0 \) can be reduced to a couple of equations
  \[
  \partial_t v + (v, \nabla_x) v = T^\frac{z}{2} \nabla_x (T^\frac{z}{2}),
  \]  
  \( (12) \)

  \[
  \partial_t T^\frac{z}{2} + (v, \nabla_x T^\frac{z}{2}) - T^\frac{z}{2} \text{div}_x v = -\frac{\Lambda}{2} \phi(t) T^\frac{z}{2},
  \]  
  \( (13) \)

where \( \phi(t) \) is given by (9).

In the 1D case this system as any system of two equations can be written in the Riemann invariants, this allows to apply the technique usual for gas dynamics, we will do this in Sec.3.2.

In the second case, for an arbitrary \( \gamma \) the equation (6) can be satisfied only for \( v(t, x) = \alpha(t)x + \beta(t) \), where \( \phi(t) \) solves (10). This case will be considered in Sec.3.1.

### 3.1. Solutions with uniform deformation, arbitrary \( \gamma \)

It is known that for usual gas dynamics equations the solutions with linear profile of velocity \( v(t, x) = \alpha(t)x + \beta(t) \), where \( \alpha(t) \) is a matrix \( n \times n \) and \( \beta(t) \) is an \( n \) - vector, \( x \) is a radius-vector of point, constitute a very important class of solutions [10]. For the system of granular hydrodynamics these solutions give a possibility to construct a singularity arising from initial data.

First of all from (12, 13) we get that in this case \( T \) has to solve the system
  \[
  (\partial_t \alpha(t) + \alpha^2(t))x + (\partial_t \beta(t) + \alpha(t) \beta(t)) = -\frac{1}{2} \nabla_x T,
  \]  
  \( (14) \)

  \[
  \partial_t T + ((\alpha x + \beta), \nabla_x T) + (\gamma - 1) T \text{tr} \alpha(t) = -\Lambda \phi(t) T,
  \]  
  \( (15) \)
and the structure of the field of velocity requires a special structure of the field of temperature, namely,

\[ T(t, x) = x^T A(t) x + (B(t), x) + C(t). \]  

(16)

Thus, we get a system of nonlinear differential equations for components of the square matrix \( \alpha(t) \), the square symmetric matrix \( A(t) \), vectors \( \beta(t) \) and \( B(t) \), the scalar functions \( C(t) \) and \( \phi(t) \), namely

\[
\begin{align*}
\alpha'(t) + \alpha^2(t) + A(t) &= 0, \\
\beta'(t) + 2\alpha(t)\beta(t) + \frac{1}{2}B(t) &= 0, \\
A'(t) + 2A(t)\alpha(t) + (\gamma - 1)\text{tr} \alpha(t)A(t) + \Lambda\phi(t)A(t) &= 0, \\
B'(t) + 2A(t)\beta(t) + B(t)\alpha(t) + (\gamma - 1)\text{tr} \alpha(t)B(t) + \Lambda\phi(t)B(t) &= 0, \\
C'(t) + (B(t), \beta(t)) &= 0.
\end{align*}
\]

(17, 18, 19, 20)

and (10). This system can be explicitly (in the simplest cases) or numerically integrated, one can study its qualitative behavior. The component of density can be found as

\[ \rho(t, x) = \frac{\phi(t)}{(x^T A(t) x + (B(t), x) + C(t))^{1/2}}, \]

(21)

\[ \rho(t, x) \sim \text{const} \cdot |x - x_*|^{-1} \] in the point \( x_* \) of the singularity formation. Therefore the singularity is integrable for \( n > 1 \). Nevertheless, the total mass is infinite for this solution, since \( \int_{\mathbb{R}^n} \rho \, dx \) diverges as \( |x| \to \infty \).

Let us consider the simplest non-rotational case: \( A(t) = a(t)I, \alpha(t) = \alpha_1(t)I, B(t) = 0, \beta(t) = 0 \), where \( I \) is the unit matrix. The system above comes to 4 equations:

\[
\begin{align*}
\phi'(t) + \frac{n}{2}(\gamma + 1)\phi(t)\alpha_1(t) - \frac{\Lambda}{2}\phi^2(t), \\
\alpha_1'(t) + \alpha_1^2(t) + a(t) &= 0, \\
a'(t) + (2 + n(\gamma - 1))\alpha_1(t) + \Lambda\phi(t)\alpha(t) &= 0, \\
C'(t) + (n(\gamma - 1)\alpha_1(t) + \Lambda\phi(t))C(t) &= 0.
\end{align*}
\]

(22, 23, 24)

We are going to find asymptotics of the solution at the point \( t = t_* > 0 \) of the singularity appearance.

Systems (22)-(24) is a polynomial system

\[ \dot{x} = f(x), \quad f: \mathbb{R}^k \to \mathbb{R}^k, \]

(25)

and we can study the occurrence of blow-up analyzing the solutions locally around their movable singularities using a set of methods based on the construction of local series. Following [5], [6], we build local series (\( \Psi \)-series) of the form:

\[ x = \Psi(\lambda, s, t) = \lambda\tau^s(1 + h(\tau, \ln \tau)), \]

where \( \tau = t_* - t \) and \( h(\tau; \ln \tau) \) is a power series in its argument which vanishes as \( \tau \to 0 \). The notation \( \lambda \tau^s \) refers to the vector whose \( i \)-th component is \( \lambda_i \tau^{s_i} \). In order to obtain the leading behavior \( \lambda \tau^s \) of the solution around \( t_* \) we look for all negatively quasihomogeneous truncations \( \tilde{f} \) of the vector field \( f = \tilde{f} + \hat{f} \) such that the dominant behavior \( x = \lambda \tau^s \), \( \lambda \in \mathbb{C}^k \) is an exact solution of the truncated system

\[ \dot{x} = \tilde{f}(x) \]

and

\[ \tilde{f}(\lambda \tau^s) \sim \lambda_i \tau^{s_i + \delta_i - 1}, \quad \delta_i \in \mathbb{Q}^k, \quad \delta_i > 0, \]

as \( \tau \to 0 \). Each truncation defines a dominant balance \( (\lambda, s) \) and every balance corresponds to the first term \( \lambda \tau^s \) in an expansion around movable singularities. For such an expansion to describe a general solution, the \( \Psi \)-series must contain \( k - 1 \)
arbitrary constants in addition to the arbitrary parameter $t_\ast$. The position in the power series where these arbitrary constants appear is given by the resonances. They are given by the eigenvalues of the matrix $R$:

$$R = -D\hat{f}(\lambda) - \text{diag}(s),$$

where $D\hat{f}(\lambda)$ is the Jacobian matrix evaluated on $\lambda$. The resonances are labeled $r_i$, $i = 1, \ldots, k$ with $r_1 = -1$. Each balance defines a new set of resonances.

**Theorem 3.1.** ([6], T.3.8) Consider a real analytic system (25) and assume that it has a balance $(\lambda, s)$ such that $r_j > 0$ for all $j > k - m + 1$, $1 \leq m \leq k$ and $\lambda \in \mathbb{R}^n$. Then there exists an $m$-dimensional manifold $S^m_\lambda \subseteq \mathbb{R}^k$ of initial conditions leading to a finite time blow-up, for all $x_0 \in S^m_\lambda$, i.e. there exists $t_\ast \in \mathbb{R}_+$ for which $|x(t; x_0)| \to \infty$ as $t \to t_\ast$.

**Corollary 3.1.** If $n \geq 2$ and $\gamma > -1 + \frac{2}{n}$, then there exists an open set $\Omega$ of initial data $x_0 = (\phi(0), \alpha_1(0), \alpha_0(0), C(0))$ such that the components $\phi(t)$ and $\alpha_1(t)$ of solution to the system (22)–(24) blow up within a finite time for all $x_0 \in \Omega$.

**Proof.** To find main terms of asymptotic at the point of singularity we consider a negatively quasihomogeneous truncation of the system (22)–(24), namely

$$\phi'(t) = -\frac{n}{2}(\gamma + 1) \phi(t) \alpha_1(t) - \frac{\Lambda}{2} \phi^2(t),$$

$$a_1'(t) = -a_1^2(t),$$

$$a'(t) = -(2 + n(\gamma - 1))a_1(t)a(t) - \Lambda \phi(t)a(t),$$

$$C'(t) = -n(\gamma - 1)\alpha_1(t)C(t) - \Lambda \phi(t)C(t).$$

The solution to the above system is the following:

$$\phi(t) = -\frac{n(\gamma + 1) - 2}{\Lambda} (t_\ast - t)^{-1}, \quad \alpha_1(t) = -(t_\ast - t)^{-1},$$

$$A(t) = A_0(t_\ast - t)^{2(n-2)}, \quad C(t) = C_0(t_\ast - t)^{2(n-1)}, \quad A_0 = \text{const}, \quad C_0 = \text{const},$$

$$s = \text{diag}(-1, -1, 2(n - 2), 2(n - 1)), \quad \lambda = (-1, -\frac{n(\gamma + 1 - 2)}{\Lambda} A_0, C_0).$$

The resonances, computed for this balance are $(n(\gamma + 1) - 2, -1, 0, 0)$. As follows from Theorem 3.1, there exists a manyfold $S^m_\lambda \subseteq \mathbb{R}^n$ such that for $\phi(0), \alpha_1(0)$ the respective solution to (22) blows up and (30) is the main term of asymptotics at the blow up point. Other components of solution to (22)–(24) can be found from linear with respect to $a(t)$ and $C(t)$ equations (28) and (29) for any initial data (it makes sense to consider $a(0) > 0, C(0) > 0$).

**Remark 3.1.** The rate of growth of the maximum of the density as $t \to t_\ast$ is $\rho(t, 0) \sim \text{const} \cdot (t - t_\ast)^{-n}$.

**Remark 3.2.** If $t_\ast < 0$, then an analogous consideration shows that there exists an open set of initial data such that the solution to system (22)–(24) remains bounded for all $t > 0$.

**Theorem 3.2.** If $n = 1$, then for any $t_\ast > 0$ there exists a family of exact solutions to system (22)–(24) depending on parameters $(\alpha_0, C_0)$, blowing up as $t \to t_\ast$. This family is physically reasonable for $\alpha_0 \in (-1, -\frac{2}{\gamma + 1})$, $C_0 > 0, \gamma > 1$. For these solutions the maximum of density has the asymptotics $\rho(t, 0) \sim \text{const} \cdot (t - t_\ast)^{\alpha_0}$ as $t \to t_\ast$. 

EXACT SOLUTIONS TO IDEAL HYDRODYNAMICS OF INELASTIC GASES 903
Proof. It can be readily checked that the balance \( s = (-1, -1, -2, -2(\alpha_0 + 1)) \), \( \lambda = (-2 + (\gamma + 1)\alpha_0)/\Lambda, \alpha_0, -\alpha_0(\alpha_0 + 1), C_0 \) gives an exact solution. The restriction on the parameters \( \alpha_0 \) and \( C_0 \) follows from the positivity of the expression under the square root in (21) and the positivity of \( \phi(0) \). □

Remark 3.3. The maximum of density as \( t \to t_\ast \) grows slower than for the solution obtained in [8].

Fig.2a presents the results of numerical computations in 2D based on system (17) –(29). The initial density has the form (21). Fig.2b shows the density near the blow-up time for \( \alpha_{11} = \alpha_{22} < 0, \alpha_{12} = \alpha_{21} = 0 \), with a concentration in a point. Fig.2c shows the density near the blow-up time for \( \alpha_{11} < 0, \alpha_{22} = \alpha_{12} = \alpha_{21} = 0 \) with a concentration along a line. The computations demonstrate a complicated behavior of solution. In particular, a vorticity can prevent the singularity formation.

![Figure 2. a,b,c](image)

3.2. Chaplygin gas, \( n = 1 \). The model of gas dynamics with the pressure given by (11) is known as the Chaplygin gas. The Chaplygin gas is now considered as a possible model for dark matter-energy [1]. This system can also be seen as the one-dimensional version of the Born-Infeld system, a non linear modification of the Maxwell equations, designed by Born and Infeld in 1934 to solve the electrostatic divergence generated by point particles in classical electrodynamics. The Chaplygin system is known to be hyperbolic, linearly degenerate, weakly stable [12]. Recently this system attracted a lot of attention, e.g [2], [11].

The system (12), (13) can be reduced to

\[
\partial_t \rho + \text{div}_x (\rho v) = 0, \quad \partial_t (\rho v) + \text{Dive}_x (\rho v \otimes v - \frac{\phi^2(t)}{\rho}) = 0,
\]

(31)

recall that \( \rho = \phi(t) T^{-\frac{1}{2}} \). It is similar to the Chaplygin gas system, the only difference is in the known multiplier \( \phi(t) \), for the Chaplygin gas \( \phi = \text{const} \). For \( n = 1 \) (31) can be written in the Riemann invariants as

\[
\partial_t s + \rho \partial_s s = \frac{\Lambda \phi(t)}{4} (r - s), \quad \partial_t r + s \partial_s r = \frac{\Lambda \phi(t)}{4} (s - r),
\]

(32)

where \( s = v - T^\frac{3}{4}, r = v + T^\frac{3}{4}, \phi(t) \) is given by (9). This system is linear degenerate, therefore provided the solution is bounded there is no possibility for the gradient catastrophe.
**Theorem 3.3.** The solution to the Riemann problem for (31) ((32)) in 1D with data

\[(v, T) = \begin{cases} 
(v_L, T_L), & x < 0, \\
(v_R, T_R), & x > 0,
\end{cases}\]  
for \( v_L > v_R \)

contains a \( \delta \)-singularity in the component of density. If

\[v_L \geq v_R + T_L^2 + T_R^2,\]

then the \( \delta \)-singularity formation begins from the initial moment of time.

**Proof.** Since the system is linear degenerate, the jumps are contact discontinuities and move along characteristics. The solution is based on the cooling state (4). If

\[v_L < v_R + T_L^2 + T_R^2,\]

then the solution is

\[(v, T) = \begin{cases} 
(v_L, T_L(t)), & x < x_{-}(t), \\
(v_M(t), T_M(t)), & x_{-}(t) < x < x_{+}(t), \\
(v_R, T_R(t)), & x > x_{+}(t),
\end{cases}\]  

with \( c = \phi^{-1}(0) \),

\[T_L(t) = \frac{cT_L^2}{(c + c)^2}, \quad T_R(t) = \frac{cT_R^2}{(c + c)^2},\]

\[x_{-}(t) = vLt - \frac{2cT_L^2}{\Lambda} \ln(\frac{\Lambda}{2c} t + 1) \quad x_{+}(t) = vRt + \frac{2cT_R^2}{\Lambda} \ln(\frac{\Lambda}{2c} t + 1),\]

\[v_M(t) = \frac{v_L + v_R + c(T_R^2 - T_L^2)(\frac{\Lambda}{2} t + c)^{-1}}{2},\]

\[T_M^2(t) = \frac{v_R - v_L + c(T_R^2 + T_L^2)(\frac{\Lambda}{2} t + c)^{-1}}{2}.\]

If \( v_L \leq v_R \), then \( x_{-}(t) < x_{+}(t) \) for all \( t > 0 \) and the solution to the Riemann problem is given by (37). If \( v_L < v_R \), then there exists a moment \( t_\ast > 0 \) such that \( x_{-}(t_\ast) = x_{+}(t_\ast) \). Moreover, in the moment \( t_{**} > 0 \) the component \( T_M \) vanishes and \( t_{**} < t_\ast \). Thus, we have to construct a new solution starting from \( t_{**} \).

We are going to introduce a \( \delta \)-singularity in the density concentrated on the jump analogously to [2]. To find a \( \delta \)-type singularity solution we have to use the system in its conservative form (31). Let us denote \( x_{\ast}(t) \) the position of the singularity and look for a solution in the form:

\[\rho(t, x) = \rho_\ast + [\rho]H(x - x_{\ast}(t)) + \theta(t)\delta(x - x_{\ast}(t)),\]

\[\rho(t, x)v(t, x) = \rho v_\ast + [\rho v]H(x - x_{\ast}(t)) + \psi(t)\delta(x - x_{\ast}(t)),\]

\[\rho(t, x)v^2(t, x) = \rho v^2_\ast + [\rho v^2]H(x - x_{\ast}(t)) + \Psi(t)\delta(x - x_{\ast}(t)),\]

\[\tau(t, x) = \tau_\ast + \tau H(x - x_{\ast}(t)), \quad \tau = \rho^{-1},\]

\([f] = f_\ast - f_\ast, \quad f_\ast \) and \( f_\ast \) are the limits of an arbitrary function from the right and from the left side of \( x_{\ast}(t) \), respectively, \( H \) is the Heaviside function. From (31) we get

\[x_{\ast}(t) = \frac{[\rho v]t_1 + \theta(t)}{[\rho]} + x_0,\]

\[\theta(t) = \sqrt{([\rho v]^2 - [\rho][\rho v^2])t_1^2 + \frac{4}{\Lambda} [\rho][\tau] \left( \phi(t_0)t_1 - \frac{2}{\Lambda} \ln \left( \frac{\Lambda\phi(t_0)}{2}t_1 + 1 \right) \right)},\]
\[ t_1 = t - t_0, \ t_0 \text{ and } x_0 \text{ are the moment and the coordinate of the } \delta \text{-singularity formation. Expanding the expression under the square root at } t = t_0 \text{ we find the necessary condition for the beginning of the concentration processes:} \]
\[ [v]^2 \geq [T^{1/2}(t_0)]^2. \]  (44)

If the initial data are such that the inequality opposite to (36) holds, i.e. \( v_L \geq v_R + T^{1/2}_L(0) + T^{1/2}_R(0) \), then \( v_L > v_R \) and
\[
(v_L - v_R)^2 \geq (T^{1/2}_L(0) + T^{1/2}_R(0))^2 \geq (T^{1/2}_R(0) - T^{1/2}_L(0))^2,
\]
such that the condition (44) is satisfied. Since in this case the solution consisting of two contact discontinuities is impossible, the only reasonable solution is given by (38)– (41), (42).

If \( u_L > u_R \), we can define again the solution of form (38)– (41), (42). In principle, initially the condition (44) may fail and this solution can exist beginning from some \( \hat{t} \), \( 0 < \hat{t} \leq t_* \), nevertheless, this moment \( \hat{t} \) always exists. Further, at least for \( t > t_* \) this solution is stable. Let us extend this solution back to \( t_* < t < t_* \). This means the assumption that the segment \([x_-(\hat{t}), x_+(\hat{t})]\) shrinks into the point \( x_0 = x_*(t_*) \) at the moment \( t_* \) and at this moment the formation of singularity begins. The point \( x_0 \) can be found from the condition \( x_*(t_*) = x_-(t_*) = x_+(t_*) \). It can be shown that the velocity of the singular front \( \dot{x}_*(t) \to \frac{v_L + v_R}{2} \) as \( t \to \infty \). □

Remark 3.4. The question on uniqueness of the solution after the “shrinking” is open.

Remark 3.5. In [9] for any spatial dimensions a simple family of solutions to the system (1) – (3) having a singularity in the density whereas other components are continuous is constructed. Moreover, a family of self-similar solutions in 1D was found.

REFERENCES


E-mail address: rozanova@mech.math.msu.su
MODELLING, SIMULATION AND OPTIMIZATION OF GAS DYNAMICS IN AN EXHAUST PIPE

INGENUIN GASSER, MARTIN RYBICKI AND WINNFRIED WOLLNER
Bundesstraße 55
20146 Hamburg, Germany

Abstract. We study the gas flow in an exhaust pipe. In particular we focus on the requirement of heating up fast the catalytic converter.

To describe the gas dynamics in the exhaust pipe we start with a set of one dimensional fluid dynamic equations for variable cross section. We derive by using a small Mach number limit an asymptotic model consisting of coupled ODEs and PDEs. For the description of the full exhaust pipe we use a network ansatz connecting various pipes with (different) constant cross section. This model still describes the main features and is computationally a few orders of magnitude faster than the original model.

Furthermore we will present a related optimization problem.

1. Introduction. The emission caused by vehicles is an import issue in these times. In order to reduce the pollutant emissions of CO, NO\textsubscript{x} and C\textsubscript{x}H\textsubscript{y} in the exhaust gas, there are catalytic converters installed in the exhaust pipe system. A crucial quantity to control the efficiency of a catalytic converter is the temperature. Due to this reason, one is interested in how to ensure a sufficient high temperature in the catalytic converters in a short time after the engine start.

A special method of heating after the engine start is the combustion of unburnt gas in the catalytic converters. Modern cars can control the ratio of oxygen and fuel in the combustion chamber of the engine. By choosing a ratio with more fuel and less oxygen some unburnt gas gets to the catalytic converters and can be used there for an exothermic reaction.

We start the modelling process with the following fully compressible, hyperbolic balance law

\[
(A\rho)_t + (A\rho u)_x = 0,
\]

\[
(A\rho u)_t + (A\rho u^2 + Ap)_x = -\frac{\xi}{4}\pi d\frac{|u|}{2} - C_c A\chi \rho u,
\]

\[
(A\rho E)_t + (A\rho uE + Aup)_x = -h\pi d(T - T_{Wall}) + q_0 A\chi \rho z K(T),
\]

\[
(A\rho z)_t + (A\rho uz)_x = -A\chi \rho z K(T),
\]

where we use the ideal gas law \( p = R\rho T \) as the closing relation. The terms on the left hand side are the Euler equations of gas dynamic in a pipe with a variable cross section \( A \). By the terms on the right hand side we describe the main physical effects which influence the dynamic of the gas. In the second equation we first have a wall friction (\( \xi \)) and secondly a local friction (\( C_c \)) due to the honeycomb structure.

2000 Mathematics Subject Classification. 35Q35, 35Q93, 80A25.

Key words and phrases. Exhaust pipe, Gas dynamics, Gas dynamics network, Small Mach number, Combustion, Optimization.
of the catalyst (locality is denoted by $\chi$). In the energy equation we have, on one hand, energy loss due to heat transfer with the wall ($h$), and, on the other hand, an energy gain due to the local exothermic reaction in the catalytic converter ($q_0$). Lastly, the reaction equation denotes the ratio of unburnt gas $z$ in the pipe. The temperature dependent reaction rate $K(T)$ is modelled by an Arrhenius’ law.

Starting from this model we want to perform two major simplification:

1. Rather than introducing artificial intervals for a smooth change of the cross section function $A$, we divide the exhaust pipe into pipes with constant cross sections and treat the whole system as a network. This has two major advantages:
   (a) Since the cross section $A$ does not have to change quickly on small intervals, which would imply large derivatives, we are able to produce numerical results with much larger step sizes in space.
   (b) Due to the necessity of coupling conditions for the network, we are able to describe the pressure losses caused by the pipe’s geometry.

2. The flow in the exhaust pipe is in a low Mach number regime, and we are not interested in the propagation of sound waves, but mainly the temperature in the catalyst. Thus, we are able to simplify the model by a small Mach number asymptotics. This will lead us to a simplified model (AM), which is computationally much faster than the hyperbolic model, since much larger step sizes in time are allowed by the CFL condition.

2. Two models: (FE) and (AM). We want to verify the accuracy of the asymptotic model (AM) by numerical comparison with the fully compressible model (FE) which is established in the engineering literature. Therefore, we present both models quickly.

First we start of the description of the gas dynamics in a single pipe with constant cross section. Then we will explain how to put those single pipes together in order to simulate the gas dynamics in the whole exhaust pipe.

2.1. Full Euler (FE) - single pipe. By treating the cross section $A$ and the diameter $d$ as constants, we obtain.

\[ \rho_t + (\rho u)_x = 0, \]
\[ (\rho u)_t + (\rho u^2 + p)_x = -\frac{\xi}{d^2} \frac{u|u|}{2} - C_c \rho u, \]
\[ (\rho E)_t + (\rho u E + up)_x = -\frac{4h}{d} (T - T_{Wall}) + q_0 \rho z K(T), \]
\[ (\rho z)_t + (\rho uz)_x = -\rho z K(T), \]

Since now each pipe with constant cross section has its own system, we need to distinguish between pipes that do have (includes blue terms), and pipes that do not have a catalyst (blue terms not included).

2.2. Asymptotic model (AM) - single pipe. We have to work more for the asymptotic model. First we need to scale the system and expand asymptotically each unknown, e.g., for the pressure $p = p_0 + \varepsilon p_1 + O(\varepsilon^2)$, where $\varepsilon = \gamma M^2$ ($\gamma$ is the adiabatic exponent and $M$ the Mach number). We then compare the coefficients that have the same order of magnitude. For the order $\varepsilon^{-1}$ we obtain

\[ (p_0)_x = 0, \]

(1)
i.e., the leading order in pressure is constant. We do the same comparison for the
order $\varepsilon^0$. In order to adopt the equations to the boundary data given for such
an application (pressure values at both ends of the pipe) we perform additionally
some simple manipulations (see [1] for details). The final version of our small Mach
number asymptotic model is the following:

$$
\begin{align*}
\rho_t + (v + Q)\rho_x &= -q\rho, \\
z_t + (v + Q)z_x &= -zK(T), \\
v_t &= \frac{1}{\int_0^1 \rho dx} \left[ p_{1r} - p_{1l} - \int_0^1 \rho Q_t \, dx - \int_0^1 \rho(v + Q)q \, dx \right. \\
&\quad \left. - C_f \int_0^1 \frac{\rho(v + Q)|v + Q|}{2} \, dx - C_c \int_0^1 \rho(v + Q) \, dx \right].
\end{align*}
$$

Remark 1. The only unknown that appears in the 1st order of its expansion is the
pressure. All other unknowns appear only in 0th order in this limit. In this case we
resign to write the sub-index.

The velocity $u$ is now split into the space independent velocity component $v$ and
the aggregated energy balance $Q$.

$$
\begin{align*}
u(x, t) &= v(t) + Q[\rho, z](x, t) = v(t) + \int_0^x q[\rho, z](y, t) dy, \\
q[\rho, z] &= \frac{1}{\gamma \rho_0} \left[-h(T - T_{Wall}) + q_0 \rho zK(T)\right].
\end{align*}
$$

We see that instead of four coupled PDEs in (FE) we have now only a system of two
(transport) PDEs and one ODE (Although this is formally an integro-differential
equation, from the perspective of our numerical treatment, we can call is an ODE).
Furthermore the boundary conditions for the pressure appear as parameters in the
system (AM).

2.3. Network - The whole exhaust pipe. In order to put things together, i.e.,
to establish connections among neighbouring pipes, we need to define coupling con-
ditions. Those coupling conditions are for all times $t$:

$$
\begin{align*}
\text{conservation of mass} : & \quad \rho^i_j(t) = \rho^{i+1}_j(t) \\
\text{conservation of ratio of unburnt gas} : & \quad z^i_j(t) = z^{i+1}_j(t) \\
\text{conservation of internal energy} : & \quad A^i u^i_j(t) = A^{i+1} u^{i+1}_j(t) \\
\text{pressure loss at junctions} : & \quad p^i_j(t) = p^{i+1}_j(t) + \varepsilon f^j_{ext}(t)
\end{align*}
$$

The super-index denotes the different pipe, whereas the sub-index denotes the spatial
evaluation ($y^{i+1}_j(t) = y^{i+1}(0, t)$ and $y^i_j(t) = y^i(L^i, t)$, where $L^i$ is the length of
the $i$-th pipe). The term $f^j_{ext}$ denotes the pressure loss at the $j$-th junction and
depends upon its geometry and the kinetic energy (see Figure 1). The formulas to
describe those pressure losses can be found in the engineering literature (see e.g.,
[3] or [4]) and are heuristically motivated.

3. Simulation. We want to compare the full Euler model (FE) and the asymptotic
model (AM) numerically. Since we want to have a ”fair competition” we have to
solve both systems with a similar sort of numerical algorithm.
pressure loss at a sudden expansion:

\[ p_j^r = p_{j+1}^r + \varepsilon f_{\text{ext}}^j \]
with
\[ f_{\text{ext}}^j = \rho_j^r \left( \frac{u_j^r}{d_j^r} \right)^2 \left( 1 - \frac{d_{j+1}^2}{d_j^2} \right)^2 \]

pressure loss at a sudden contraction:

\[ p_j^r = p_{j+1}^r + \varepsilon f_{\text{ext}}^j \]
with
\[ f_{\text{ext}}^j = \frac{1}{2} \rho_j^{r+1} \left( \frac{u_j^{r+1}}{d_j} \right)^2 \left( 1 - \frac{d_{j+1}^2}{d_j^2} \right) \]

Figure 1. Coupling conditions for the pressure

3.1. **Numerical treatment.** We solve both models by explicit upwinding. Furthermore, we have to compute the necessary data, i.e., the pressure values, at the junctions. For the (FE) model we solve linearised Riemann problems at each junction. For the (AM) model a more sophisticated ansatz is needed (for details see [1]). It leads to a linear system which delivers, in contrast to the hyperbolic case, the pressure values at all junctions at once.

So, in terms of mathematics we are dealing in with a local problem in the full Euler case, whereas in the asymptotic case we have a global problem, with instantaneous influences from all over the network. This is due to the fact that we have a finite speed of propagation in the full Euler model and infinite speed of propagation in the asymptotic model. The infinite speed of propagation in the asymptotic model is a consequence of the small Mach number limit.

3.2. **Numerical costs.** Comparing the two CFL-conditions for the upwind schemes for the full Euler equations (3) and for the asymptotic model (4), one can already guess at first sight, that the time step sizes in the full Euler case will be much smaller.

\[ \lambda_{\text{max}} \frac{\Delta t}{\Delta x} \leq c_N \quad \text{with} \quad \lambda_{\text{max}} := \max |\lambda| \quad (3) \]
\[ u_{\text{max}} \frac{\Delta t}{\Delta x} \leq c_N \quad \text{with} \quad u_{\text{max}} := \max |u| \quad (4) \]

The parameter \( c_N \) denotes the Courant number. Assuming we have positive velocities, the maximum Eigenvalue \( \lambda_{\text{max}} \) for the full Euler equations is \( \lambda_{\text{max}} = u_{\text{max}} + c \), where \( c \) denotes the speed of sound. Starting from the same spatial grid, this would lead to the following relation between the time step size for the asymptotic model (\( \Delta t_{\text{AM}} \)) and the time step size for the full Euler equations (\( \Delta t_{\text{FE}} \)):

\[ \Delta t_{\text{AM}} = \frac{\lambda_{\text{max}}}{u_{\text{max}}} \Delta t_{\text{FE}} = \left( 1 + \frac{1}{M} \right) \Delta t_{\text{FE}}, \]

where \( M = \frac{u_{\text{max}}}{c} \) denotes the Mach number. One can see that the smaller the Mach number is, the more time efficient the algorithm for the asymptotic model is, due to larger step sizes in time.
Additionally one can count the number of floating point operation (flops), which are necessary for the computation of the next time step at one spatial grid point. Although this is a very rough estimate, one can deduce, that the number of flops for this computation are about 5 times higher for the hyperbolic model. This gives us a lower estimate for the numerical effort of the full Euler model:

\[ \text{work}_{FE} \geq 5 \left( 1 + \frac{1}{M} \right) \cdot \text{work}_{AM}. \]

3.3. Numerical example. We simulate an engine start, i.e., the initial conditions are \( \rho(x,0) = 1.2 \text{ kg/m}^3 \), \( z(x,0) = 0 \) and \( u(x,0) = 0 \text{ m/s} \). The constant boundary conditions for the density and the ratio of unburnt gas at the left end of the pipe are \( \rho(0,t) = 0.4 \text{ kg/m}^3 \) and \( z(0,t) = 0.1 \). As pressure difference we have \( \Delta \rho = 0.001 \text{ bar} \). The result of the numerical simulation is presented in Figure 2, where the green lines represent the numerical solution of (FE) and the blue lines the numerical solution of (AM). The geometry of the exhaust pipe is shown in the figures. The pipes filled with grey color are those with catalytic converters.

We can observe a good qualitative agreement of both solutions. Although we did not compare our results to real experiments, due to lack of data, we could compare the results with papers that did so and could observe a good agreement (see [6, 7, 8]).

The measured computational effort for (FE) is of the order \( 10^4 \) larger than the effort for (AM), where simulations in real time are possible for (AM).

4. Optimization. We want to do some optimization with the asymptotic model (AM). First we need to declare quantities we want to optimize. In case
of the exhaust pipe application this would be the temperature of the catalytic converter \((T_c)\). It is desired to reach an optimal temperature level \(T_{\text{opt}}\) at which the transformation of harmful gases into less harmful gases works best. Therefore, we need an additional equation to determine the catalyst’s temperature in time. We will use the following simple ODE:

\[
\frac{d}{dt} T_c(t) = -h_c(T_c(t) - T_{\text{Gas}}(t)), \tag{5}
\]

where \(T_{\text{Gas}}(t)\) denotes the average gas temperature in the catalytic converter at time \(t \in [0, t_{\text{end}}]\). This has an effect on the energy balance (2). Therefore we need to add a term in \(q\), such that the energy loss (or gain) due to heat transfer with the catalytic converter would be modelled. This term would be basically the right hand side of the ODE (5) with opposite sign.

4.1. The cost and Lagrange functional. We use the boundary condition for the ratio of unburnt gas \(z_l(t)\) as the control variable of the temperature in the catalytic converter. This is done because the unburnt gas is responsible for the exothermic reaction which heats up the catalyst. We want to punish the temperature difference between current and optimal temperature in the catalyst, as well as the use of unburnt gas in our cost functional \((J)\).

\[
J(T_c, z_l) = \frac{1}{2} \int_0^{t_{\text{end}}} (T_c(t) - T_{\text{opt}})^2 dt + \sigma \cdot \frac{1}{2} \int_0^{t_{\text{end}}} (z_l(t))^2 dt
\]

The parameter \(\sigma\) can be chosen by the user and determines the penalization of the control variable \(z_l\), i.e., in our application the cost of fuel.

The task is now: Find \(z_l, \rho, z, v, T_c\) such that \(J(T_c, z_l)\) is minimized, subject to the system \((AM+5)\) and its initial, boundary, and coupling conditions.

In order to fulfill the task, we are going to use a gradient based method. For the calculation of the gradient we employ adjoint calculus. Therefore, we need to calculate the Lagrangian functional first. The Lagrangian functional \(L\) consists of the cost functional minus the space-time integral of the Lagrangian multipliers or adjoint variables \(\xi_i\) times the constraints, which are the three equations of \((AM)\), \((5)\) and the initial, boundary, and coupling conditions. Due to the shortage of space we resign from displaying the Lagrangian functional here.

The next step is to compute the Fréchet-derivatives of \(L\), with respect to the adjoint variables \(\xi_1, \xi_2, \xi_3, \xi_4\) (leads to the constraints, i.e., the state system \((AM+5)\)), the prime variables \(\rho, z, v, T_c\) (leads to the adjoint system) and the control \(z_l\) (leads to the optimality condition). The constraints, adjoint system and the optimality condition form a optimality system or KKT-system, which can be used for our optimization algorithm.

A detailed computation of the above describe procedure will be part of forthcoming work \([2]\). However, the formal computation is standard and can be found in the literature (e.g., \([5]\)).

Remark 2. Since we do not know anything about the existence or uniqueness of \((AM+5)\), we just assume all variables to be smooth, i.e., depend differentiable on the control, in order to execute the above mentioned calculations.

4.2. Numerical algorithm (Steepest descend). First we guess an initial control \(z_l^{(0)}\) and repeat the following steps for \(i = 0, 1, 2, \ldots\) until satisfactory convergence is achieved,
4.3. Examples. We want to present two examples. One in which fuel is cheap ($\sigma = 0.1$) and we start with a low usage of it ($z(0)_l = 0.01$), and one with in which fuel is expensive ($\sigma = 1$) and the initial guess for the control is high ($z(0)_l = 0.1$). The pipe’s geometry is shown in the adjoining figure.

The geometry for the following examples is shown in the adjoining graphic. We choose the optimal temperature to be $T_{opt} = 600$ K and want to observe the first 60 s after the engine start, i.e. $t_{end} = 60$ s.

The results are shown in Table 1. One can observe that the expected behaviour, i.e., a reduction of the control variable in the case that fuel is expensive and vice versa. We remark that the convergence occurred after only 3 iterations of the gradient decent algorithm. Thus, we have refrained from using more advanced optimization algorithms.

Table 1. Numerical results of optimization example

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$J$</th>
<th>$J_{z_l}$</th>
<th>$J_{T_c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>118.873</td>
<td>3.000</td>
<td>118.574</td>
</tr>
<tr>
<td>1</td>
<td>69.106</td>
<td>31.863</td>
<td>40.183</td>
</tr>
<tr>
<td>2</td>
<td>47.542</td>
<td>73.588</td>
<td>40.183</td>
</tr>
<tr>
<td>9</td>
<td>35.306</td>
<td>171.433</td>
<td>18.162</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$J$</th>
<th>$J_{z_l}$</th>
<th>$J_{T_c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>328.412</td>
<td>300.000</td>
<td>28.412</td>
</tr>
<tr>
<td>1</td>
<td>111.408</td>
<td>65.232</td>
<td>46.177</td>
</tr>
<tr>
<td>2</td>
<td>98.437</td>
<td>29.959</td>
<td>68.478</td>
</tr>
<tr>
<td>3</td>
<td>98.303</td>
<td>27.405</td>
<td>70.898</td>
</tr>
</tbody>
</table>

1. solve with $z_l^{(i)}$ the constraints to obtain $\rho^{(i)}, z^{(i)}, v^{(i)}$ and $T_c^{(i)}$;
2. solve with $\rho^{(i)}, z^{(i)}, v^{(i)}$ and $T_c^{(i)}$ the adjoint system to obtain $\xi_1^{(i)}, \xi_2^{(i)}, \xi_4^{(i)}$;
3. use results of step 1 and 2 to compute the gradient of the cost functional;
4. use a line search algorithm (e.g. Armijo) to compute a new control $z_l^{(i+1)}$.

5. Conclusion. To reduce the complexity of the gas dynamic models for exhaust pipes we derived by a small Mach number asymptotics an alternative model and used a network ansatz to describe the pipe’s geometry. By comparison to established models for exhaust pipe flow we could verify that this new simple model still keeps the main physical features of the application. The network ansatz gives the opportunity to include pressure loss terms (for both models: (FE) and (AM)). The new asymptotic model is computationally much (orders of magnitude) faster than the standard (fully compressible) model. The numerical simulation of the asymptotic model is so inexpensive, that real time simulations are possible even on so-called Smartphone (Implementation was done for a Samsung Galaxy S2 i9100, see [9]).
Furthermore, we can use the asymptotic model for optimization tasks. The derivation of an optimality system through an adjoint based methods seems to work satisfactory. This opens the possibility to answer control and optimize question (also for similar problem like gas pipe lines or any sort of fluid networks) numerically in a short time.

REFERENCES


E-mail address: gasser@math.uni-hamburg.de
E-mail address: martin.rybicki@math.uni-hamburg.de
E-mail address: winnifried.wollner@math.uni-hamburg.de
A COMBINED HYBRIDIZED DISCONTINUOUS GALERKIN / HYBRID MIXED METHOD FOR VISCOUS CONSERVATION LAWS

Jochen Schütz
IGPM, RWTH Aachen University
Templergraben 55
52062 Aachen, Germany

Michael Woopen and Georg May
AICES, RWTH Aachen University
Schinkelstraße 2
52062 Aachen, Germany

Abstract. Recently, we have proposed a method for solving steady-state convection-diffusion equations, including the full compressible Navier-Stokes equations [19]. The method is a combination of a mixed Finite Element method for the diffusion terms, and a Discontinuous Galerkin method for the convection term. The method is fully implicit, and the globally coupled unknowns are the hybrid variables, i.e., variables having support on the skeleton of the mesh only. This reduces the amount of overall degrees of freedom tremendously. In this paper, we extend our method to be able to cope with time-dependent convection-diffusion equations, where we use a dual time-stepping method in combination with backward difference schemes.

1. Introduction. Based on recent work on high-order methods, especially Discontinuous Galerkin methods [1, 6, 10], we have proposed a method for solving steady-state convection-diffusion equations in [19], based on the work by Egger and Schöberl [7]. The method combines a mixed Finite Element method for the diffusion terms, and a Discontinuous Galerkin method for the convection terms. It turned out to be actually very similar to a hybridized DG method proposed by Nguyen et al. [16, 17, 18], see also [20] for a comparison of both methods with respect to asymptotic performance.

For stationary and (weakly-)instationary applications in aerodynamics, implicit methods are very popular, as they allow for large time-steps and thus an efficient solution process [12, 15]. Arguably, most popular implicit methods rely on the use of a Newton-type solver, i.e., one needs the Jacobian of the discretization. Typically, for Discontinuous Galerkin methods, the size of this matrix is $O(N \cdot p^{2d})$, where $N$ is the number of elements in the triangulation, $p$ is the underlying degree of polynomial and $d$ is the spatial dimension. Available memory then usually poses a severe restriction on both $N$ and $p$.

A known way to reduce the size of the Jacobian that has gained some attention recently is the use of hybridization [3, 5]. Roughly speaking, instead of considering

2000 Mathematics Subject Classification. 35M33, 65M20, 65M60.
ansatz functions having support in the interior of the elements, one considers ansatz functions whose support is the skeleton of the mesh. For a hybridized Discontinuous Galerkin method, the Jacobian is of size $O(\tilde{N} \cdot p^{2(d-1)})$, where $\tilde{N}$ is the number of edges. Typically, this yields a reduction of the Jacobian size, which implies less storage requirements and, usually, a faster iterative solution process.

In this paper, we extend the method proposed in [19] to cope with time-dependent problems. Time discretization relies on $A(\alpha)$-stable backward difference schemes. We give the definition of the method and show numerical results.

2. Underlying Equations. Let $\Omega \subset \mathbb{R}^2$ be a domain. For a given system size $m$, fluxes $f : \mathbb{R}^m \rightarrow \mathbb{R}^{m \times 2}$ and $f_v : \mathbb{R}^m \times \mathbb{R}^{m \times 2} \rightarrow \mathbb{R}^{m \times 2}$, we consider viscous balance laws, given in mixed form as

$$\sigma = \nabla w, \quad w_t + \nabla \cdot (f(w) - f_v(w, \sigma)) = g \quad \forall (x, t) \in \Omega \times (0, \infty),$$

equipped with suitable initial and boundary conditions.

A case of particular interest is that of time-dependent Navier-Stokes equations, where the unknowns $w := (\rho, \rho u, \rho v, E)^T$ are density, momentum and total energy, and the fluxes $f \equiv (f_1, f_2)$ and $f_v \equiv (f_v^1, f_v^2)$ are defined by

$$f_1 = (\rho u, p + \rho u^2, \rho uv, u(E + p))^T, \quad f_2 = (\rho v, \rho uv, p + \rho v^2, v(E + p))^T,$$

$$f_v^1 = (0, \tau_{11}, \tau_{21}, \tau_{11}u + \tau_{12}v + kT_{x_1}), \quad f_v^2 = (0, \tau_{12}, \tau_{22}, \tau_{21}u + \tau_{22}v + kT_{x_2}).$$

Here $p$ is the pressure, related to the other variables via the ideal gas law. Furthermore, $\tau$ denotes the stress tensor, $T$ is the temperature, and $k$ is the thermal conductivity coefficient. For the Navier-Stokes equations, $g \equiv 0$.

Boundary conditions on a surface, e.g., an airfoil, are set as adiabatic, no-slip boundary conditions, i.e., one sets $(u, v) \equiv 0$ and $n \cdot \nabla T \equiv 0$. In- and outflow boundary conditions are set via a characteristic splitting of the convective flux.

3. Method. In the sequel, we work on a regular triangulation of $\Omega \subset \mathbb{R}^2$. For the necessary notation, we make the following definition:

**Definition 3.1** (Triangulation). Let $\Omega$ be regularly triangulated as $\Omega = \bigcup_{k=1}^N \Omega_k$. We define an edge $e_k$ as an intersection of two neighboring elements, or an element with the physical boundary $\partial \Omega$, having positive one-dimensional measure. $\Gamma$ denotes the collection of all these intersections, while $\Gamma_0 \subset \Gamma$ denotes those $e_k \in \Gamma$ that do not intersect the physical boundary $\partial \Omega$ of the domain. We define $\hat{N} := |\Gamma|$ to be the number of edges in $\Gamma$.

To simplify the notation, we introduce the following abbreviations for integration:

$$(f_1, f_2) := \sum_{k=1}^N \int_{\Omega_k} f_1 \cdot f_2 \, dx,$$

$$(f_1, f_2) = \sum_{k=1}^{\hat{N}} \int_{e_k} f_1 \cdot f_2 \, d\sigma, \quad (f_1, f_2)_{|\partial \Omega_k} := \sum_{k=1}^N \int_{\partial \Omega_k} f_1 \cdot f_2 \, d\sigma.$$

The method to be presented depends on a triple of function spaces. For a given triangulation, we thus make the following definition:
Definition 3.2. For the approximation of $\sigma_h \approx \nabla w$, $w_h \approx w$ and $\lambda_h \approx w|_\Gamma$, we consider the following spaces:

$$V_h := \{ f \in L^2(\Omega) | f|_{\partial\Omega_k} \in \Pi^p(\Omega_k) \quad \forall k = 1, \ldots, N \}^m$$
$$H_h := \{ f \in L^2(\Omega) | f|_{\partial\Omega_k} \in \Pi^q(\Omega_k) \quad \forall k = 1, \ldots, N \}^{2m}$$
$$M_h := \{ f \in L^2(\Gamma) | f|_{e_k} \in \Pi^q(e_k) \quad \forall k = 1, \ldots, \tilde{N} \}^m.$$

Recall that $m$ denotes the dimension of the system, i.e., $m = 4$ for the Navier-Stokes equations, and $m = 1$ for a scalar equation. $\Pi^p(U)$ is the space of polynomials up to order $p$ on a domain $U$. On the relation of $p$ and $q$, see Remark 1.

Remark 1. For $q = p + 1$, our method is inspired by a Hybrid Mixed method [2], while for $q = p$, it is indeed a hybridized Discontinuous Galerkin method [16, 17, 20]. We demonstrate numerical results for both choices.

We start by semi-discretizing (1) in a straightforward, DG-like manner as

$$(\sigma_h - \nabla w_h, \tau_h) - (\lambda_h - w_h, \tau_h \cdot n)\sigma_h = 0 \quad \forall \tau_h \in H_h$$

$$(w_h, \varphi_h) - (f(w_h) - f_v(w_h, \sigma_h), \nabla \varphi_h) + \langle (\tilde{f} - \tilde{f}_v) \cdot n, \varphi_h \rangle = (g, \varphi_h) \quad \forall \varphi_h \in V_h$$

$$\langle (\tilde{f} - \tilde{f}_v) \cdot n, \mu_h \rangle = 0 \quad \forall \mu_h \in M_h$$

with numerical fluxes

$$\tilde{f} := f(\lambda_h) - \beta (\lambda_h - w_h^-) n, \quad \tilde{f}_v := f_v(\lambda_h, \sigma_h), \quad \gamma (\lambda_h - w_h^-) n.$$
which cannot be resolved well by a low-order scheme. In Fig. 1 on the right, we
into the asymptotic regime. This is due to the rich spatial structure of the solution,
which usually has less degrees of freedom than the corresponding DG discretization.

4.1. Linear convection-diffusion equation. We consider a test case that has
this problem is diffusion-dominated, as the convective flux vanishes. Away from
the domain \( \Omega \) is defined by \( \Omega := [0, 5] \times [-5, 5] \).

Numerical Results.

If we apply BDF to the scheme defined in (4), we obtain
\[
\alpha_0 T(w_h^{n+1}, \varphi_h) + \Delta t N(\sigma_h^{n+1}, w_h^{n+1}, \lambda_h^{n+1}, \tau_h, \varphi_h, \mu_h) = -\sum_{i=1}^{l} \alpha_i T(w_h^{n+1-i}, \varphi_h),
\]
which has to hold for all \((\sigma_h, \varphi_h, \mu_h) \in H_h \times V_h \times M_h\).

For BDF schemes of order \( l \geq 2 \), it is necessary to have a 'startup' phase, i.e.,
one has to compute \((\sigma_h^k, w_h^k, \mu_h^k)\) for \( k = 1, \ldots, l - 1 \) with a different method. We
use the following strategy: For BDF2, it is enough to compute \((\sigma_h^1, w_h^1, \mu_h^1)\) by
an implicit Euler step, as this method generates (in one step) approximations of
\( O(\Delta t^2) \). Unfortunately, for BDF3, such a simple strategy is not sufficient any more.
To compute \((\sigma_h^2, w_h^1, \mu_h^2)\) and \((\sigma_h^3, w_h^2, \mu_h^3)\), we use a BDF2 scheme with time step
\( \Delta t := (\Delta t)^{3/2} \). This ensures that both quantities are approximations to the exact
quantities of \( O(\Delta t^3) \).

Obviously, the scheme in its current form has a lot of degrees of freedom. However,
the unknowns associated to \( \sigma_h \) and \( w_h \) on a cell \( \Omega_k \) are not directly coupled to
the unknowns on another cell \( \Omega_{k'} \), they are only implicitly coupled via \( \lambda \). Thus, via
local solution procedures (or, on the linear algebra level, via static condensation)
[4], it is possible to express both \( \sigma_h \) and \( w_h \) as functions of \( \lambda_h \) only. Therefore, for
each time step we obtain a nonlinear system of equations
\[
M(\lambda_h, \mu_h) = b(\mu_h) \quad \forall \mu_h \in M_h,
\]
which usually has less degrees of freedom than the corresponding DG discretization.

4. Numerical Results.

4.1. Linear convection-diffusion equation. We consider a test case that has
also been investigated by Nguyen et al. [16]. It is a scalar and linear convection-
diffusion equation with the convective flux given by \( f(w) = (-4y, 4x)^T w \), and the
diffusive flux given by \( f_d(w, \nabla w) = 0.1 \nabla w \). The source term \( g \) is set to 0, and the
domain \( \Omega \) is defined by \( \Omega := [-0.5, 0.5]^2 \). Note that in the vicinity of the origin,
this problem is diffusion-dominated, as the convective flux vanishes. Away from
(0,0), convection gets more and more dominant. Based on the initial distribution,
the exact solution can be chosen to be a rotating Gaussian distribution. We use
a hybrid mixed method for this test case, i.e., we set \( g = p + 1 \). For \( p = 0, 1, 2 \),
we perform a numerical convergence study. For \( p = 0 \), we use the implicit Euler
scheme, for \( p = 1 \) we use a BDF2 scheme, and for \( p = 2 \), we use a BDF3 scheme. In
cases, the optimal order for \( w_h \)-convergence should be \( p + 1 \), which is observed,
see Fig. 1 on the left. Note that for \( p = 0 \), the scheme needs much more time to get
into the asymptotic regime. This is due to the rich spatial structure of the solution,
which cannot be resolved well by a low-order scheme. In Fig. 1 on the right, we
have plotted \( \sigma_h \)-convergence. For this test case, the order to be expected is \( p + 1 \)
(which is also achieved), although \( \sigma_h \) is approximated in a space of polynomial order
\( p + 1 \). This lack of optimality is due to the need for stabilization of the convective

<table>
<thead>
<tr>
<th>Name</th>
<th>( l )</th>
<th>( \alpha_0 )</th>
<th>( \alpha_1 )</th>
<th>( \alpha_2 )</th>
<th>( \alpha_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>implicit Euler</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BDF2</td>
<td>2</td>
<td>3/2</td>
<td>-2</td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>BDF3</td>
<td>3</td>
<td>11/6</td>
<td>-3</td>
<td>3/2</td>
<td>-1/3</td>
</tr>
</tbody>
</table>

Table 1. BDF schemes
terms, and is a common feature in mixed discretizations. For implications and ways to overcome this when diffusion is dominating, we refer to [7, 20].

Figure 1. Convergence for the linear test case: \( w_h \) convergence (left) and \( \sigma_h \) convergence (right).

4.2. Burgers equation. The second test case under consideration is the multidimensional (viscous) Burgers equation with convective flux \( f(w) = \frac{1}{2}(w^2, w^2) \) and viscous flux from Section 4.1, given by \( f_v(w, \nabla w) = 0.1 \nabla w \). The source term \( g \) is set in such a way that the exact solution for problem (1) can be given by

\[
    w(x, y, t) := e^{-t} \left( x + e^{10x} - \frac{1}{1 - e^{10}} \right) \left( y + e^{10y} - \frac{1}{1 - e^{10}} \right).
\]

For this problem, we also use a hybrid mixed method, and perform for \( p = 0, 1, 2 \) a numerical convergence study. Again, temporal and spatial orders are chosen correspondingly. In all cases, convergence orders of \( p + 1 \) are achieved for both \( w_h \) and \( \sigma_h \).

Figure 2. Convergence for Burgers equation: \( w_h \) convergence (left) and \( \sigma_h \) convergence (right).

4.3. Navier-Stokes equations. The flow past a circular cylinder is a classical example of a bluff body flow. When the Reynolds number is smaller than 50 the flow is steady and symmetric about the centerline of the wake. Even at small values of the Reynolds number, say \( Re = 10 \), the flow separates from the surface of the cylinder and forms a pair of bound vortices in the near wake. At \( Re = 50 \) this configuration becomes unstable and the process of vortex shedding begins, resulting in the well-known Karman vortex street.

We examine one configuration, defined by a Mach number \( Ma = 0.2 \) and a Reynolds number (based on the diameter of the cylinder) \( Re = 180 \). This test
case was also investigated in [8]. The employed mesh consists of 2916 elements and extends to 20 diameters away from the cylinder (see Fig. 3). At the cylinder wall no-slip conditions are applied; the outer boundary is modeled by characteristic far field conditions. We use the hybridized Discontinuous Galerkin method, i.e. \( p = q \). The computations are initialized with free stream conditions.

![Figure 3. Close-up view of the employed mesh for the cylinder test case.](image)

In Tables 2 and 3, the Strouhal numbers and the temporally averaged drag coefficients for various spatial and temporal discretizations are listed. Both numbers compare very well with shrinking time step to values from the literature (see Table 4). In Fig. 4, the Mach number is plotted at four temporal instances of the shedding cycle. The periodic nature of vortex shedding can be observed remarkably well (compare the first and the third, and the second and the fourth plot, respectively).

<table>
<thead>
<tr>
<th>( \Delta t = 1 )</th>
<th>( \Delta t = 5 )</th>
<th>( \Delta t = 10 )</th>
<th>( \Delta t = 5 )</th>
<th>( \Delta t = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p = 1 )</td>
<td>0.1898</td>
<td>0.1585</td>
<td>0.1205</td>
<td>0.1931</td>
</tr>
<tr>
<td>( p = 2 )</td>
<td>0.1898</td>
<td>0.1618</td>
<td>0.1255</td>
<td>0.1964</td>
</tr>
<tr>
<td>( p = 3 )</td>
<td>0.1898</td>
<td>0.1618</td>
<td>0.1255</td>
<td>0.1964</td>
</tr>
</tbody>
</table>

**Table 2.** Strouhal number for various spatial and temporal discretizations

<table>
<thead>
<tr>
<th>( \Delta t = 1 )</th>
<th>( \Delta t = 5 )</th>
<th>( \Delta t = 10 )</th>
<th>( \Delta t = 5 )</th>
<th>( \Delta t = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p = 1 )</td>
<td>1.3448</td>
<td>1.2449</td>
<td>1.0675</td>
<td>1.4604</td>
</tr>
<tr>
<td>( p = 2 )</td>
<td>1.3640</td>
<td>1.2486</td>
<td>1.0744</td>
<td>1.4972</td>
</tr>
<tr>
<td>( p = 3 )</td>
<td>1.3634</td>
<td>1.2490</td>
<td>1.0727</td>
<td>1.4795</td>
</tr>
</tbody>
</table>

**Table 3.** Temporally averaged drag coefficient for various spatial and temporal discretizations

5. **Conclusions and Outlook.** We have presented a hybrid mixed method for the computation of time-dependent convection-diffusion equations, and we have demonstrated performance by numerical studies. One disadvantage of BDF methods is that they do not allow for (arbitrary) high-order discretizations, as they become unstable for \( l \geq 7 \). Therefore, one near-future project is the incorporation of (embedded) DIRK schemes into our method, see [13].
<table>
<thead>
<tr>
<th>Experiment</th>
<th>$c_D$</th>
<th>$Sr$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gopinath [8]</td>
<td>1.3406</td>
<td>0.1866</td>
</tr>
<tr>
<td>Henderson [11]</td>
<td>1.336</td>
<td>-</td>
</tr>
<tr>
<td>Williamson [22]</td>
<td>-</td>
<td>0.1919</td>
</tr>
</tbody>
</table>

Table 4. Mean drag coefficients and Strouhal numbers from the literature

Figure 4. Four subsequent snapshots of the vortex shedding cycle ($p = 3$, BDF2, $\Delta t = 1$).

Furthermore, stability of our method has not yet been investigated theoretically. In the scalar, linear case, it is known that for the limiting cases of the underlying equation (1), i.e., $f \equiv 0$ or $f_v \equiv 0$, the semi-discretization in (4) is $L^2$-stable. ($f_v \equiv 0$: See [14] for the purely convective DG case. $f \equiv 0$: Substitute $(\tau_h, \varphi_h, \mu_h) = (\sigma_h, w_h, \lambda_h)$ in (4) and use standard arguments to obtain $\|\sigma\|_{L^2(\Omega)} + \frac{1}{\pi} \|w_h\|_{L^2(\Omega)} = 0$.) However, this has yet to be shown for the full convection-diffusion equation. Also for the fully discretized scheme, stability has still to be shown.

REFERENCES


E-mail address: schuetz@igpm.rwth-aachen.de
E-mail address: woopen@aices.rwth-aachen.de
E-mail address: may@aices.rwth-aachen.de
FINITE VOLUME SCHEMES ON 2D NON-UNIFORM GRIDS

GABRIELLA PUPPO
Dip. di Scienza e Alta Tecnologia, Università Insubria
Via Valleggio 11
22100, Como, Italy

MATTEO SEMPLICE
Dip. di Matematica, Università di Torino
Via Carlo Alberto 10
10123 Torino, Italy

Abstract. In this work we briefly describe a technique to define second order finite volume schemes on non uniform cartesian grids. The purpose is to couple this scheme with an error indicator to drive mesh adaptivity. In this context, it is crucial that the underlying scheme uses a very compact stencil. Here we illustrate an algorithm that matches this goal, without giving up accuracy, while relying on a non oscillatory reconstruction.

1. Introduction. We are interested in the integration in 2D of systems of conservation laws of the form

\[ u_t + f_x(u) + g_y(u) = 0 \] (1)

where the system is assumed to be hyperbolic, namely any linear combination of the Jacobian matrices \( f'(u) \) and \( g'(u) \) is diagonalizable with real eigenvalues. The purpose of this work is to integrate such a system with a finite volume, second order scheme, using a locally adaptive grid with a cartesian structure. The advantage of this procedure is that schemes based on cartesian grids are easier to parallelize than schemes based on unstructured grids, because it is easy to store the grid in a tree structure, allowing for easier communication between the grid elements. It is frequently said that the main drawback of cartesian grids is due to the difficulty in dealing with curved boundaries, but this problem can be solved using a combination of level sets to describe the boundary and the ghost fluid method, see [1], [2], and it will not be discussed further in this work.

The scheme described here will be applied in the context of adaptive grid refinement, driven by an error indicator based on the entropy production, as in [3]. As opposed to the rectangular patches of finer grids used in [4], our approach allows to refine every single cell independently of its neighbors and thus, even in the cartesian setting, the location of cells in the neighborhood and their sizes can vary in a wide pool of possible patterns. Moreover, the amplitude of each cell may change as a result of the error diagnosed within a time step, and in particular the solution just computed in a cell that undergoes refinement must be recomputed on the refined grid.

2000 Mathematics Subject Classification. Primary: 65M08; Secondary: 65M50.

Key words and phrases. Finite Volume Schemes, Conservation Laws, Non-uniform cartesian grids.

This work is partly supported by PRIN-2009588FHJ002.
cells to improve its accuracy. Consequently the stencil on which the scheme is based must be as compact as possible, so that the necessity of recomputing the solution spreads as little as possible on neighboring cells. The reconstruction proposed here is targeted on this particular requirement.

2. Reconstruction. We suppose that the computational domain is covered with a non uniform grid composed of non overlapping square cells. Each cell has a side which is a submultiple by a power of 2 of a basic length $h$. The cells are labelled with a single integer index, $j = 1, \ldots, N$. The cell $V_j$ is centered around the point $(x_j, y_j)$ and has a width $h_j = 2^{-l_j} h$, where $l_j$ denotes the current cell level of refinement, $l_j = 0, \ldots, l_{\text{max}}$, and $l_{\text{max}}$ is the maximum level of refinement. The grid structure is stored in a tree, as described in \[3\]. The leaves of the tree represent the active cells. Note that the following procedure can be extended to rectangular cells, with some tedious modifications.

The time step starts with a knowledge of the cell averages of the unknown $u$ on each cell, $u_j$, for $j = 1, \ldots, N$. The first task is to compute a piecewise linear reconstruction from the cell averages. Suppose that $v(x, y)$ is a smooth function with cell averages $\bar{u}_j$ on the current grid. In this case the reconstruction ought to be given by

$$u(x, y) = \sum_{j=1}^{N} P_j^1(x, y) \chi_j(x, y),$$

where $\chi_j$ is the characteristic function of the $j$-th cell, and

$$P_j^1(x, y) = \bar{u}_j + \sigma_j^x (x - x_j) + \sigma_j^y (y - y_j),$$

with

$$\sigma_j^x = \partial_x v|(x_j, y_j) + O(h_j), \quad \sigma_j^y = \partial_y v|(x_j, y_j) + O(h_j), \quad (2)$$

so that $u(x, y) - v(x, y) = O(h_j^2)$. Moreover, the reconstruction must be non-oscillatory. The reconstruction is computed starting from the large cells in the grid. We illustrate the evaluation of the slope $\sigma_j^x$ in the two cases illustrated in Fig. 1, which account for the typical cases that may occur. We need a left and a right slope, $\sigma_j^L$ and $\sigma_j^R$, from which the actual slope will be computed via the MinMod function, or some other limiter. Referring to the figure, in both cases, the left slope is simply given by $\sigma_j^L = (\bar{u}_j - \bar{u}_1)/(x_j - x_1)$.

![Figure 1. Computing the horizontal slopes](image)

The right slope is different in the two cases shown in the figure. In the example appearing on the left, the cell $j$ interfaces two smaller cells, with the same level of refinement. Clearly the centers of these cells have the same abscissas, and one can define

$$\sigma_j^R = c_2(\bar{u}_j - \bar{u}_2)/(x_j - x_2) + c_3(\bar{u}_j - \bar{u}_3)/(x_j - x_3),$$
where $c_2$ and $c_3$ are two positive weights adding up to 1. It is clear that if the weights are chosen equal (thus in this case, $c_2 = c_3 = \frac{1}{2}$), $\sigma^R_j$ satisfies the requirement (2). The same behaviour occurs in the case in which the cell $j$ interfaces with a larger number of cells, but all of them have the same level of refinement: we just need to modify the weights, which will be $c_r = 2^{-l_r - l_j}$. The second case (right of Fig. 1) is trickier. Here, the cells on the right do not have a uniform size, and their cell centers do not have the same abscissas. We still compute

$$
\sigma^R_j = c_2 \frac{u_j - u_2}{x_j - x_2} + c_3 \frac{u_j - u_3}{x_j - x_3} + c_4 \frac{u_j - u_4}{x_j - x_4}
$$

$$
= \left( \sum_{k=2}^{4} c_k \right) v_x + \left( \sum_{k=2}^{4} c_k \frac{y_j - y_k}{x_j - x_k} \right) v_y + O(h)
$$

The coefficients in the first sum must add up to 1, while the sum of the coefficients in the second term must be zero. We further choose $c_3 = c_4$. In this fashion, recalling that $h_3 = h_4 = \frac{1}{4} h_j$; $h_2 = \frac{1}{2} h_j$, one easily finds $c_2 = \frac{6}{11}, c_3 = c_4 = \frac{5}{22}$, which give the correct weights to ensure that the slope in the large cell has the desired accuracy, as in (2). Clearly, Fig. 1 does not represent all possible cases. What is left is the possibility that the large cell $j$ borders with cells of mixed type that have an even lower degree of refinement. In this case, it is still possible to compute constant weights, depending on the local pattern of refinement, but not on the solution itself, which ensure that the correct accuracy is matched, but this would complicate the scheme exceedingly. Alternatively, one could enlarge the stencil, including also the cells immediately to the right of $V_3$ and $V_4$ (namely the cells $V_5$ and $V_6$ in the figure), but this alternative would make the stencil less compact.

On the other hand, we note that $c_3$ is close to $\frac{1}{4}$, while $c_2$ is quite close to $\frac{1}{2}$. Thus in general we choose:

$$
c_i = \begin{cases} 
1 & \text{if } h_i = h_j \\
\frac{6}{11} & \text{if } h_i/h_j = \frac{1}{2} \\
\frac{5}{22} & \text{if } h_i/h_j = \frac{1}{4} \\
\frac{1}{2^{l_i-l_j}} & \text{if } h_i/h_j = 2^{l_j-l_i} < \frac{1}{4},
\end{cases}
$$

where $j$ denotes the index of the cell on which the reconstruction is sought, and $i$ is the index of the neighboring cell which contributes to the slope in the cell $V_j$. The right slope will then be computed as

$$
\sigma^R_j = \left( \sum_i c_i \frac{u_j - u_i}{x_j - x_i} \right) / \sum_i c_i,
$$

where the sum is extended to all cells on the right of the cell $V_j$. In this fashion, accuracy is preserved exactly in the two common cases in which an edge of a cell $V_j$ is either facing cells which are all of the same size (even if they are much smaller) or is facing cells which are at most two levels finer (in any possible disposition). On the other hand accuracy is almost preserved in the unlikely case in which a cell is surrounded by cells with a very uneven refinement pattern. In all cases, the stencil consists only of the cells that are adjacent to the current cell.

Once the reconstruction of each cell of a given level $l$ has been obtained, the algorithm proceeds to the evaluation of the reconstruction of cells of level $l + 1$. For the sake of illustration, we still consider the cases of Fig. 1. We suppose we have
the reconstruction in the cell \(j\) and we compute the reconstruction in the cell \(V_2\). The left slope will be given by
\[
\sigma_{L}^{j} = \frac{P_1(x_j, y_j - h_2/2) - \pi_2}{x_j - x_2},
\]
while for the right slope we use the algorithm described above. In this case too accuracy is preserved. This completes the description of the evaluation of the reconstruction.

Equation (1) is written in finite volume formulation, namely
\[
\frac{d\pi_j}{dt} = -\frac{1}{h^2} \int_{\partial V_j} [F, G]^T(t) \cdot n,
\]
where \(F\) and \(G\) denote the numerical fluxes, consistent with \(f\) and \(g\) respectively, \(\partial V_j\) is the boundary of the cell \(V_j\), and \(n\) is the external normal to \(\partial V_j\). The line integrals are computed with the mid point rule, and the quadrature nodes depend on the cells surrounding the cell \(V_j\). To compute the numerical fluxes, we need two boundary extrapolated data, across each quadrature node, which are computed from within each cell using the reconstruction. The location where the boundary extrapolated data are computed is shown by the crosses in Fig. 1. Once the boundary extrapolated data are known, we compute the numerical fluxes using the Local Lax Friedrichs formula. For the time integration, a second order explicit TVD Runge Kutta scheme is applied.

The main focus of this paper is on the numerical integration of 2D problems on non uniform grids, but we also provide, as a comparison, results obtained using an adaptive grid, in which the level of refinement and coarsening is computed dynamically, following the solution, with the aid of an a-posteriori error indicator. In this paper, we consider the error indicator based on the numerical residual in an entropy inequality for the original equation (1), see [3] and [5]. Let \(\eta\) and \(\psi\) be an entropy-entropy flux pair for (1). Then we define the density of entropy production for the cell \(V_j\) during the time step \(t^n, t^{n+1}\) as
\[
S_j^n = \frac{1}{\Delta t} \left( \eta(\bar{u}^{n+1}_j) - \eta(\bar{u}^{n+1}_j) + \frac{\Delta t}{h^2} \sum_{i=1}^{2} \int_{\partial V_j} \Psi(i) \right)
\]
where \(\Psi(i)\) are numerical entropy fluxes, consistent with \(\psi\), computed corresponding to the \(i\)-th Runge Kutta stage value of the solution \(u\). The integrals are evaluated with the same quadrature points as in the time advancement scheme, and the numerical entropy fluxes are based on the boundary extrapolated data already computed to advance the solution in time. Here, we will use a variant of the Local Lax Friedrichs formula. For more details, see [3, 5]. For alternative error indicators, see for example [6] and [7].

3. An example. We consider a linear advection problem \(u_t + a(x, y)u_x + b(x, y)u_y = 0\), in which a bump rotates around the origin. The computational domain is the square \([-2, 2] \times [-2, 2]\). The initial condition is
\[
u_0(x, y) = (1 - 4r^2)^3 \chi_{(r \leq \frac{1}{2})}, \quad r = ||(x, y) - (0.75, 0.75)||_2,
\]
that is \(u_0\) is a peaked hump, centered in the top right quadrant, that is prolongated to zero with a \(C^2\) degree of smoothness. Note that \(\max(u_0) = 1\), and that the maximum is quite narrow. The rotation field is \(a(x, y) = -2\pi y, b(x, y) = 2\pi x\). In this fashion, the bump completes a full rotation in 1 unit of time. During its
rotation, the hump will intersect the grid with no predefined angle. We devised
this test to remove any effect due to a biased alignment of the grid with the data,
which often pollutes results with cartesian meshes.

Fig. 2 shows the solution obtained after one complete rotation. The column on
the left shows the solution obtained on a uniform grid, with $h = 4/2^i, i = 4, 5, 6$
from top to bottom. From the color bar we note that the height of the peak
decreases slightly faster than linearly, as a consequence of the artificial diffusion
induced by the MinMod limiter. Moreover, the effect of the numerical error is also
apparent in the deformation of the peak in the direction of its advection, especially
on the coarse grid. The right column contains the solution obtained on several non
uniform grids based on 3 levels of refinement. In the figure at the top the level of refinement is chosen randomly (grid $R$), and naturally this provides the worst case. The shape of the bump is similar to the coarse grid case, although the spreading is more contained. Note also that the height of the peak is less smeared than on the coarse grid. In the middle row, the grid is refined outside a circle of radius 0.75, so that there is a always discontinuity in the grid size at the center of the hump during the whole revolution (grid $C$). Here one can note that there is no apparent distortion due to grid effects (the hump preserves its shape, although of course it is more diffused in the portion solved by the coarse grid). Finally, the figure at the bottom is obtained with the adaptive grid algorithm. Here the patch on which the grid is refined travels with the hump, and the quality of the solution is the same as the one provided when the uniform fine grid is used throughout (grid $A$).

![Figure 3. Error as a function of time on several grids. The dashed lines refer to uniform grids, while the green curves refer to non uniform grids with the structure illustrated on the left of the figure.](image)

Fig. 3 shows the behavior of the error as a function of time when the grid has the shape seen in the left part of the figure. The dashed lines correspond to the error computed on several uniform grids with mesh size $h = 4/2^l$: from the top the bottom curve (which corresponds to the minimum error) $l = 6, \ldots 10$. The green curves are the errors obtained on the grid shown on the left, and they involve three levels of refinement, i.e. $h = 4/2^l, 4/2^{l+1}, 4/2^{l+2}$. Here $l = 6$ for the curve at the top, while $l = 8$ for the curve at the bottom. The behavior of the green curve shows clearly that the error grows faster while the hump crosses the coarse patches in the grid, while it grows slowly when the patch is located on the refined parts. In any case, the behavior of the error remains smooth, with no oscillations or accuracy losses due to the irregularity in the grid.

Finally, Fig. 4 contains the behavior of the error with time for the solutions obtained on the three non uniform grids shown on the right of Fig. 2. The dashed lines refer to results obtained on uniform grids, as in the previous figure. The green curves are obtained on each of the grids illustrated in Fig. 2, but they involve cells with different levels of refinement, as in the previous figure. Clearly the random grid $R$ provides the highest error, which is quite close to the one corresponding to its coarsest grid. It is noteworthy that the adaptive grid yields an error which almost coincides with the error found on its finest grid, although the number of grid points is much smaller.
Figure 4. Error versus time for the three grids shown in Fig. 2, for different levels of refinement. From left to right, grids $R$, $C$ and $A$. The dashed lines refer to uniform grids, while the green curves refer to solutions obtained on non uniform grids.

Acknowledgments. All simulations were performed with dune-fv, a module for computing solutions of conservation laws written by the second author and depending on the DUNE [8] and the ALUGRID [9] libraries. The source code (GPL licence) is available from the web-pages [8, 9] and upon request to the second authos (for dune-fv).

REFERENCES


E-mail address: gabriella.puppo@uninsubria.it
E-mail address: matteo.semplice@unito.it
THEORETICAL STUDY OF ENTROPY DISSIPATION OF MOVING MESHES

MARIA LUKÁČOVÁ-MEDVID’OVÁ AND NIKOLAOS SFAKIANAKIS

Institute for Mathematics
Johannes Gutenberg - University
Staudingerweg 9
Mainz, 55099, Germany

Abstract. Non-uniform grids and mesh adaptation have been a growing part of numerical simulation over the past years. It has been experimentally noted that mesh adaptation leads not only to locally improved solution but also to numerical stability of the underlying method. There have been though only few results on the mathematical analysis of these schemes (see for example [15]) due to the lack of proper tools that incorporate both the time evolution and the mesh adaptation step of the overall algorithm.

In this paper we provide a method to perform the analysis of the mesh adaptation method, including both the mesh reconstruction and evolution of the solution. We moreover employ this method to extract sufficient conditions -on the adaptation of the mesh- that stabilize a numerical scheme in the sense of the entropy dissipation.

1. Introduction. Hyperbolic conservation laws appear in various applications. For example, fundamental physical laws, the conservation of mass momentum and energy, lead to the Euler equations of gas dynamics. Further examples arise in traffic flows, shallow water flows, magnetohydrodynamics and biology, see, e.g. [13, 20, 9, 8].

Let us consider a scalar conservation law in one space dimension,

\[ u_t + f(u)_x = 0, \quad x \in [a, b], \quad t \in [0, T], \]

(1)

with initial data \( u_0 \in L^\infty([a, b]) \). In order to simplify the presentation we assume e.g. periodic boundary conditions.

Adaptivity is a main theme in modern scientific computing of complex physical phenomena. It is important to investigate the behaviour of adaptive schemes for hyperbolic problems, such as (1), which exhibit several interesting and not trivial characteristics. In this work we study the behaviour of certain geometrically driven adaptive algorithms when combined with the important class of entropy conservative schemes introduced in [17, 18].

In every time step \( t = t^n \) the mesh we consider is:

\[ M^n = \{ a = x_1^n < \cdots < x_N^n = b \} \]

2000 Mathematics Subject Classification. Primary: 35Lxx; Secondary: 65Mxx.

Key words and phrases. Hyperbolic conservation laws, adaptive mesh reconstruction, entropy dissipation, finite volume schemes.

The second author is supported by the Alexander von Humboldt Foundation and the Center of Computational Sciences in Mainz.
with variable space step sizes $h^n_i = x^n_{i+1} - x^n_i$, $i = 1 \ldots N - 1$; the mesh is reconstructed in every time step $t^n$. Further, we consider a numerical approximation $U^n$ of the exact solution $u$ over the mesh $M^n_x$ at time $t = t^n$ given as a
$$U^n = \{u^n_1, \ldots, u^n_N\}.$$ 

The construction and evolution of our non-uniform meshes and the time evolution of the approximate solutions is dictated by the Main Adaptive Scheme (MAS) which is described by the following procedure:

- in every time step, construct new mesh according to the prescribed adaptivity criterion,
- reconstruct the numerical solution over the new mesh,
- evolve the numerical solution in time using the numerical scheme.

MAS will be discussed in details in Section 2; we note here that the number of spatial nodes is fixed and that the reconstruction of the mesh is realized by moving its points according to the geometry of the numerical solution.

The use of non-uniform adaptively redefined meshes, in the context of finite differences, was first studied, among others, in [10], [7], and [19]. The approach that we follow, for the mesh reconstruction step of MAS (Step 1) was first introduced in [4]. Applications of MAS on several problems, point out a strong stabilisation property emanating from the mesh reconstruction [3], [4], [2], [15], [16]. These stabilization properties led [2] to combine MAS with the marginal class of entropy conservative schemes. The later were first introduced in [17] and further studied in [12, 18, 14]. They are semi-discrete numerical schemes which satisfy an exact entropy equality. On one hand these schemes are interesting on their own right, since they appear in the context of zero dispersion limits, complete integrable systems and computation of non-classical shocks. On the other hand they are important as building blocks for the construction of entropy stable schemes [17, 18], [14].

We note that classical techniques for the analysis of numerical schemes, only include the time evolution step of the procedure. In order though to have a complete picture of the quality of the numerical solution under a mesh adaptation procedure a broader analysis is needed. In this direction the work done in [16] has provided some constructive analysis tools. In the present paper though we are able to combine in one relation the effect of both the time evolution and the mesh adaptation. Let us point out that our approach allows us to represent the effects of both the adaptive mesh reconstruction as well as finite volume scheme in one conservative update relation over a reference uniform mesh (21). This approach allows to apply the entropy stability analysis and derive a sufficient mesh adaptation criterion to control entropy production.

2. Main Adaptive Scheme (MAS). Using non-adaptive meshes -both uniform and non-uniform- the evolution of the numerical solution is dictated solely by the solution update. On the contrary, in the adaptive mesh case, two more phenomena need to be taken into account; the construction of the new mesh and the solution update. These steps comprise the Main Adaptive Scheme (MAS):

**Definition 2.1 (MAS).** Given mesh $M^n_x$ and approximations $U^n$,

1. (Mesh Reconstruction). Construct new mesh $M^{n+1}_x$
2. (Solution Update). Reconstruct $U^n$ over $M^{n+1}_x$ to obtain $\hat{U}^n$.
3. (Time Evolution). Evolve $\hat{U}^n$ in time to compute $U^{n+1}$ over $M^{n+1}_x$. 

It is worth noting that in the case of a fixed mesh, uniform or non-uniform, MAS reduces to just the time evolution step (Step 3.) which is what is usually considered as a numerical scheme.

In the mesh reconstruction step (Step 1.) the mesh nodes are relocated according to the geometric information contained in the numerical solution. The basic idea is geometric:

\textit{in areas where the numerical solution is more smooth/flat the density of the nodes is low, in areas where the numerical solution is less smooth/flat the density of the nodes should be higher.}

In fact, the mesh reconstruction process can be chosen in any suitable way. One possibility is to use the monitor function which reflects the curvature of the numerical solution. For details we refer [16, 4].

Further, we consider the solution update procedure (Step 2. of MAS). The numerical solution \( U^n = \{ u^n_i, i = 1 \ldots N \} \), \( M^x_n = \{ C^n_i, |C^n_i| = h^n_i, i = 1 \ldots N \} \), we obtain in the case of finite volume scheme

\[ u^{n+1}_i = \hat{u}^n_i - \frac{\Delta t}{h^i_{n+1}} \left( \hat{F}^n_{i+1/2} - \hat{F}^n_{i-1/2} \right). \] (2)

Here \( \hat{U}^n = \{ \hat{u}^n_i, i = 1 \ldots N \} \) is a reconstructed \( U^n \) over \( M^{x+1}_n \) and the numerical flux \( F \) is decorated with \( \hat{\cdot} \) since it is computed over the updated values \( \hat{\cdot} \). The numerical flux itself can be any numerical flux valid for non-uniform grids. We refer to [16] and [4] for more details regarding both the implementation of numerical schemes over non-uniform meshes and their properties.

2.1. Reference uniform mesh. A schematic representation of MAS (Definition 2.1) in the form of mesh-solution pairs is the following

\[ \{ M_x^n, U^n \} \xrightarrow{\text{mesh adapt.}} \{ M_x^{n+1}, \hat{U}^n \} \xrightarrow{\text{num. scheme}} \{ M_x^{n+1}, U^{n+1} \}, \] (3)

where in the first part we have considered the Steps 1 and 2 of MAS and in the second part the Step 3.

In parallel to MAS and (3) we define a new set of mesh-solution pairs where the meshes are uniform, constant in time, of the same cardinality as \( M_x^n \) and discretizing the same physical domain.

Definition 2.2 (Reference uniform mesh-solution pair). Let \( \{ M_x, U \} \) and \( \{ \bar{M}, \bar{V} \} \) be two mesh-solution pairs with \( M_x = \{ C_i, |C_i| = h_i, i = 1 \ldots N \} \), \( \bar{M} = \{ \bar{C}_i, |\bar{C}_i| = \Delta x, i = 1 \ldots N \} \), \( U = \{ u_i, i = 1 \ldots N \} \), and \( \bar{V} = \{ v_i, i = 1 \ldots N \} \). We call \( \{ M, V \} \) the reference uniform mesh-solution pair to \( \{ M_x, U \} \) if

- the meshes \( M_x \) and \( \bar{M} \) discretize the same physical domain, and
- the following per-cell mass conservation is satisfied for every \( i = 1, \ldots, N \)

\[ \Delta x v_i = h_i u_i. \] (4)
Geometric conservation law (GCL). Let us consider a a time dependent cell \( C(\tau) = (x_1(\tau), x_2(\tau)) \). We look for an appropriate conservation law,

\[
u_t(x, t) + \xi(u(x, t), x)_x = 0,
\]

that expresses the mass conservation of \( u \) over the moving cell \( C(\tau) \). Thus, by the Leibniz rule,

\[
\frac{d}{d\tau} \int_{x_1(\tau)}^{x_2(\tau)} u(x, \tau) dx = u(x_2(\tau), \tau)x'_2(\tau) - u(x_1(\tau), \tau)x'_1(\tau) + \int_{x_1(\tau)}^{x_2(\tau)} u_t(x, \tau) dx. \tag{6}
\]

If the mass of \( u \) over \( C(\tau) \) remains constant with respect to \( \tau \), the following condition holds

\[
\int_{x_1(\tau)}^{x_2(\tau)} u_t(x, \tau) dx = -u(x_2(\tau), \tau)x'_2(\tau) + u(x_1(\tau), \tau)x'_1(\tau). \tag{7}
\]

Integrating (5) over \( C(\tau) \) we obtain

\[
\int_{x_1(\tau)}^{x_2(\tau)} u_t(x, \tau) dx + \int_{x_1(\tau)}^{x_2(\tau)} \xi(u(x, \tau), x)_x dx = 0.
\]

Now, using (7) we get

\[
\xi(u(x_2(\tau), \tau), x_2(\tau)) - \xi(u(x_1(\tau), \tau), x_1(\tau)) = u(x_2(\tau), \tau)x'_2(\tau) - u(x_1(\tau), \tau)x'_1(\tau).
\]

A suitable flux function \( \xi \) hence is

\[
\xi(u(x(\tau), \tau), x(\tau)) = u(x(\tau), \tau)x'(\tau). \tag{8}
\]

Therefore, the strong formulation of (5) reads

\[
u_t(x, \tau) + (u(x, \tau)x_t)_x = 0, \tag{9}
\]

which is referred in the literature as the Geometric Conservation Law, see e.g. [6].

As previously announced we can attain the per-cell mass conservation property (4) by discretizing the corresponding GCL.

**Lemma 2.3.** The per-cell mass conservation label (4) is a consequence of the geometric conservation law (9).

**Proof.** For every given cell-value pair \( C_i \), \( u_i \) and the respective reference pair \( \bar{C}_i \), \( v_i \) as provided in the Definition 2.2— we set the moving cell \( C(\tau) = (x_1(\tau), x_2(\tau)) \) for \( \tau \in [\tau_1, \tau_2] \) to be a linear interpolation of \( C_i \) and \( \bar{C}_i \)

\[
C(\tau) = \frac{\tau - \tau_1}{\tau_2 - \tau_1} \bar{C}_i + \frac{\tau_2 - \tau}{\tau_2 - \tau_1} C_i.
\]

Now, to attain a discrete version of (9) we integrate it over \( C(\tau) \)

\[
\int_{x_1(\tau)}^{x_2(\tau)} u_t(x, \tau) dx + \int_{x_1(\tau)}^{x_2(\tau)} (u(x(\tau), \tau)x_t(x, \tau))_x dx = 0,
\]

and invoke (6) to get

\[
\frac{d}{d\tau} \int_{x_1(\tau)}^{x_2(\tau)} u(x, \tau) dx - u(x_2(\tau), \tau)x'_2(\tau) + u(x_1(\tau), \tau)x'_1(\tau)
\]

\[+ \int_{x_1(\tau)}^{x_2(\tau)} (u(x(\tau), \tau)x_t(x, \tau))_x dx = 0.
\]
We discretize explicitly in \([\tau_1, \tau_2]\), set \(\Delta \tau = \tau_2 - \tau_1\), and recall that \(|C_i| = h_i\), and \(C_i = \Delta x\) to get
\[
\frac{1}{\Delta \tau} (\Delta x u_i - h_i u_i) - u_i \frac{x_2(\tau_2) - x_2(\tau_1)}{\Delta \tau} + u_i \frac{x_1(\tau_2) - x_1(\tau_1)}{\Delta \tau} + u_i \frac{x_2(\tau_2) - x_2(\tau_1)}{\Delta \tau} - u_i \frac{x_1(\tau_2) - x_1(\tau_1)}{\Delta \tau} = 0
\]
or simply \(\Delta x v_i = h_i u_i\).

We point out that in the theoretical analysis we will use the reference uniform mesh-solution pair to combine the effects of mesh adaptivity and the numerical update.

In view of the Definition 2.2, and after applying the per-cell mass conservation (4) on the schematic representation (3) of MAS, we get for \(i = 1, \ldots, N\):
\[
h_i^n u_i^n = \Delta x v_i^n \xrightarrow{\text{mesh adapt.}} h_i^{n+1} \hat{v}_i^n = \Delta x \hat{v}_i^n \xrightarrow{\text{num. scheme}} h_i^{n+1} u_i^{n+1} = \Delta x v_i^{n+1}.
\]

Now, by invoking (10) after multiplying with \(h_i^{n+1}\), we can rewrite the scheme (2) over the uniform reference mesh
\[
\hat{v}_i^{n+1} = \hat{v}_i^n - \frac{\Delta t}{\Delta x} (\hat{F}_i^{n+1/2} - \hat{F}_i^{n-1/2}),
\]
where \(\hat{F}\) is the numerical flux function from (2) written in variables \(\hat{v}\).

3. Entropy stability. Before stating the main theoretical result we introduce the following notations:
\[
\Delta v_i^{n+1/2} := v_i^{n+1} - v_i^n,
\]
\[
B_i^{n+1/2} := \frac{f(v_i^{n+1}) - f(v_i^n)}{\Delta v_i^{n+1/2}},
\]
\[
Q_i^{n+1/2} := \frac{f(v_i^{n+1}) + f(v_i^n) - 2\hat{F}_i^n}{\Delta v_i^{n+1/2}},
\]
\[
D_i^{n+1/2} := Q_i^{n+1/2} - Q_i^{n+1/2},
\]
\[
\Delta x_i^{n+1/2} := x_i^{n+1/2} - x_i^{n+1/2},
\]
\[
H_i^{n} := \frac{(\Delta x_i^{n+1/2})}{h_i^n} v_i^n - \frac{(\Delta x_i^{n+1/2})}{h_i^{n+1}} v_i^{n+1},
\]
\[
M_i^{n} := \hat{v}_i^n (H_i^{n+1/2} - H_i^{n+1/2}).
\]

Now, we proceed with the main theoretical result.

**Theorem 3.1.** We use the notations (12a)-(12g) and assume that the following condition holds
\[
M_i^n \leq \frac{\Delta t}{4\Delta x} \left( D_{i-1/2}^n - K^3 \frac{\Delta t}{\Delta x} (B_{i-1/2}^n + Q_{i-1/2}^n + D_{i-1/2}^n)^2 \right) (\Delta v_{i-1/2}^n)^2 + \left( D_{i+1/2}^n - K^3 \frac{\Delta t}{\Delta x} (B_{i+1/2}^n - Q_{i+1/2}^n - D_{i+1/2}^n)^2 \right) (\Delta v_{i+1/2}^n)^2.
\]
where $\Delta x$, $\Delta t$ are respectively the space and time steps that correspond to the numerical scheme (2). The mesh adaptation procedure (3) is used, where the corresponding reference uniform mesh is given in the Definition 2.2. Then the mesh adaptation procedure MAS (3) with the numerical scheme (2) for the time evolution step is entropy stable.

**Proof.** The numerical scheme for the uniform variables reads, cf. (11)

$$v_i^{n+1} = \hat{v}_i^n - \frac{\Delta t}{\Delta x}(\hat{F}_i^{n+1/2} - \hat{F}_i^{n-1/2}).$$

We subtract $v_i^n$ to develop the respective incremental form

$$v_i^{n+1} - v_i^n = \hat{v}_i^n - v_i^n - \frac{\Delta t}{\Delta x}(2\hat{F}_i^{n+1/2} - 2\hat{F}_i^{n-1/2}).$$ (15)

Equivalently

$$v_i^{n+1} - v_i^n = \hat{v}_i^n - v_i^n$$

$$- \frac{\Delta t}{2\Delta x}(f(v_i^{n+1}) + f(v_i^n) - f(v_i^{n+1}) + f(v_i^n)) - f(v_i^{n-1}) + f(v_i^{n-1}) - f(v_i^{n-1}) + f(v_i^{n-1}))$$

$$= \hat{v}_i^n - v_i^n$$

$$- \frac{\Delta t}{2\Delta x}((B_i^{n+1/2} - Q_i^{n+1/2}) \Delta v_{i+1/2} + (B_i^{n-1/2} + Q_i^{n-1/2}) \Delta v_{i-1/2}).$$ (16)

We point out that the term $\hat{v}_i^n - v_i^n$ is new and accounts for the mesh reconstruction and solution update steps of the MAS (Definition 2.1).

We now express $\hat{v}_i^n - v_i^n$ in a conservative form with respect to $\{v_{i}^n\}$. Accordingly the size of $C_i$ changes as:

$$h_i^{n+1} = h_i^n + \Delta x_i^{n+1/2} - \Delta x_i^{n-1/2},$$ (17)

and the the mass of $u$ over $C_i$ as:

$$h_i^{n+1}u_i^n = h_i^n u_i^n - (\Delta x_i^{n+1/2} - u_i^n + (\Delta x_i^{n+1/2} + u_{i+1}^n)$$

$$+ (\Delta x_i^{n+1/2} - u_{i-1}^n - (\Delta x_i^{n+1/2} + u_i^n)$$ (18)

recasts to:

$$\hat{v}_i^n = v_i^n - \frac{\Delta x_i^{n+1/2}}{h_i^n} - v_i^n + \frac{\Delta x_i^{n+1/2}}{h_i^{n+1}} v_{i+1}^n + \frac{\Delta x_i^{n+1/2}}{h_i^{n-1}} v_{i-1}^n - \frac{\Delta x_i^{n+1/2}}{h_i^n} v_i^n,$$

which can be written as a conservative difference

$$\hat{v}_i^n - v_i^n = H_i^{n+1/2} - H_i^{n+1/2}$$ (19)

for $H_i^{n+1/2} = \frac{\Delta x_i^{n+1/2}}{h_i^n} v_i^n - \frac{\Delta x_i^{n+1/2}}{h_i^{n+1}} v_i^n$. 


Replacing (19) in (16) we obtain
\[ v_{i}^{n+1} - v_{i}^{n} = H_{i-1/2}^{n} - H_{i+1/2}^{n} \]
\[ - \frac{\Delta t}{2\Delta x} \left( \left( B_{i+1/2}^{n} - Q_{i+1/2}^{n} \right) \Delta v_{i+1/2}^{n} + \left( B_{i-1/2}^{n} + Q_{i-1/2}^{n} \right) \Delta v_{i-1/2}^{n} \right) \]
which can be analogously written as a conservative update over the reference uniform mesh by replacing (19) in (15)
\[ v_{i}^{n+1} = v_{i}^{n} - \frac{\Delta t}{\Delta x} \left( H_{i+1/2}^{n} + \hat{H}_{i+1/2}^{n} - H_{i-1/2}^{n} - \hat{H}_{i-1/2}^{n} \right). \]  
(20)

We note that the conservative difference \(H_{i-1/2}^{n} - H_{i+1/2}^{n}\) accounts for the mesh reconstruction and the solution update step of the MAS.

In order to simplify the presentation of the rest of the proof we assume that the entropy and the conservative variables (\(\tilde{v}\) and \(v\), respectively) coincide, i.e., we choose \(U(u) = \frac{1}{2}u^2\) for the entropy function. To recover the entropy-entropy flux representation of (20), we multiply it by the entropy variables \(\tilde{v}_{i}^{n}\), yielding
\[ U(v_{i}^{n+1}) - U(v_{i}^{n}) + \frac{\Delta t}{\Delta x} \left( G_{i+1/2}^{n} - G_{i-1/2}^{n} \right) = M_{i}^{n} - \frac{\Delta t}{\Delta x} \mathcal{E}_{i}^{(x)} + \mathcal{E}_{i}^{(FE)}(\Delta v_{i+1/2}^{n}). \]  
(22)

where \(G\) is the numerical entropy flux. We have further following [18],
\[ \mathcal{E}_{i}^{(x)} = \frac{1}{4} \left( D_{i-1/2} \Delta v_{i-1/2}^{2} + D_{i+1/2} \Delta v_{i+1/2}^{2} \right) \]
\[ \mathcal{E}_{i}^{(FE)}(x) \leq \frac{K^3}{4} \left( \frac{\Delta t}{\Delta x} \right)^2 \left( (B_{i+1/2} - Q_{i+1/2})^2 \Delta v_{i+1/2}^{2} + (B_{i-1/2} + Q_{i-1/2})^2 \Delta v_{i-1/2}^{2} \right), \]
\[ = \frac{K^3}{4} \left( \frac{\Delta t}{\Delta x} \right)^2 \left( (B_{i+1/2} - Q_{i+1/2})^2 \Delta v_{i+1/2}^{2} + (B_{i-1/2} + Q_{i+1/2}^*)^2 \Delta v_{i-1/2}^{2} \right) \]
\[ + (B_{i-1/2} + Q_{i+1/2}^*)^2 \Delta v_{i+1/2}^{2} \]

Now, (22) reduces to
\[ U(v_{i}^{n+1}) - U(v_{i}^{n}) + \frac{\Delta t}{\Delta x} \left( G_{i+1/2}^{n} - G_{i-1/2}^{n} \right) \leq \]
\[ M_{i}^{n} - \frac{\Delta t}{4\Delta x} \left( D_{i-1/2} \Delta v_{i-1/2}^{2} + D_{i+1/2} \Delta v_{i+1/2}^{2} \right) \]
\[ + \frac{K^3}{4} \left( \frac{\Delta t}{\Delta x} \right)^2 \left( (B_{i+1/2} - Q_{i+1/2}^*)^2 \Delta v_{i+1/2}^{2} + (B_{i-1/2} + Q_{i+1/2}^*)^2 \Delta v_{i-1/2}^{2} \right), \]

Hence, the sufficiency condition for entropy stability reads
\[ M_{i}^{n} \leq \frac{\Delta t}{4\Delta x} \left( D_{i-1/2}^{n} - K^3 \frac{\Delta t}{\Delta x} (B_{i-1/2} + Q_{i-1/2})^2 \Delta v_{i-1/2}^{n} \right) \left( (\Delta v_{i-1/2}^{n})^2 \right) \]
\[ + \left( D_{i+1/2}^{n} - K^3 \frac{\Delta t}{\Delta x} (B_{i+1/2} - Q_{i+1/2}^*)^2 \Delta v_{i+1/2}^{n} \right) \left( (\Delta v_{i+1/2}^{n})^2 \right). \]  
\[ \square \]
Example 1. To gain further insight into (17), (19) we refer to Figure 1 and provide three special examples.

- Cell moves to the left.

This means that \( \Delta x_{i+1/2} = -r_i^n \), \( \Delta x_{i-1/2} = -r_{i-1}^n \), so \( h_i^{n+1} = h_i^n - r_i^n + r_{i-1}^n \). The mass of \( u \) satisfies \( h_i^{n+1} \dot{u}_i^n = r_{i-1}^n u_{i-1}^n + l_i^n u_i^n \); passing in \( v \) variables \( \dot{v}_i^n = \frac{r_{i-1}^n}{h_i^{n-1}} v_{i-1}^n + \frac{l_i^n}{h_i^n} v_i^n \) or
\[
\dot{v}_i^n - v_i^n = \frac{r_{i-1}^n}{h_i^{n-1}} v_{i-1}^n - \frac{r_i^n}{h_i^n} v_i^n. \tag{23}
\]

- Cell moves to the right.

This means that \( \Delta x_{i+1/2} = m_i^{n+1} \), \( \Delta x_{i-1/2} = l_i^n \), so \( h_i^{n+1} = h_i^n + m_i^{n+1} - l_i^n = r_i^n + l_{i+1}^n \). In this case, the mass of \( u \) satisfies \( h_i^{n+1} \dot{u}_i^n = r_i^n u_i^n + l_{i+1}^n u_{i+1}^n \), or in \( v \) variables \( \dot{v}_i^n = \frac{r_i^n}{h_i^n} v_i^n + \frac{l_{i+1}^n}{h_{i+1}^n} v_{i+1}^n \), or
\[
\dot{v}_i^n - v_i^n = -\frac{l_i^n}{h_i^n} v_i^n + \frac{l_{i+1}^n}{h_{i+1}^n} v_{i+1}^n. \tag{24}
\]
Cell moves to both directions. Similarly, $\Delta x_{i+1/2} = l_i^n + r_i^{-1}$, so $h_{i+1}^{n+1} = h_i^n + l_i^n + r_i^{-1}$.

Moreover $h_i^{n+1} \hat{u}_i^n = r_i^{-1} u_i^{n-1} + l_i^n u_i^n$, or $\hat{v}_i^n = r_i^{-1} v_i^{n-1} + v_i^n + \frac{l_i^n}{h_i^n} v_i^{n+1}$, or

$$\hat{v}_i^n - v_i^n = \frac{r_i^{-1} u_i^{n-1} - l_i^n u_i^n}{h_i^n} + \frac{l_i^n}{h_i^n} v_i^{n+1}. \quad (25)$$

Let us point that now the conditions (23)-(25) are simple enough in order to be tested for a moving mesh algorithm. We will report numerical experiments in a future work.

4. Conclusions. We provide in this work a new framework of studying the combined effect of mesh adaptation and time evolution of a numerical solution. This new method, has the benefit of being described over a uniform “underlying” grid that resolves the physical domain with the same number of discretization nodes as the numerical solution itself. To exhibit properties of this technique we study the dissipation of entropy due to the adaptation of the mesh. We have derived a sufficient condition for the mesh movement in order to guarantee that the overall procedure dissipates the entropy.

Acknowledgments. The authors wish to thank E. Tadmor and Ch. Makridakis for the very useful discussions and suggestions.

REFERENCES


[12] LeFloch, Ph. and Rohde, Ch., J. Numerical Analysis SIAM, High-order schemes, entropy inequalities and non-classical shocks, 2000


E-mail address: lukacova@uni-mainz.de
E-mail address: sfakiana@uni-mainz.de
Almost Global Existence of Classical Discontinuous Solutions to General Quasilinear Hyperbolic Systems of Conservation Laws with Small BV Initial Data

Zhi-Qiang Shao
Department of Mathematics
Fuzhou University, Fuzhou 350002, China

Abstract. In the present paper the author investigates the global structure stability of Riemann solutions for general quasilinear hyperbolic systems of conservation laws under small BV perturbations of the initial data, where the Riemann solution only contains shocks and contact discontinuities, the perturbations are in BV but they are assumed to be $C^1$-smooth, with bounded and possibly large $C^1$-norms. The author obtains the almost global existence and lifespan of classical discontinuous solutions to a class of the generalized Riemann problem, which can be regarded as a small BV perturbation of the corresponding Riemann problem. Some applications to quasilinear hyperbolic systems of conservation laws arising in physics, particularly to one-dimensional compressible Euler equations, are also given.

1. Introduction and main result. Consider the following quasilinear hyperbolic system of conservation laws:

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f(u) = 0, \quad x \in \mathbb{R}, \quad t > 0, (1.1)$$

where $u = (u_1, \ldots, u_n)^T$ is the unknown vector-valued function of $(t, x)$, $f : \mathbb{R}^n \to \mathbb{R}^n$ is a given $C^3$ vector function of $u$.

It is assumed that system (1.1) is strictly hyperbolic, i.e., for any given $u$ on the domain under consideration, the Jacobian $A(u) = \nabla f(u)$ has $n$ real distinct eigenvalues

$$\lambda_1(u) < \lambda_2(u) < \ldots < \lambda_n(u). (1.2)$$

Let $l_i(u) = (l_{i1}(u), \ldots, l_{in}(u))^T$ (resp. $r_i(u) = (r_{i1}(u), \ldots, r_{in}(u))^T$) be a left (resp. right) eigenvector corresponding to $\lambda_i(u) (i = 1, \ldots, n)$:

$$l_i(u)A(u) = \lambda_i(u)l_i(u) \quad (\text{resp. } A(u)r_i(u) = \lambda_i(u)r_i(u)). (1.3)$$

We have

$$\det|l_{ij}(u)| \neq 0 \quad \text{(equivalently, } \det|r_{ij}(u)| \neq 0). (1.4)$$

2000 Mathematics Subject Classification. Primary: 35L65, 35L45; Secondary: 35L67.

Key words and phrases. Generalized Riemann problem, quasilinear hyperbolic system of conservation laws, piecewise $C^1$ solution, shock wave, contact discontinuity, lifespan, almost global existence.

This work is supported by the Scientific Research Foundation of the Ministry of Education of China (Grant No. 02JAZ790014), the Natural Science Foundation of Fujian Province of China (Grant No. 2012J01006) and the Science and Technology Developmental Foundation of Fuzhou University (Grant No. 2004-XQ-16).
Without loss of generality, we may assume that on the domain under consideration
\[ l_i(u)l_j(u) \equiv \delta_{ij} \quad (i, j = 1, \ldots, n) \] (1.5)
and
\[ r_i^T(u)r_i(u) \equiv 1 \quad (i = 1, \ldots, n), \] (1.6)
where \( \delta_{ij} \) stands for the Kronecker’s symbol.

Clearly, all \( \lambda_i(u) \), \( l_{ij}(u) \) and \( r_{ij}(u) \) (\( i, j = 1, \ldots, n \)) have the same regularity as \( A(u) \), i.e., \( C^2 \) regularity.

We also assume that on the domain under consideration, each characteristic field is either genuinely nonlinear in the sense of Lax (cf. [1]):
\[ \nabla \lambda_i(u)r_i(u) \neq 0 \] (1.7)
or linearly degenerate in the sense of Lax:
\[ \nabla \lambda_i(u)r_i(u) \equiv 0. \] (1.8)

We are interested in the generalized Riemann problem for system (1.1), which is a Cauchy problem with a piecewise \( C^1 \) initial data of the form:
\[ t = 0 : u = \begin{cases} u_0^-(x), & x \leq 0, \\ u_0^+(x), & x \geq 0, \end{cases} \] (1.9)
where \( u_0^-(x) \) and \( u_0^+(x) \) are \( C^1 \) vector functions defined for \( x \leq 0 \) and \( x \geq 0 \) respectively with
\[ u_0^-(0) \neq u_0^+(0). \] (1.10)

Problem (1.1) and (1.9) may be regarded as a perturbation of the corresponding Riemann problem (1.1) and
\[ t = 0 : u = \begin{cases} \hat{u}_-, & x \leq 0, \\ \hat{u}_+, & x \geq 0, \end{cases} \] (1.11)
in which
\[ \hat{u}_\pm = u_0^\pm(0). \] (1.12)

Let
\[ \theta = |\hat{u}_- - \hat{u}_+|. \] (1.13)
When \( \theta > 0 \) is suitably small, by Lax [1], the Riemann problem (1.1) and (1.11) admits a unique self-similar solution composed of \( n + 1 \) constant states \( \bar{u}^{(0)} = \bar{u}_-, \bar{u}^{(1)}, \ldots, \bar{u}^{(n-1)}, \bar{u}^{(n)} = \bar{u}_+ \) separated by shocks, centered rarefaction waves (corresponding characteristics are genuinely nonlinear) or contact discontinuities (corresponding characteristics are linearly degenerate). As in [2], this kind of solution is simply called the Lax’s Riemann solution of the system (1.1).

For the self-similar solution of the Riemann problem of general quasilinear hyperbolic systems of conservation laws, the local nonlinear structure stability has been proved by Li and Yu [3] for one-dimensional case, and by Majda [4] for multidimensional case. If system (1.1) is strictly hyperbolic and genuinely nonlinear, Li and Zhao [5] proved the global structure stability of the self-similar solution containing only \( n \) shocks under perturbation (1.9) satisfying (1.12). In their work they do not require the amplitude of the self-similar solution is small, although the existence of the self-similar solution with non-small amplitude still remains open. If system (1.1) is strictly hyperbolic and linearly degenerate, Li and Kong [6] proved the global structure stability of the self-similar solution with small amplitude under perturbation (1.9) satisfying (1.12). In this case the self-similar solution contains only \( n \) contact discontinuities. Since both genuinely nonlinear system and
linearly degenerate system are only two extreme cases, many physical systems (for
example, the system of one-dimensional gas dynamics, the system of traffic flow
on a road network using the Aw-Rascle model, etc.) do not belong to these two
cases. Therefore, a general consideration is needed for general hyperbolic systems
of conservation laws. Recently, Kong [2] investigated this kind of systems whose
characteristic families might be either genuinely nonlinear or linearly degenerate.
Precisely speaking, under certain reasonable hypotheses Li and Zhao [5] obtained
the following well-known result.

**Theorem 1.1.** Suppose that system (1.1) is strictly hyperbolic and genuinely non-
linear. Suppose furthermore that \( u_0(x) \) and \( u_0^+(x) \) are all \( C^1 \) vector functions on
\( x \leq 0 \) and on \( x \geq 0 \) respectively, \( f(u) \) is a \( C^2 \) vector function and
\[
\theta \triangleq |\hat{u}_+ - \hat{u}_-| = |u_0^+(0) - u_0^-(0)| > 0
\]
is suitably small. Suppose finally that the self-similar solution \( u = U(x) \) of the
Riemann problem (1.1) and (1.11) is composed of \( n + 1 \) constant states \( \hat{u}^{(0)} = \hat{u}_-, \hat{u}^{(1)}, \ldots, \hat{u}^{(n)}, \hat{u}^{(n)} = \hat{u}_+ \) and \( n \) non-degenerate typical shocks \( x = \hat{F}^i t \) \( (i = 1, \ldots, n) \):
\[
u = U\left(\frac{x}{t}\right) = \begin{cases}
\hat{u}^{(0)}, & x \leq \hat{F}^1 t, \\
\hat{u}^{(i)}, & \hat{F}^i t \leq x \leq \hat{F}^{i+1} t \ (i = 1, \ldots, n - 1), \\
\hat{u}^{(n)}, & x \geq \hat{F}^n t.
\end{cases}
\]
Then there exists a positive constant \( \varepsilon \) so small that if
\[
|u_0^-(x) - u_0^-(0)|, \quad |u_0^-(x)| \leq \frac{\varepsilon}{1 + |x|}, \quad \forall x \leq 0,
\]
\[
|u_0^+(x) - u_0^+(0)|, \quad |u_0^+(x)| \leq \frac{\varepsilon}{1 + |x|}, \quad \forall x \geq 0,
\]
then problem (1.1) and (1.9) admits a unique global classical discontinuous solution
\( u = u(t, x) \) only containing \( n \) shocks \( x = x_i(t) (x_i(0) = 0) \) \( (i = 1, \ldots, n) \), such that
\( u(t, x) \) belongs to \( C^1 \) on each domain \( D^i \) \( (i = 0, 1, \ldots, n) \) and \( x_i(t) \) \( (i = 1, \ldots, n) \) to
\( C^2 \) on \( t \geq 0 \) with
\[
|u(t, x) - u(0, 0)| \leq \frac{K\varepsilon}{1 + t}, \quad \forall (t, x) \in D^i \ (i = 0, 1, \ldots, n),
\]
\[
\left|\frac{\partial u}{\partial x}(t, x)\right|, \left|\frac{\partial u}{\partial t}(t, x)\right| \leq \frac{K\varepsilon}{1 + t}, \quad \forall (t, x) \in D^i \ (i = 0, 1, \ldots, n),
\]
\[
|x_i'(t) - x_i'(0)| \leq \frac{K\varepsilon}{1 + t}, \quad \forall t \geq 0 \ (i = 1, \ldots, n),
\]
where
\[
D^0 = \{(t, x)|t \geq 0, x \leq x_1(t)\},
\]
\[
D^i = \{(t, x)|t \geq 0, x_i(t) \leq x \leq x_{i+1}(t)\} \ (i = 1, \ldots, n - 1),
\]
\[
D^n = \{(t, x)|t \geq 0, x \geq x_n(t)\}
\]
and \( K \) is a positive constant independent of \( t \). Moreover, \( u(0, 0) = \hat{u}^{(i)} \) on the
domain \( D^i \) \( (i = 0, 1, \ldots, n) \) and \( x_i'(0) = \hat{F}^i \) \( (i = 1, \ldots, n) \). Therefore, as a global
perturbation, \( u(t, x) \) possesses a similar structure to that of the self-similar solution
to Riemann problem (1.1) and (1.11) on \( t \geq 0 \).
Remark 1. Recently, under certain reasonable hypotheses Kong [2, 7] proved that the Lax’s Riemann solution of general \( n \times n \) quasilinear hyperbolic system of conservation laws is globally structurally stable if and only if it contains only non-degenerate shocks and contact discontinuities, but no rarefaction waves and other weak discontinuities. Shao [8, 9] also studied that the global structure stability and instability of this kind of Lax’s Riemann solution with small amplitude in a half space.

However, it is well known that the BV space is a suitable framework for one-dimensional Cauchy problem for the hyperbolic systems of conservation laws (see Bressan [10], Glimm [11]), the result in Bressan [12] suggests that one may achieve global smoothness even if the \( C^1 \) norm of the initial data is large. So the following question arises naturally: can we obtain the global existence and uniqueness of piecewise \( C^1 \) solution containing only shocks and contact discontinuities to a class of the generalized Riemann problem, which can be regarded as a small BV perturbation of the corresponding Riemann problem, for system (1.1) with the following piecewise \( C^1 \) initial data:

\[
\begin{align*}
  t = 0 : u &= \begin{cases} 
    \hat{u}_- + u_-(x), & x \leq 0, \\
    \hat{u}_+ + u_+(x), & x \geq 0,
  \end{cases}
\end{align*}
\]  

(1.14)

where \( u_\pm(x) \in C^1 \) with bounded and possibly large \( C^1 \) norm, but of small bounded variation, such that

\[
  ||u_-(x)||_{C^1}, \ ||u_+(x)||_{C^1} \leq M,
\]

(1.15)

for some \( M > 0 \) bounded but possibly large, and also such that

\[
  \int_{-\infty}^{\infty} |u'_+(x)|dx, \int_{0}^{\infty} |u'_-(x)|dx \leq \varepsilon,
\]

(1.16)

for some \( \varepsilon > 0 \) sufficiently small? Here, it is important to mention that the global existence of weak solutions to a strictly hyperbolic system of conservation laws in one space dimension when the initial data is a small BV perturbation of a solvable Riemann problem has been proved by Schochet [13], unfortunately his method is not useful to show that the solutions are still either contact discontinuities or shocks. An analogous result on stability of a strong shock wave under perturbations of small bounded variation is stated by Corli and Sable-Tougeron [14]. In this paper we exploit to some extent the ideas of Bressan [12], we will develop the method of using continuous Glimm’s functional to provide a new, concise proof of an estimate on the lifespan of the piecewise \( C^1 \) solution to the generalized Riemann problem under consideration mentioned above. The basic idea we will use here is to combine the techniques employed by Li-Kong [5], especially both the decomposition of waves and the global behavior of waves on the discontinuity curves, with the method of using continuous Glimm’s functional. However, we must modify Glimm’s functional in order to take care of the presence of shock waves. This makes our new analysis more complicated than those for the \( C^1 \) solutions of the Cauchy problem for linearly degenerate quasilinear hyperbolic systems in Bressan [12], Zhou [15], Dai and Kong [16].

As in [17], the aim of this paper is to study the global structure stability of Lax’s Riemann solution containing only shocks and contact discontinuities (particularly shocks are present). In this case, we shall first get a lower bound of the lifespan of the piecewise \( C^1 \) solution to the generalized Riemann problem.
To do so, we consider the generalized Riemann problem for the system (1.1) with the following piecewise $C^1$ initial data:

$$ t = 0 : u = \begin{cases} \widehat{u}_- + \varepsilon u_-(x), & x \leq 0, \\ \widehat{u}_+ + \varepsilon u_+(x), & x \geq 0, \end{cases} $$  \hspace{1cm} (1.17)

where $\varepsilon$ ($0 < \varepsilon \ll |\widehat{u}_+ - \widehat{u}_-|$) is a small parameter, $u_-(x)$ and $u_+(x)$ are $C^1$ vector functions defined on $x \leq 0$ and $x \geq 0$ respectively, which satisfy

$$ u_-(0) = u_+(0) = 0, $$

$$ ||u_-(x)||_{C^1}, ||u_+(x)||_{C^1} \leq K_1 $$

and

$$ \int_0^{+\infty} |u_+(x)| dx, \int_{-\infty}^0 |u_-(x)| dx \leq K_2, $$

where $K_1$ and $K_2$ are positive constants independent of $\varepsilon$.

Introduce

$$ J_S \triangleq \{ j \mid j \in \{1, \ldots, n\}, \text{ } j\text{-wave in } u = U\left(\frac{x}{t}\right) \text{ is a shock wave}\}, $$

$$ J \triangleq \{ j \mid j \in \{1, \ldots, n\}, \lambda_j(u) \text{ is genuinely nonlinear}\} $$

and

$$ I \triangleq \{ i \mid i \in \{1, \ldots, n\}, \lambda_i(u) \text{ is linearly degenerate}\}. $$

Then, the assumption that each characteristic field is either genuinely nonlinear or linearly degenerate gives

$$ I \cup J = \{1, \ldots, n\}. $$

To state our result precisely, we now introduce the concept of the lifespan of the piecewise $C^1$ solution to the generalized Riemann problem (1.1) and (1.17) as follows.

**Definition 1.2.** The existence of piecewise $C^1$ local solutions to the generalized Riemann problem (1.1) and (1.17) is guaranteed by the monograph Li-Yu [3]. The life span is defined to be the supremum of the time $T$ such that a Li-Yu solution exists for $0 < t \leq T$. This definition will be coincide with the usual definition of the life span of $C^1$ solution (without shocks).

Our main results can be summarized as follows.

**Theorem 1.3.** Suppose that system (1.1) is strictly hyperbolic and each characteristic field is either genuinely nonlinear or linearly degenerate. Suppose furthermore that $u_-(x)$ and $u_+(x)$ are all $C^1$ vector functions on $x \leq 0$ and on $x \geq 0$ respectively satisfying (1.18)-(1.20) and

$$ \theta = |\widehat{u}_+ - \widehat{u}_-| = |u_0^+(0) - u_0^-(0)| > 0 $$

is suitably small. Suppose finally that the self-similar solution $u = U(\frac{x}{t})$ of the Riemann problem (1.1) and (1.11) consists of $k$ shock waves and $n - k$ contact discontinuities for some integer $k$ ($1 \leq k \leq n$). Then for small $\theta > 0$, there exists a constant $\varepsilon_0 > 0$ so small that for any fixed $\varepsilon \in (0, \varepsilon_0]$, the lifespan $T(\varepsilon)$ of the piecewise $C^1$ solution to the generalized Riemann problem (1.1) and (1.17) satisfies

$$ T(\varepsilon) \geq K_3 \varepsilon^{-1}, $$

where $K_3$ is a positive constant independent of $\varepsilon$. Moreover, when $u = u(t, x)$ blows up in a finite time, $u = u(t, x)$ itself is bounded on the domain $[0, T(\varepsilon)] \times \mathbb{R}$, while the first-order derivatives of $u = u(t, x)$ tend to be unbounded as $t \searrow T(\varepsilon)$. 
Remark 2. Our result implies that classical discontinuous solutions to the generalized Riemann problem under consideration exists almost globally in time. We refer to Kong [18] for the definition of an almost global solution.

Remark 3. Suppose that (1.1) is a non-strictly hyperbolic system with characteristics with constant multiplicity, say, on the domain under consideration,
\[
\lambda_1(u) \equiv \cdots \equiv \lambda_p(u) < \lambda_{p+1}(u) < \cdots < \lambda_n(u) \quad (1 \leq p \leq n).
\]
Then the conclusion of Theorem 1.3 still holds (cf. [16]).

Remark 4. The relation to the Li-Kong paper is as follows: the Li-Kong initial data will be shown to satisfy (1.17)-(1.20). Because, the Li-Kong solution is a global solution and naturally satisfies the conditions for almost global solutions.

Some of the results related to these topics are listed below. Chen et al. [19-22] investigated the asymptotic stability of Riemann waves for hyperbolic conservation laws. Hsiao and Tang [23] investigated the construction and qualitative behavior of the solution of the perturbed Riemann problem for the system of one-dimensional isentropic flow with damping. Xin et al. [24, 25] proved the nonlinear stability of contact discontinuities in systems of conservation laws. Smoller et al. [26] investigated the instability of rarefaction shocks in systems of conservation laws. For the overcompressive shock waves, Liu [27] proved the nonlinear stability and instability. Bressan and LeFloch [28] investigated the structural stability and regularity of entropy solutions to hyperbolic systems of conservation laws. Lions et al. [29] proved the existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates. Recently, \(L^1\) stability for hyperbolic systems of conservation laws was proved by Bressan, Liu and Yang [30], within the class of solutions with small total variation (see also [10, 31-33]). Their results were extended in Lewicka [34] to the case where the initial data is a BV perturbation of a possibly large Riemann data. Liu and Xin [35] proved the nonlinear stability of discrete shocks for systems of conservation laws. Dafermos [36] studied the entropy and the stability of classical solutions of hyperbolic systems of conservation laws. For a relaxation system in several space dimensions, Luo and Xin [37] proved the nonlinear stability of shock fronts. Liu and Xin [38] investigated the nonlinear stability of rarefaction waves for compressible Navier-Stokes equations. Hsiao and Pan [39] investigated the nonlinear stability of rarefaction waves for a rate-type viscoelastic system. Moreover, the nonlinear stability of an undercompressive shock for complex Burgers equation was studied by Liu and Zumbrun [40]. For the viscous conservation laws, the theory of nonlinear stability of shock waves was established (see [41, 42] and the references therein).

2. Applications. In this section, two concrete examples are given to show some applications of our main results, Theorem 1.3. They are both the 1-dimensional compressible Euler equations of gas dynamics (cf. Chen and Frid [20]) and the system of traffic flow on a road network using the Aw-Rascle model (cf. [43, 44]), for which our work is of importance. It is well-known that these systems are strictly hyperbolic and that their fields are either genuinely nonlinear or linearly degenerate, so the theorems are obviously applicable. Here, we skip the details.

REFERENCES


E-mail address: zqshao@fzu.edu.cn
Abstract. We consider a nonlinear system of conservation laws arising in petroleum engineering. We are modeling the injection of a mixture of gas and oil, in any proportion, into a porous medium filled with a similar mixture. The two mixtures may have different temperatures. We will focus on a particularly unusual feature found in this model: for open set of Riemann data the solution is given by a single wave group, i.e., there is no constant intermediate state. The key aspect supporting this feature is the existence of structurally stable doubly sonic shock waves, which robustly connect slow rarefaction waves to fast rarefaction waves. The solutions are constructed around a curve of coinciding characteristic speeds, intrinsically associated to most bifurcations in the Riemann solutions for this class of models.

1. Introduction. We consider systems of conservation laws in the form:

$$\varphi \partial_t G(w) + \partial_x u F(w) = 0,$$

which models compositional thermal two-phase flow in a porous medium. This class of models is useful in applications such as advanced recovery of gas or oil, see [2], and clean-up of polluted sites, see [8]. Here, the reduced state space consists of oil saturation ($s$) and temperature ($T$) above a reference temperature ($T_{ref}$), which we write as:

$$w \in \Omega = \{(s,T) \mid 0 \leq s \leq 1, T > T_{ref}\}.$$  

The variable $u$ denotes the Darcy speed and $\varphi$ is the constant porosity of the medium. To fully describe a state in system (1) one needs to specify $(w, u) \in \Omega \times \mathbb{R}^+$. However, the variable $u$ can be recovered from $w$, as done in [6]; for that reason we study the wave structure in $\Omega$, which we will call state space. The vector valued accumulation function $G = (G_1, G_2, G_3)^T$ and flux function $F = (F_1, F_2, F_3)^T$ are smooth on the entire state space.

A remarkable property satisfied by models in this class is that they can be written in the following form (after rescaling time and dropping $\varphi$):

$$\partial_t \left( \alpha(T)s + \beta(T)(1-s) + \gamma(T) \right) + \partial_x u \left( \alpha(T)f(s,T) + \beta(T)(1-f(s,T)) \right) = 0.$$
Regarding the thermal branch of the Hugoniot locus (motivated by Proposition 2) line the saturation branch of the Hugoniot locus, and denote this set by $H^\xi$ be such a smooth self-similar solution conveniently parametrized by $\xi$ where $D$ denotes differentiation relative to $\xi$; see more details in [7]. Let $(w(\xi), u(\xi))$ be such a smooth self-similar solution conveniently parametrized by $\xi = x/t$ and $r_i$ a right eigenvector defined by (2). Together, they must satisfy the ODE in $\xi$:

$$(\dot{w}(\xi), \dot{u}(\xi)) = r_i(w(\xi), u(\xi)),$$

for a suitable initial datum. The family $i$ is one of the two possible choices:

**Proposition 1.** There are two eigenpairs associated to the generalized eigenvalue problem (2), namely:

(s) $\lambda_s(w, u) = u\partial_s f(s, T)$ and $r_s = (1, 0, 0)^T$;

(e) $\lambda_e(w, u) = \frac{a(T)f(s, T) + b_1(T)}{a(T)s + b_2(T)}$ and $r_e = (\hat{r}_e(s, T), u\lambda'_3(s, T))^T$,

where $w = (s, T)$ and $\hat{r}_e(w) = (X_1(w), X_2(w))^T$ is the restriction of $r_e$ to $\Omega$.

**Remark 1.** The quantities $a(T), b_1(T)$ and $b_2(T)$ are positive, as in the examples [6], [13]; see there expressions for $r_e$ and these quantities in terms of $\alpha, \beta$ and $\gamma$.

The family (s) produces pure transport of fluid without changes in temperature and is henceforth called the saturation transport family. Family (e) is associated to thermal transport and typically produces waves associated to evaporation/condensation phenomena; it will be called the thermal transport family.

A bounded discontinuous self-similar solution of (1) with a jump connecting the pairs $(w^-, u^-), (w^+, u^+)$ must satisfy the Rankine-Hugoniot relation:

$$\mathcal{H}(w^-, u^-, w^+, u^+; \sigma) \equiv u^+ F(w^+) - u^- F(w^-) - \sigma \left(G(w^+) - G(w^-)\right) = 0,$$  

(4)

where the wave speed is denoted by $\sigma$. For a fixed initial state $(w^-, u^-)$ a typical issue is the locus of states $(w^+, u^+)$ that satisfy (4). Here, one first calculates the set of states $w^+$ in which the determinant of the matrix $(F(w^+), F(w^-), G(w^+) - G(w^-))$ is zero; the resulting set is called the Hugoniot locus of the base state $w^-$, which will be denoted as $\mathcal{H}(w^-) \subset \Omega$. Afterwards, the Darcy speed $u^+$ can be obtained for any triplet $(w^-, u^-, w^+), w^+ \in \mathcal{H}(w^-)$ by using (4). Here, we call a jump admissible if it satisfies the Liu criterion [10]; admissible jumps are called shock waves. Regarding the shape of the Hugoniot locus, we have [7]:

**Proposition 2.** For any pair $(w^-, u^-) \in \Omega \times \mathbb{R}^+$, the locus $\mathcal{H}(w^-)$ always contains a straight line in state space. On this line, the following equalities hold:

$$T^+ = T^-, \quad u^+ = u^- \quad \text{and} \quad \sigma = u^+ \frac{f^+ - f^-}{s^+ - s^-}.$$  

(5)

Shocks connecting the base state $w^-$ to points $w^+$ in this line satisfy the Lax inequalities for the $s$-family: $\lambda_s(w^-, u^-) \geq \sigma \geq \lambda_s(w^+, u^+)$, where $u^+ = u^-, u^-$ is given and only one inequality is allowed to become an equality. We therefore call this line the saturation branch of the Hugoniot locus, and denote this set by $\mathcal{H}_s(w^-)$. Regarding the thermal branch of the Hugoniot locus (motivated by Proposition 2)
we may ask if it makes sense to take the limit in the expression of the Rankine-Hugoniot relation (4) divided by $T^+ - T^-$, when $T^+$ tends to $T^-$, i.e., if the following function is well defined:

$$h_e(w^+; w^-) = \begin{cases} 
\frac{\det \left( F(w^+); F(w^-); G(w^+) - G(w^-) \right)}{T^+ - T^-}, & T^+ \neq T^-; \\
\lim_{T^+ \to T^-} \frac{\det \left( F(w^+); F(w^-); G(w^+) - G(w^-) \right)}{T^+ - T^-}, & \text{otherwise.}
\end{cases}$$

(6)

One can prove that the function $h_e$ defined in (6) is smooth. Let $\mathcal{H}_e(w^-)$ consist of points $w^+$ in the zero set of $h_e(\cdot; w^-)$. We have:

**Proposition 3.** For any base state $w^-$ the Hugoniot locus can be decomposed as:

$$\mathcal{H}(w^-) = \mathcal{H}_s(w^-) \bigcup \mathcal{H}_e(w^-).$$

(7)

Of course, the set $\mathcal{H}_e(w^-)$ is the thermal branch of the Hugoniot locus; its topology will be discussed in the next section.

3. Elementary waves and curves. Since the seminal works [9] and [4], strict hyperbolicity and genuine nonlinearity were mainly recognized as two key assumptions for the well-posedness of the Cauchy problem for systems of conservation laws. In the class of systems we consider both are violated; yet, as far as the Riemann problem is concerned, examples of existence and continuous dependence of the solution with respect to the initial data are known, see [6] and [12]. For our models, strict hyperbolicity is lost when:

**Proposition 4.** The two characteristic speeds coincide along a smooth curve, transverse to the constant field $\hat{r}_s = (1, 0)^T$, which disconnects state space. We denote this coincidence curve by $C$.

Away from the coincidence curve we may solve unambiguously the ODE stated in (3), for each characteristic family. However, these orbits can (and typically will) cross the coincidence curve; this is the case we will tackle now. In what follows, we will only focus on the flow induced by the eigenvectors of the thermal family; the other one lies along straight lines. A fundamental observation is that:

**Proposition 5.** Along the coincidence curve the kernel of the matrix $J$, defined in (2b), has dimension 1, generically.

This and the obvious observation that the dimension of the kernel of $J$ can only be one or two, which can be seen directly from (2b), lead to the following:

**Definition 3.1.** A point on the coincidence curve is called singular if the dimension of the kernel of $J$ equals 2. We denote the set of singular points by $S$.

**Remark 2.** Notice that, by Propositions 4 and 5, the singular points form a discrete subset of the coincidence curve. Researchers interested in models for three-phase flow in porous media should contrast this behavior with those in [5], [1].

Now we are in position to state:

**Proposition 6.** The thermal family eigenvector-field can be represented by a smooth, non-vanishing vector field in $\Omega \setminus S$. 

Proposition 6 allows us to construct smooth thermal self-similar solutions of (1), provided that genuine nonlinearity holds almost everywhere for this family and the set where it fails is transverse to the thermal vector field. This is the case as will be shown now. We use the notation $I_e$ for the set of points $(w, u)$ on which genuine nonlinearity is lost in the thermal family, i.e., the equality $\nabla \lambda_e(w, u) \cdot r_e(w, u) = 0$ is satisfied. The set $I_e$ is called the inflection locus (of the thermal family); it possesses the following properties:

**Proposition 7.** In state space, the thermal inflection locus is a smooth curve such that $I_e \supset C$. However, in a neighborhood of a singular point the inflection and the coincidence curves are the same, i.e., every singular point possesses a neighborhood $V$ in which $I_e \cap V = C \cap V$.

Together with Propositions 4 and 5, Proposition 7 gives the properties we need for constructing thermal rarefaction waves in a neighborhood of a singular point. Actually, this construction can be extended to the whole state space.

Of course, singular points also have great influence on the thermal branches of the Hugoniot locus. The reader may already expect that:

**Proposition 8.** If $w^*$ is a singular point then $\partial_{w^+} h_e(w^+; w^*) \bigg|_{w^+ = w^*} = 0$.

The criticality stated in Proposition 8 allows two generic situations: in a neighborhood of a singular point the $e$-branches of the Hugoniot locus are diffeomorphic either to circles or to hyperbolae. To construct the Riemann solution we propose, the singular points are required to satisfy the following non-degeneracy conditions; both can be verified in physical models such as [6] and [12].

**Assumption 3.2.** If $w^* \in S$ is a singular point then:

1. $H_e(w^*) = \{w^*\}$. There is a neighborhood $V$ of the singular point such that the curve $H_e(w^-)$ is diffeomorphisic to a circle for any state $w^- \in V \setminus S$. Moreover, for any $w^- \in V \setminus S$ the locus $H_e(w^-)$ is tangent to the $e$-family integral curves only at $w^-$. 

2. The formula for $\hat{r}_e$ can be chosen so that $\hat{r}_e(w^*) = 0$, preserving the smoothness of the thermal eigenvector field. The eigenvalues of $D\hat{r}_e(w^*)$ are a pair of complex conjugate numbers with non-zero real part.

From now on we use bold capital roman letters to denote points of type $P = (w^P, u^P) \in \Omega \times \mathbb{R}^+$. A distinctive property of this class of systems is that, to obtain the solution of the Riemann problem, one just needs to describe the self-similar solutions in $\Omega$; the variable $u$ can be obtained from its boundary value in the PDE problem and from $w$, see [6]. Thus we will represent the elementary wave-curves in $\Omega$; we often use the notation $P$ instead of $w^P \in \Omega$. We illustrate rarefaction curves and shock branches in Figure 1.

A typical Riemann solution can be decomposed into a sequence of constant states and wave-groups: groups of elementary waves (shocks or rarefactions) that move together as a single entity. We introduce the notation $P_1 \xrightarrow{w} P_2$ to express the fact that the states $P_1, P_2$ are connected by an elementary wave $w$. Here, elementary shocks and rarefactions are saturation or thermal transport waves; they may be slow or fast depending on the ordering between the characteristic speeds, which varies: see Proposition 4. Deciding if a wave is a $s$-wave or an $e$-wave for rarefactions is a
Figure 1. Left: rarefaction curves; slow waves are dashed-dotted, fast waves are solid. Right: shocks, the same graphical convention. We use the superscripts $-, +$ for slow, fast waves and the subscripts $s, e$ for saturation, thermal waves.

matter of verifying which is the associated eigenvalue; for shocks one must verify in which branch of the Hugoniot-locus lies the pair of left, right states.

To construct the solution of the Riemann problem, we will use a generalization of Liu’s wave-curve method; see [10], [3], [11] and [12]. We define $\mathcal{W}^-(L)$, the slow wave curve emanating from state $L$, as the set in state space that can be reached through a speed-increasing succession of admissible elementary slow waves beginning at $L$. We define $\mathcal{W}^+(R)$, the fast wave curve reaching state $R$, as the set in state space from which one can reach the state $R$ through a speed-increasing succession of admissible elementary fast waves. Notice the asymmetry in these definitions. In the classical theory for $2 \times 2$ Liu systems, the Riemann solution between $L$ and $R$ (if it exists) would be obtained by finding the intersection of $\mathcal{W}^-(L)$ with $\mathcal{W}^+(R)$: the solution consists of elementary waves from the slow wave-curve, succeeded by the constant state $M \in \mathcal{W}^-(L) \cap \mathcal{W}^+(R)$ and, finally, elementary waves from the fast wave-curve. In the specific case we present here there would be no such intersection: a doubly characteristic shock (to be defined in the next paragraph) must be used as a bridge between the two wave-curves.

The construction of the Riemann solution relies on the fact that for any state $w^-$ in a neighbourhood of the singular point and any $u^- > 0$ there is a companion $w^+ \in \mathcal{H}_s(w^-) \cap \mathcal{H}_e(w^-)$ such that the shock between $w^-$ and $w^+$ is doubly characteristic with respect to the thermal speed: $\lambda_e(w^-, u^-) = \sigma = \lambda_e(w^+, u^+)$, with $u^+ = u^-$ by Proposition 2 and $\sigma$ is the shock speed in (4). Thus we introduce the following notation: for a set $\gamma$ (typically, a curve) in state space, the mixed extension set is

$$\mathcal{E}(\gamma) = \{ w^+ \in \Omega \mid \exists w^- \in \gamma; w^+ \in \mathcal{H}_s(w^-) \text{ and } \sigma = \lambda_e(w^-, u^- = 1) \}. \quad (8)$$

One can see that the extension set is independent of the choice $u^- = 1$. Of course, $w^-$ and $w^+$ have the same temperature in (8). We will also need the following:

Proposition 9. Let $P^-- (w^-, u^-) \neq P^+ (w^+, u^+)$ be at the same temperature; $P^-$, $P^+$ and $\sigma$ satisfying (4). Then $\sigma = \lambda_e(P^+)$ if and only if $\sigma = \lambda_e(P^-)$.
4. The Riemann solution. In this section we provide a sketch for the proof of:

**Theorem 4.1.** In a neighborhood \( W \) of a singular point \( w^* \) satisfying Assumption 3.2, there are open sets \( U, V \subset W \) such that the Riemann solution for any pair \( L = (w^L, u^L), R = (w^R, -) \) with \( (w^L, w^R) \in U \times V \) possesses no intermediate constant state.

A few remarks are in order: first, we cannot assign both \( u^L \) and \( u^R \) as Riemann data for system (1) in general, see [7]; this is not a property particular to the Riemann solutions we are showing. Second, this is a part of the full Riemann solution in the neighborhood of a singular point, see [12]: the complete solution satisfies Liu’s criterion [10], is \( L^1_{loc} \) continuous with respect to the Riemann data and is structurally stable in the sense of [11]. Albeit superficially similar to the doubly sonic (characteristic) transitional shock waves predicted in [11], in our case, the doubly characteristic shock waves vary when we allow the Riemann data to change.

**Proof.** By Proposition 7 we can restrict ourselves to a neighborhood of the singular point \( w^* \), such that \( \mathcal{I}_s = C \). One can further verify that this neighborhood can be chosen so that \( s \mapsto \lambda_s(s, T) \) is monotone (increasing, for definiteness), i.e., the \( s \)-family is genuinely nonlinear. Moreover, on isotherms, the saturation \( s \) increases (decreases) along \( s \)-rarefactions (\( s \)-shocks), see Propositions 1 and 2. Both eigenvalues are positive, recall Remark 1, and on the left (right) side of \( C \) we have \( \lambda_s < \lambda_e \) (\( \lambda_s > \lambda_e \)). We choose a \( R \) state on the right side of the coincidence curve \( C \), above the singular point; the set of \( L \) states will be specified soon.

The fast wave curve reaching \( R \) from below, \( \mathcal{W}^+(R) \), contains a fast thermal rarefaction curve, \( R^+_s \), on the left side of the coincidence curve \( C \) followed by a fast saturation rarefaction curve, \( R^+_e \), on the right side of \( C \) (see Figure 2a). These rarefaction curves are tangent at their intersection \( O \), which lies on \( C \), see Proposition 5; thus these rarefactions form a wave-group.

**Figure 2.** Left: \( \mathcal{W}^+(R) \), the fast wave curve reaching \( R \) and \( \mathcal{E}(R^+_e) \), the (doubly) characteristic extension of \( R^+_e \). Right: \( R^-_e \), slow rarefaction curve emanating from \( L \) and \( S^d_s \), saturation doubly characteristic shock between the slow and fast wave-curves.
Now we focus on the boundaries of the set from which we will select \( L \) states. First: the integral curve containing \( R_+^e \) also contains a slow thermal rarefaction curve beginning at a point \( Q \in \mathcal{C} \) and reaching \( O \), which we will denote as \( \gamma_{QO} \) and, of course, lies entirely on the right side of \( \mathcal{C} \), see the dashed curve on Figure 2a. Second: from (8), \( \mathcal{E}(R_+^e) \) is the set from which emanate characteristic s-shocks reaching \( R_+^e \). We claim that \( \mathcal{E}(R_+^e) \) is a smooth curve (see next section), which is drawn as the bold curve on Figure 2a. Moreover, there is an open set \( Z \in \Omega \) such that \( Z \cap \mathcal{E}(R_+^e) \) is transverse to the \( e \)-family vector field. By construction, the set \( Z \) must be on the right side of \( \mathcal{E}(R_+^e) \); one may see that it is on the left side of \( \gamma_{QO} \).

Finally, we choose a \( L \in Z \) such that there is a slow \( e \)-rarefaction \( R_+^e \) emanating from it and reaching \( Z \cap \mathcal{E}(R_+^e) \) transversally at a point \( M_1 \), see Figure 2b. From (8), there is a state \( M_2 \) in \( R_+^e \) such that the shock speed \( \sigma \) between \( M_1 \) and \( M_2 \) satisfies \( \sigma = \lambda_e(M_2) \); by Proposition 9, it also satisfies:

\[
\lambda_e(M_1) = \sigma = \lambda_e(M_2).
\]

Under these conditions there is a Riemann solution for the data \( L, R \), built so that the speed admissibility condition is satisfied. The construction is illustrated in Figure 2b, the profiles and the \( x/t \) diagram are depicted in Figure 3; the Riemann solution is:

\[
L \xrightarrow{R_+^e} M_1 \xrightarrow{S_{d}^e} M_2 \xrightarrow{R_+^e} O \xrightarrow{R_+^e} R.
\]

The construction outlined above is fairly generic; in the next section we will address why the extension (8) maps integral curves on the left side of \( \mathcal{C} \) to smooth curves on the right side of \( \mathcal{C} \).

5. Final facts. For condition (4) to hold in an open set of state space it is necessary that the secondary bifurcation of the Hugoniot locus does exist generically, [3]. We
reserve the capital letter $W$ for a suitably chosen neighborhood of the singular point in $\Omega$ and define:

$$\mathcal{D} = \left\{ W \times W \times \mathbb{R} \right\} \setminus \Delta, \quad \Delta = \left\{ (w^-, w^+, u^+) \in \Omega \times \Omega \times \mathbb{R}^+ \mid w^- = w^+ \right\}.$$  \hspace{1cm} (9)

Using Proposition 9, one can see that the set of pairs of states between which doubly characteristic shocks can be constructed is exactly the zero level of the smooth map:

$$\mathcal{F} : \mathcal{D} \rightarrow \mathbb{R}^4, \quad (w^-, w^+, u^+) \mapsto \left( \begin{array} {c} H(w^-, u^- = 1; w^+, u^+; \lambda_c^+(w^+)[G] \end{array} \right), \hspace{1cm} (10)$$

where $\lambda_c^+ = \lambda_c(w^+, u^+)$, $[G] = G(w^+) - G(w^-)$, $I_c$ is the left eigenvector of the thermal family and the function $H$ is given in (4). Such zero level is the secondary bifurcation locus $B = \mathcal{F}^{-1}\{0\}$. From $B$ we can define the projection maps $\pi_+, \pi_- : B \rightarrow W$, $\pi_+(\text{resp.} \pi_-) : (w^-, w^+, u^+) \mapsto w^+$ (resp. $w^-$). We are ready to state:

**Theorem 5.1.** The secondary bifurcation locus $B$ is a smooth two dimensional manifold in $\mathcal{D}$. The projections $\pi_+, \pi_-$ map $B$ diffeomorphically onto $W \setminus \mathcal{C}$.

Since one can use the map $\pi_- \circ \pi_+^{-1}$ in $W \setminus \mathcal{C}$ as the extension map of the fast $e$-rarefactions, Theorem 5.1 gives immediately the properties we sought.

**REFERENCES**


E-mail address: jd@impa.br, marchesi@impa.br
A NOTE ON THE INITIAL-BOUNDARY VALUE PROBLEM FOR
CONTINUITY EQUATIONS WITH ROUGH COEFFICIENTS

GIANLUCA CRIPPA
Departement Mathematik und Informatik, Universität Basel
CH-4051 Basel, Switzerland

CARLOTTA DONADELLO
Laboratoire de Mathématiques, Université de Franche-Comté
F-25030 Besançon Cedex, France

LAURA V. SPINOLO
IMATI-CNR
I-27100 Pavia, Italy

Abstract. In this note we establish the existence of solutions of initial-boundary
value problems for continuity equations with low regularity coefficients. We
also announce a uniqueness result and some related counterexamples.

1. Introduction. We are concerned with the continuity equation
\[ \partial_t u + \text{div} \,(bu) = cu + f, \tag{1} \]
where \( b : ]0,T[ \times \Omega \to \mathbb{R}^d \), \( c : ]0,T[ \times \Omega \to \mathbb{R} \) and \( f : ]0,T[ \times \Omega \to \mathbb{R} \) are given functions
and the unknown is \( u : ]0,T[ \times \Omega \to \mathbb{R} \). Finally, \( \Omega \subseteq \mathbb{R}^d \) is an open set and \( \text{div} \) denotes
the divergence computed with respect to the space variable only. The investigation
of (1) in the case when \( b \) has low regularity is the object of several recent research
papers. Here we only quote the two milestones provided by the works by DiPerna
and Lions [11] and by Ambrosio [1], which deal with the case when \( b \) enjoys Sobolev
and \( BV \) regularity, respectively. We refer to the lecture notes by Ambrosio and
and uniqueness results for the Cauchy problem obtained by coupling (1) with an
initial datum in the case when \( \Omega = \mathbb{R}^d \). We also point out that these results are
motivated by applications to different classes of nonlinear PDEs, see the lecture
notes by De Lellis [10] and the informal overview by Crippa and Spinolo [9] for the
applications concerning systems of conservation laws.

This note aims at establishing existence of solutions of the initial-boundary value
problem for (1) under weak regularity assumptions on \( b \). More precisely, if all
functions were smooth up to the boundary then the formulation of our problem
would read as follows:
\[
\begin{cases}
\partial_t u + \text{div} \,(bu) = cu + f & \text{in } ]0,T[ \times \Omega \\
u = \tilde{g} & \text{on } \Gamma^- \\
u = \hat{u} & \text{on } \{0\} \times \Omega,
\end{cases}
\tag{2}
\]

2000 Mathematics Subject Classification. Primary: 35F16.
Key words and phrases. Continuity equation, initial-boundary value problem, transport equation, low regularity coefficients, existence.
where $\tilde{u}$ and $\tilde{g}$ are bounded smooth functions and $\Gamma^-$ is the subset of $]0,T[\times\partial\Omega$ where the characteristics are entering the domain $]0,T[\times\Omega$. Note, however, that if $b$ and $u$ are not sufficiently regular (if, for instance, $u$ is only an $L^\infty$ function), then their values on zero-measure sets are not well defined. However, in [8] (see also § 2 in here) we introduce a distributional formulation of (2) and we consequently provide a definition of distributional solution, see Definition 2.2 below. This is done by relying on the theory of normal traces for low regularity vector fields, see the works by Anzellotti [4], Chen and Frid [6], Chen, Torres and Ziemer [7] and Ambrosio, Crippa and Maniglia [3]. The main result of this note reads as follows.

**Theorem 1.1.** Let $\Omega \subseteq \mathbb{R}^d$ be an open set with uniformly Lipschitz continuous boundary. Assume that the following conditions hold:

- $b \in L^\infty([0,T[\times\Omega;\mathbb{R}^d)$ and $\text{div} \ b \in L^\infty([0,T[\times\Omega)$;
- $c \in L^\infty([0,T[\times\Omega)$ and $f \in L^\infty([0,T[\times\Omega)$.

Then for every $\tilde{u} \in L^\infty(\Omega)$ and $\tilde{g} \in L^\infty(\Gamma^-)$ there is a distributional solution of problem (2).

Three remarks are here in order. First, we refer to the book by Leoni [12, Definition 12.10] for the definition of open set with uniformly Lipschitz continuous boundary.

Second, the proof of Theorem 1.1 closely follows an argument due to Boyer [5]. The main novelties of Theorem 1.1 compared to the analysis in [5] are: (i) we replace the condition $\text{div} \ b \equiv 0$ with $\text{div} \ b \in L^\infty$ and (ii) we remove the assumptions that $c \equiv 0$ and that $\Omega$ is bounded.

Finally, in [8] we prove that the solution of (2) is unique provided that $b$ enjoys $BV$ regularity up to the boundary of $\Omega$. We also discuss some examples showing that, if $BV$ regularity is violated, then (2) admits, in general, infinitely many solutions. In particular, this happens even if $b$ enjoys $BV$ regularity in every open set compactly contained in $\Omega$ but the $BV$ regularity deteriorates at the boundary $\partial\Omega$.

This note is organized as follows: in § 2 we provide the distributional formulation of problem (2) and in § 3 we give the proof of Theorem 1.1.

1.1. **Notation.**

- $L^n$ : the $n$-dimensional Lebesgue measure.
- $H^m$ : the $m$-dimensional Hausdorff measure.
- $\text{div} \ b$ : the distributional divergence of the vector field $b : \mathcal{O} \to \mathbb{R}^d$, computed with respect to the space variable only.
- $\nabla \varphi$ : the gradient of the Sobolev function $\varphi : \mathcal{O} \to \mathbb{R}$, computed with respect to the space variable only.
- $L^p(\partial\Omega) := L^p(\partial\Omega, H^{d-1})$.
- $L^p([0,T[\times\partial\Omega) := L^p([0,T[\times\partial\Omega, \mathcal{L}^1 \otimes H^{d-1})$.
- $\bar{\Omega}$ : the closure of the set $\Omega \subseteq \mathbb{R}^d$.

2. **Distributional formulation of problem (2).** We first observe that by interpreting the first and the last line of (2) in the sense of distributions we obtain

$$
\int_0^T \int_\Omega u(\partial_t \eta + b \cdot \nabla \eta) \, dx \, dt + \int_\Omega \tilde{u} \eta(0,\cdot) \, dx + \int_0^T \int_\Omega (c \, u \eta + f \eta) \, dx \, dt = 0
$$

for every $\eta \in C^\infty_c([0,T[\times\Omega)$.

The following results is proven in [8] and provides a distributional interpretation of (2) under the solely assumptions that $b$, $\text{div} b$, $c$ and $f$ are all bounded.
Lemma 2.1. Let $\Omega \subseteq \mathbb{R}^d$ be an open set with uniformly Lipschitz continuous boundary. Assume that $b \in L^\infty([0,T] \times \Omega; \mathbb{R}^d)$ satisfies $\text{div} \, b \in L^\infty([0,T] \times \Omega)$. Then the following implications hold:

i) there is a unique function, which we denote by $\text{Tr} \, b$, that belongs to the space $L^\infty([0,T] \times \partial \Omega)$ and satisfies, for every $\varphi \in C_c^\infty([0,T] \times \mathbb{R}^d)$,
\[
\int_0^T \int_{\Omega} \varphi \, \text{div} \, b \, dx \, dt + \int_0^T \int_{\partial \Omega} \partial_t \varphi + b \cdot \nabla \varphi \, dx \, dt = \int_0^T \int_{\partial \Omega} \varphi \, \text{Tr} \, b \, dH^{d-1} \, dt - \int_{\Omega} \varphi(0, \cdot) \, dx.
\]
(4)

Also, the function $\text{Tr} \, b$ satisfies the inequality $\|\text{Tr} \, b\|_{L^\infty} \leq \|b\|_{L^\infty}$.

ii) Assume moreover that $c, f \in L^\infty([0,T] \times \Omega)$ and that $u \in L^\infty([0,T] \times \Omega)$ satisfies (3). Then there is a unique function, which we denote by $\text{Tr} \,(bu)$, that belongs to $L^\infty([0,T] \times \partial \Omega)$ and satisfies, for every $\varphi \in C_c^\infty([0,T] \times \mathbb{R}^d)$,
\[
\int_0^T \int_{\Omega} u(\partial_t \varphi + b \cdot \nabla \varphi) \, dx \, dt + \int_0^T \int_{\partial \Omega} (cu + f) \, \varphi \, dx \, dt = \int_0^T \int_{\partial \Omega} \varphi \, \text{Tr} \,(bu) \, dH^{d-1} \, dt - \int_{\Omega} \bar{u} \, \varphi(0, \cdot) \, dx.
\]
(5)

Also, the function $\text{Tr} \,(bu)$ satisfies the inequality $\|\text{Tr} \,(bu)\|_{L^\infty} \leq \|b\|_{L^\infty} \|u\|_{L^\infty}$.

Some remarks are here in order. First, if the functions are smooth up to the boundary, then the Gauss-Green formula implies that $\text{Tr} \, b = b \cdot \bar{n}$ and $\text{Tr} \,(bu) = ub \cdot \bar{n}$, where $\bar{n}$ is the outward pointing unit normal vector to $\partial \Omega$. Second, in [8] we exhibit an example where $\text{Tr} \, b \equiv 0$ but $\text{Tr} \,(bu) \equiv 1$. The vector field $b$ in the example enjoys $BV$ regularity in every open set compactly contained in $\Omega$, but the $BV$ regularity deteriorates at the boundary. Finally, based on Lemma 2.1, we can rigorously define the sets $\Gamma^-$ and $\Gamma^{0+}$ by setting
\[
\Gamma^- := \{(t, x) \in [0,T] \times \partial \Omega : \text{Tr} \, b < 0\}, \quad \Gamma^{0+} := \{(t, x) \in [0,T] \times \partial \Omega : \text{Tr} \, b \geq 0\}.
\]
(6)

We can now introduce the definition of distributional solution of (2).

Definition 2.2. Let $\Omega \subseteq \mathbb{R}^d$ be an open set with uniformly Lipschitz continuous boundary and assume that $b \in L^\infty([0,T] \times \Omega; \mathbb{R}^d)$, $\text{div} \, b \in L^\infty([0,T] \times \Omega)$, $c \in L^\infty([0,T] \times \Omega)$ and $f \in L^\infty([0,T] \times \Omega)$. A distributional solution of problem (2) is a function $u \in L^\infty([0,T] \times \Omega)$ satisfying (3) such that the equality $\text{Tr} \,(bu) = \bar{g} \, \text{Tr} \, b$ holds on $\Gamma^-$. 

3. Proof of Theorem 1.1. In the following, we establish the existence of functions $u \in L^\infty([0,T] \times \Omega)$ and $\beta \in L^\infty(\Gamma^{0+})$ such that, for every $\varphi \in C_c^\infty([0,T] \times \mathbb{R}^d)$,
\[
\int_0^T \int_{\Omega} u(\partial_t \varphi + b \cdot \nabla \varphi) \, dx \, dt + \int_0^T \int_{\Omega} (cu \varphi + f \varphi) \, dx \, dt = \int_{\Gamma^-} \varphi \, \bar{g} \, \text{Tr} \, b \, dH^{d-1} \, dt + \int_{\Gamma^{0+}} \varphi \beta \, dH^{d-1} \, dt - \int_{\Omega} \bar{u} \varphi(0, \cdot) \, dx.
\]
(7)

By comparing the previous expression with (3) and (5) and by recalling Definition 2.2, we infer that $u$ is a distributional solution of problem (2) and that
\[
\text{Tr} \,(bu)(t, x) = \begin{cases}
\bar{g} \, \text{Tr} \, b & (t, x) \in \Gamma^- \\
\beta & (t, x) \in \Gamma^{0+}.
\end{cases}
\]

The proof of Theorem 1.1 is divided into three steps: in § 3.1 we introduce a second order approximation and we state an existence result for the approximate problem.
In § 3.2 we establish a priori bounds on the family of approximate solutions and in § 3.3 we pass to the limit and obtain a distributional solution of (2).

3.1. **Second order approximation.** We introduce a family of approximate problems, whose classical formulation is the following:

\[
\begin{cases}
\partial_t u_\varepsilon + \text{div}(b u_\varepsilon) = \varepsilon \Delta u_\varepsilon + c u_\varepsilon + f & \text{on } ]0,T[ \times \Omega \\
\varepsilon \frac{\partial u_\varepsilon}{\partial n} + (u_\varepsilon - \bar{u}) [\text{Tr } b]^- = 0 & \text{on } ]0,T[ \times \partial \Omega \\
u_\varepsilon = \bar{u} & \text{on } \{0\} \times \Omega.
\end{cases}
\]

(8)

In the previous expression, \( \varepsilon > 0 \) is a parameter, \( \bar{n} \) as usual denotes the outward pointing unit normal vector to \( \partial \Omega \) and \( [\text{Tr } b]^- \) is the negative part of the function \( \text{Tr } b \). In this section we assume that \( \bar{u}, \bar{g} \) and \( f \), besides being bounded, are also square integrable (see the statement of Lemma 3.2). First, we provide the definition of weak solution of (8). To this end, we introduce the following notation:

- \( V \) : the Sobolev space \( W^{1,2}(\Omega) \).
- \( V^* \) : the dual space of \( V \), endowed with the standard dual norm.
- \( (F, u) \) : the duality between \( F \in V^* \) and \( u \in V \).
- The bilinear form \( B_\varepsilon(t, \cdot) : V \times V \rightarrow \mathbb{R} \) is defined for \( \mathcal{L}^1 \)-a.e. \( t \in ]0,T[ \) as

\[
B_\varepsilon(t, u, v) := -\int_\Omega u b(t, \cdot) \nabla v \, dx + \varepsilon \int_\Omega \nabla u(t, \cdot) \nabla v \, dx
\]

\[
-\int_\Omega c u(t, \cdot) v \, dx + \int_{\partial \Omega} u v [\text{Tr } b]^+(t, \cdot) \, d\mathcal{H}^{d-1},
\]

(9)

where \( [\text{Tr } b]^+ \) denotes the positive part of the function \( \text{Tr } b \).

- The functional \( F(t) \in V^* \) is defined for \( \mathcal{L}^1 \)-a.e. \( t \in ]0,T[ \) by setting

\[
\langle F(t), v \rangle := \int_{\partial \Omega} \bar{v} [\text{Tr } b]^- (t, \cdot) \, d\mathcal{H}^{d-1} + \int_\Omega f(t, \cdot) v \, dx
\]

(10)

Note that continuity of \( v \mapsto \langle F(t), v \rangle \) follows from the square integrability of \( f \) and from the fact that, under the regularity assumptions we impose on \( \partial \Omega \), the trace map is continuous \( V \rightarrow L^2(\partial \Omega) \), see for example Leoni [12, Theorem 15.23].

The following definition is classical, see for example Salsa [14, §9.3.1].

**Definition 3.1.** A weak solution of (8) is a function

\[ u_\varepsilon : [0,T] \rightarrow V \]

such that

1. \( u_\varepsilon \in L^2([0,T]; V) \) and \( \dot{u}_\varepsilon \in ([0,T]; V^*) \), where \( \dot{u}_\varepsilon \) denotes the distributional derivative of \( u_\varepsilon \).
2. For \( \mathcal{L}^1 \)-a.e. \( t \in ]0,T[ \),

\[
\langle \dot{u}_\varepsilon(t), v \rangle + B_\varepsilon(t, u_\varepsilon(t), v) = \langle F(t), v \rangle \quad \forall v \in V.
\]

(11)

3. \( u_\varepsilon(0) = \bar{u} \)

Note that requirement 3 above makes sense since by using requirement 1 we infer that \( u_\varepsilon \in C^0([0,T]; L^2(\Omega)) \), see for example Salsa [14, Theorem 7.22]. Also, note that by the bold letter \( u_\varepsilon \) we denote the function taking values in \( V \), while \( u_\varepsilon \) is the real-valued function \( u_\varepsilon(t, \cdot) = u_\varepsilon(t) \).
Remark 1. By relying on standard arguments, we get that any weak solution of (8) satisfies, for every $\phi \in C_c^\infty([0,T[\times \mathbb{R}^d)$,

$$
\int_0^T \int_\Omega u_\varepsilon (\partial_t \phi + b \nabla \phi) \, dx \, dt - \varepsilon \int_0^T \int_\Omega \nabla u_\varepsilon \nabla \phi \, dx \, dt + \int_0^T \int_\Omega (c u_\varepsilon + f) \phi \, dx \, dt = - \int_\Omega \phi(0,x) \bar{u} \, dx - \int_0^T \int_{\partial \Omega} \bar{g}_\varepsilon \mathrm{Tr} b \, d\mathcal{H}^{d-1} \, dt.
$$

(12)

(13)

The following lemma provides an existence and uniqueness result for (8).

Lemma 3.2. Assume that $b$ and $c$ verify the same assumptions as in the statement of Theorem 1.1 and assume moreover that $\bar{g} \in L^\infty(\Gamma^-) \cap L^2(\Gamma^-)$, $\bar{u} \in L^\infty(\Omega) \cap L^2(\Omega)$ and $f \in L^2([0,T[ \times \Omega) \cap L^\infty([0,T[ \times \Omega)$. Then for any given $\varepsilon > 0$ problem (8) admits a unique weak solution, in the sense of Definition 3.1.

The proof of Lemma 3.2 follows by a classical Faedo-Galerkin method (see for instance [14, §9.3.2]).

3.2. A priori estimates. In this section we establish the estimates we need to study the convergence $\varepsilon \to 0^+$ of the family $u_\varepsilon$ solving (8).

Lemma 3.3. Let $u_\varepsilon$ be the weak solution of problem (8). Then

$$
\|u_\varepsilon(t)\|^2_{L^2(\Omega)} \leq \left(\|\bar{u}\|^2_{L^2(\Omega)} + \int_{\Gamma^-} \bar{g}^2 \mathrm{Tr} b \, d\mathcal{H}^{d-1} \, ds + \|f\|^2_{L^2} \right) \exp \left(\|\mathrm{div} b\|_{L^\infty} + 2\|c\|_{L^\infty} + 1\right)t.
$$

(13)

(14)

for every $t \in [0,T]$. Also,

$$
\int_0^T \int_\Omega |\sqrt{\varepsilon} \nabla u_\varepsilon(t,x)|^2 \, dx \, dt \leq C,
$$

(14)

where $C$ is a constant only depending on $T$, $\|b\|_{L^\infty}$, $\|\mathrm{div} b\|_{L^\infty}$, $\|\bar{g}\|_{L^2}$, $\|\bar{u}\|_{L^2}$, $\|c\|_{L^\infty}$ and $\|f\|_{L^2}$.

Proof. First, we recall (see e.g. [14, Theorem 7.22]) that

$$
\int_0^t \langle \dot{u}_\varepsilon(s), u_\varepsilon(s) \rangle \, ds = \frac{1}{2} \left(\|u_\varepsilon(t)\|^2_{L^2(\Omega)} - \|u_\varepsilon(0)\|^2_{L^2(\Omega)} \right) \quad \forall t \in [0,T].
$$

(15)

Next, we suitably choose the test functions in (4) and we use the density of the space $C_c^\infty(\mathbb{R}^d)$ in $W^{1,1}(\Omega)$ and the continuity of the trace operator from $W^{1,1}(\Omega)$ onto $L^1(\partial \Omega)$, obtaining that for $L^1$-a.e. $t \in [0,T]$

$$
\int_\Omega w \mathrm{div} b(t,\cdot) \, dx + \int_\Omega b(t,\cdot) \cdot \nabla w \, dx = \int_{\partial \Omega} w \mathrm{Tr} b(t,\cdot) \, d\mathcal{H}^{d-1} \quad \forall w \in W^{1,1}(\Omega).
$$

(16)
Lemma 3.4. We now choose \( v := u_\varepsilon(s) \) as a test function in (11), we integrate over \([0, t]\) and we use (16) with \( w = u_\varepsilon(s, \cdot) \). After straightforward computations, we arrive at

\[
\|u_\varepsilon(t)\|_{L^2}^2 + 2\varepsilon \int_0^t \int_\Omega |\nabla u_\varepsilon|^2 dx ds + \int_0^t \int_{\partial \Omega} u_\varepsilon^2 [\text{Tr} b]^+ d\mathcal{H}^{d-1} ds \\
+ \int_0^t \int_{\partial \Omega} (u_\varepsilon - \bar{g})^2 [\text{Tr} b]^- d\mathcal{H}^{d-1} ds \\
= \|\bar{u}\|_{L^2}^2 - \int_0^t \int_\Omega u_\varepsilon^2 \text{div} b dx ds + 2 \int_0^t \int_\Omega c u_\varepsilon^2 dx ds \\
+ \int_0^t \int_{\partial \Omega} \bar{g}^2 [\text{Tr} b]^- d\mathcal{H}^{d-1} ds + 2 \int_0^t \int_\Omega f u_\varepsilon dx ds \\
\leq \|\bar{u}\|_{L^2}^2 + (\|\text{div} b\|_{L^\infty} + 2\|c\|_{L^\infty} + 1) \int_0^t \|u_\varepsilon(s)\|_{L^2}^2 ds \\
+ \int_0^t \int_{\partial \Omega} \bar{g}^2 [\text{Tr} b]^- d\mathcal{H}^{d-1} ds + \|f\|_{L^2}^2.
\]

Hence, by relying on the Gronwall Lemma we get (13) and then (14). \( \square \)

We now establish a maximum principle

**Lemma 3.4.** Let \( u_\varepsilon \) be the weak solution of problem (8), then

\[
\|u_\varepsilon\|_{L^\infty} \leq \left( M + \|f\|_{L^\infty} T \right) \exp \left( (\|\text{div} b\|_{L^\infty} + \|c\|_{L^\infty}) T \right),
\]

where

\[
M := \max\{\|\bar{g}\|_{L^\infty}, \|\bar{u}\|_{L^\infty}\}.
\]

By a slight abuse of notation, in the following we denote by \( m_\varepsilon \) the function \( m_\varepsilon(t, x) := \left( u_\varepsilon(t, x) + (M + \|f\|_{L^\infty} T) e^{B_1} \right)^2 \).

By a slight abuse of notation, in the following we denote by \( m_\varepsilon \) the function \( 2 \left[u_\varepsilon(t, x) + (M + \|f\|_{L^\infty} T) e^{B_1} \right]^- \). Note that, for \( \mathcal{L}^1\)-a.e. \( t \in [0, T] \), \( m_\varepsilon \in L^1(\Omega) \) and \( m_\varepsilon \in V = W^{1,2}(\Omega) \) since

\[
[u_\varepsilon + (M + \|f\|_{L^\infty} T) e^{B_1}]^- \leq \max\{-u_\varepsilon, 0\}.
\]

In the following we also use the formula

\[
\|u_\varepsilon m_\varepsilon\|_{L^\infty} = 2 \left[u_\varepsilon + (M + \|f\|_{L^\infty} T) e^{B_1} \right] m_\varepsilon - M e^{B_1} m_\varepsilon - \|f\|_{L^\infty} t e^{B_1} m_\varepsilon
\]

\[
= 2 m_\varepsilon - M e^{B_1} m_\varepsilon - \|f\|_{L^\infty} t e^{B_1} m_\varepsilon.
\]

ii) We define the function \( m_\varepsilon \) by setting

\[
m_\varepsilon(t, x) := \left( u_\varepsilon(t, x) + (M + \|f\|_{L^\infty} T) e^{B_1} \right)^2.
\]

Also,

\[
\|u_\varepsilon [\text{Tr} b]^+\|_{L^\infty} \leq \left( M + \|f\|_{L^\infty} T \right) \|b\|_{L^\infty} \exp \left( (\|\text{div} b\|_{L^\infty} + \|c\|_{L^\infty}) T \right).
\]

**Proof.** The argument is divided into four steps.

**Step 1.** We introduce some preliminary notation and remarks.

i) We set \( B := \|\text{div} b\|_{L^\infty} + \|c\|_{L^\infty} \).

ii) We define the function \( m_\varepsilon \) by setting

\[
m_\varepsilon(t, x) := \left( u_\varepsilon(t, x) + (M + \|f\|_{L^\infty} T) e^{B_1} \right)^2.
\]

iii) We choose a sequence of smooth cut-off functions \( \{\psi_n\}_{n \in \mathbb{N}} \). More precisely, we require that, for every \( n \in \mathbb{N}, \psi_n \in C^\infty_c(\mathbb{R}^d) \) and

\[
0 \leq \psi_n(x) \leq 1, \quad |\nabla \psi_n(x)| \leq 1/n, \quad \forall x \in \mathbb{R}^d \quad \text{and} \quad \psi_n \equiv 1 \text{ on } B_n(0).
\]
In the previous expression, \( B_n(0) \) denotes the ball of radius \( n \) and center at 0 in \( \mathbb{R}^d \).

iv) Finally, we observe that for every \( n \in \mathbb{N} \) and \( \mathcal{L}^1 \)-a.e. \( t \in ]0, T[ \)

\[
\int_0^t (\dot{u}_\varepsilon(s) + B \left( M + \|f\|_{L^\infty} s \right) e^{B_s} + \|f\|_{L^\infty} e^{B_s} m_\varepsilon' \psi_n) \, ds = \int_\Omega m_\varepsilon(t, x) \psi_n(x) \, dx - \int_\Omega \left( \left[ \tilde{u} + M \right] - \right)^2 \psi_n(x) \, dx.
\]

This formula can be established by relying on an approximation argument, see for example the analysis in the book by Lions and Magenes [13, Section 2.2].

**Step 2.** We use equation (11). First, we observe that by applying (16) with \( n \), \( v \) we obtain

\[
\int \langle \dot{u}_\varepsilon \cdot v \rangle \, dx + \int \left( u_\varepsilon v \div b \right) \, dx + \int v b \cdot \nabla u_\varepsilon \, dx + \varepsilon \int \nabla u_\varepsilon \nabla v \, dx
\]

\[
\quad + \int_{\partial \Omega} u_\varepsilon v \left[ \text{Tr} b \right]^- \, d\mathcal{H}^{d-1} - \int_{\partial \Omega} c u v \, dx
\]

\[
= \int_{\partial \Omega} \tilde{g} v \left[ \text{Tr} b \right]^- \, d\mathcal{H}^{d-1} + \int f v \, dx \quad \forall v \in V \text{ and } \mathcal{L}^1 \text{-a.e. } t \in ]0, T[.
\]

(21)

Next, we fix \( n \in \mathbb{N} \) and we apply (16) with \( w = v u_\varepsilon(t, \cdot) \psi_n \), obtaining

\[
\int_\Omega m_\varepsilon \psi_n \div b \, dx + \int_\Omega m_\varepsilon' \psi_n b \cdot \nabla u_\varepsilon \, dx = \int_{\partial \Omega} m_\varepsilon \psi_n \left[ \text{Tr} b \right] \, d\mathcal{H}^{d-1} - \int_\Omega m_\varepsilon b \cdot \nabla \psi_n \, dx.
\]

(22)

Hence, by plugging \( v := m_\varepsilon'(t, \cdot) \psi_n \) as a test function in (21) and by using (20) and (22) we obtain

\[
(\dot{u}_\varepsilon, m_\varepsilon' \psi_n) + \int_\Omega m_\varepsilon \psi_n \div b \, dx - \int_\Omega M e^{B_t} m_\varepsilon' \psi_n \div b \, dx
\]

\[
- \int_\Omega \|f\|_{L^\infty} t e^{B_t} m_\varepsilon' \psi_n \div b \, dx + \int_\Omega m_\varepsilon \psi_n \left[ \text{Tr} b \right]^+ \, dx + 2 \varepsilon \int_\Omega \chi_\varepsilon |\nabla u_\varepsilon|^2 \psi_n \, dx
\]

\[
- \frac{1}{2} \int_\Omega m_\varepsilon' \psi_n \, dx + \int_\Omega M e^{B_t} m_\varepsilon' \psi_n c \, dx + \int_\Omega \|f\|_{L^\infty} t e^{B_t} m_\varepsilon' \psi_n c \, dx + R_n
\]

\[
= \int_{\partial \Omega} \psi_n \left[ \text{Tr} b \right]^- \left( m_\varepsilon + m_\varepsilon' \left( \bar{g} + (M + \|f\|_{L^\infty} t) e^{B_t} - (u_\varepsilon + (M + \|f\|_{L^\infty} t) e^{B_t}) \right) \right) \, d\mathcal{H}^{d-1}
\]

\[
+ \int_\Omega f m_\varepsilon' \psi_n,
\]

(23)

where \( \chi_\varepsilon \) is the characteristic function of the set where \( u_\varepsilon + (M + \|f\|_{L^\infty} t) e^{B_t} \leq 0 \) and

\[
R_n := - \int_\Omega m_\varepsilon b \cdot \nabla \psi_n \, dx + \varepsilon \int_\Omega m_\varepsilon' \nabla u_\varepsilon \cdot \nabla \psi_n \, dx.
\]
Step 3. We establish an estimate on the time integral of the first and third term in (23). We fix $\tau \in [0, T]$ and we integrate in time to get
\[
\int_0^\tau (\dot{u}_\varepsilon, m_\varepsilon \psi_n) dt - \int_0^\tau \int_\Omega m_\varepsilon \epsilon^B t \psi_n \text{div } b \, dx dt - \int_0^\tau \int_\Omega \|f\|_{L^\infty} t \epsilon^B m_\varepsilon \psi_n \text{div } b \, dx dt \\
+ \int_0^\tau \int_\Omega M \epsilon^B m_\varepsilon \psi_n \epsilon^B t \psi_n c \, dx dt + \int_0^\tau \int_\Omega \|f\|_{L^\infty} t \epsilon^B m_\varepsilon \psi_n c \, dx dt - \int_0^\tau \int_\Omega f m_\varepsilon \psi_n \, dx dt \\
= \int_0^\tau (u_\varepsilon + B Me^B t + \|f\|_{L^\infty} t \epsilon^B, m_\varepsilon \psi_n) dt \\
- \int_0^\tau \int_\Omega m_\varepsilon \psi_n Me^B t (B + \text{div } b - c) \, dx dt \\
- \int_0^\tau \int_\Omega \|f\|_{L^\infty} t \epsilon^B (B + \text{div } b - c) m_\varepsilon \psi_n \\
- \int_0^\tau \int_\Omega \|f\|_{L^\infty} e^B m_\varepsilon \psi_n \, dx dt - \int_0^\tau \int_\Omega f m_\varepsilon \psi_n \, dx dt \\
\geq \int_\Omega m_\varepsilon(\tau, x) \psi_n(x) \, dx.
\]
To get the last inequality, we have used that $\tilde{u} + M \geq 0$, $B + \text{div } b - c \geq 0$, $\|f\|_{L^\infty} e^B t + f \geq 0$ and that $m'_\varepsilon \leq 0$.

Step 4. We conclude the proof of Lemma 3.4. First, we point out that the convexity of the function $z \mapsto (|z|^2 - 2)^2$ implies that
\[
m_\varepsilon + m'_\varepsilon \left(\tilde{\varphi} + (M + \|f\|_{L^\infty} t) e^B t - (u_\varepsilon + (M + \|f\|_{L^\infty} t) e^B t)\right) \\
\leq \left(\tilde{\varphi} + (M + \|f\|_{L^\infty} t) e^B t\right) - (M + \|f\|_{L^\infty} t) e^B t\right)^2 = 0.
\]
Next, we observe that
\[
\int_{\partial \Omega} m_\varepsilon \psi_n [\text{Tr } b]^+ \, dH^{d-1} + 2\varepsilon \int_\Omega \chi_\varepsilon |\nabla u_\varepsilon|^2 \psi_n \, dx \geq 0.
\]
Hence, by time integrating (23) and combining (24) with the above observations we obtain
\[
\int_\Omega m_\varepsilon(\tau, x) \psi_n(x) \, dx + \int_0^\tau \int_\Omega m_\varepsilon \psi_n (\text{div } b - 2\varepsilon) \, dx dt + \int_0^\tau R_n \, dt \leq 0. \tag{25}
\]
Next, we observe that, by recalling estimates (13), (14) and the bounds on $\psi_n$, we get
\[
\lim_{u \to +\infty} \int_0^\tau R_n \, dt \to 0.
\]
Hence, by letting $u \to +\infty$ in (25) and applying the Gronwall Lemma, we conclude that $u_\varepsilon(\tau, x) + e^B t (M + \|f\|_{L^\infty} t) \geq 0$ for $L^{d+1}$-a.e. $(\tau, x) \in \Omega$. By using an analogous argument, we obtain that $u_\varepsilon(\tau, x) \leq e^B t (M + \|f\|_{L^\infty} t)$ for $L^{d+1}$-a.e.
we deduce that, for every \( \varepsilon > 0 \) lying on Lemma 3.4, we infer that the families \( \{ u_{\varepsilon} \} \) satisfy

\[
\text{estimate (14), we can pass to the limit in (12) and obtain a couple } (u, \beta) \text{ satisfying (7).}
\]

For example, we can take \( \bar{u} \) supported by the ERC Starting Grant ConsLaw (P.I. Stefano Bianchini).

Acknowledgments. The authors would like to thank Wladimir Neves for pointing out reference [5]. Donadello and Spinolo also thank the University of Basel for the kind hospitality during their visits. Crippa’s research is partially supported by the Swiss National Science Foundation grant 140232. Donadello acknowledges partial support from the ANR grant CoToCoLa. Spinolo’s participation to HYP2012 was supported by the ERC Starting Grant ConsLaw (P.I. Stefano Bianchini).

REFERENCES


E-mail address: gianluca.crippa@unibas.ch
E-mail address: Carlotta.Donadello@univ-fcomte.fr
E-mail address: spinolo@imati.cnr.it
ON INVISCID LIMITS FOR THE NAVIER-STOKES EQUATIONS
WITH SLIP BOUNDARY CONDITIONS INVOLVING
THE VORTICITY

Stefano Spirito
Universität Basel, Department Matematik und Informatik
Rheinsprung 21, CH-4051
Basel, Switzerland

Luigi C. Berselli
Università degli Studi di Pisa, Dipartimento di Matematica
Via F. Buonarroti I/c, I-56127
Pisa, Italy

Abstract. In this note we consider the inviscid limit for the Navier-Stokes
equations under different slip boundary conditions of Navier’s type and we
show how this influences the convergence rate in the energy norm. The role of
the initial data is also emphasized in connection with the vanishing viscosity
limit.

1. Introduction. In this note the \( L^\infty(0,T;L^2(\Omega)) \) convergence as \( \nu \) goes to 0 of
(Leray-Hopf) weak solutions of the Navier-Stokes equations to the unique local
smooth solution of the Euler equations is studied. In particular, we consider the
case when the inviscid limit is performed in a bounded domain \( \Omega \subset \mathbb{R}^3 \) with smooth
boundary \( \Gamma := \partial \Omega \neq \emptyset \) and we comment and slightly extend similar results recently
proved by Xiao and Xin in \[24\] and obtained independently in the Ph.D. thesis of
the first author. The Navier-Stokes equations in a bounded domain \( \Omega \subset \mathbb{R}^3 \) with
initial datum \( u_0 \) read as follows:

\[
\begin{align*}
\partial_t u^\nu - \nu \Delta u^\nu + (u^\nu \cdot \nabla) u^\nu + \nabla p^\nu &= 0 \quad \text{in } \Omega \times (0,T], \\
\nabla \cdot u^\nu &= 0 \quad \text{in } \Omega \times (0,T], \\
u^\nu(0,x) &= u_0 \quad \text{in } \Omega.
\end{align*}
\]

When the Dirichlet conditions are imposed, \( u^\nu = 0 \) on \( \Gamma \times (0,T] \), it is not known if
there is convergence of the vanishing viscosity even in weak norms towards smooth
solutions of the Euler equations with the same initial data

\[
\begin{align*}
\partial_t u^E + (u^E \cdot \nabla) u^E + \nabla p^E &= 0 \quad \text{in } \Omega \times (0,T], \\
\nabla \cdot u^E &= 0 \quad \text{in } \Omega \times (0,T], \\
u^E \cdot n &= 0 \quad \text{on } \Gamma \times (0,T], \\
u^E(0,x) &= u_0 \quad \text{in } \Omega,
\end{align*}
\]

2000 Mathematics Subject Classification. Primary: 35Q30, 35B25, 35B30; Secondary: 76D05.
Key words and phrases. Navier-Stokes equations, Euler equations, vanishing viscosity limits.
see for example [10] and [16]. In fact, even if both \( u^E \) and \( u^\nu \) are very smooth and both exist in \([0, T]\) (for some positive \( T \) independent of the viscosity) to show, at least, that
\[
\text{as } \nu \to 0 \quad u^\nu(t) \to u^E(t) \text{ in } L^2(\Omega), \quad \text{uniformly in } t \in [0, T],
\]
certain extra-assumptions are needed to be assumed. Some necessary and sufficient conditions, related with the dissipation of energy in a boundary-strip of width \( \nu \) can be found in [13]. See also [20, 21, 14]. Better results can be obtained in the case of the Navier-Stokes equations with Navier’s boundary conditions:
\[
u \cdot n = 0, \quad [D(u^\nu)n + \beta u^\nu]_{\tan} = 0 \quad \text{on } \Gamma \times (0, T),
\]
where \( D(u^\nu) = \frac{1}{2}[\nabla u^\nu + (\nabla u^\nu)^T] \) is the deformation tensor, \( \beta \geq 0 \) is a constant (the friction coefficient) and \([D(u^\nu)n + \beta u^\nu]_{\tan} \) is the tangential component of the vector \( D(u^\nu)n + \beta u^\nu \). In particular in [12] the authors show the convergence of a Leary weak solutions of the Navier-Stokes equations to the unique local smooth solution of (2). These boundary conditions was originally proposed by Navier [17] and studied analytically (in the stationary case) starting from Solonnikov and Ščadilov [18] and Beirão da Veiga [1]. More details on the role of Navier’s boundary conditions especially for numerical simulations, and some of the crucial differences between the two dimensional and three dimensional case, can be found in the review paper [7].

Our aim is to study the convergence under some different slip-without-friction boundary conditions, involving the vorticity. More precisely we will study the following initial-boundary value problem
\[
\partial_t u^\nu - \nu \Delta u^\nu + (u^\nu \cdot \nabla) u^\nu + \nabla p^\nu = 0 \quad \text{in } \Omega \times (0, T],
\]
\[
\nabla \cdot u^\nu = 0 \quad \text{in } \Omega \times (0, T],
\]
\[
u \cdot n = 0 \quad \text{on } \Gamma \times (0, T),
\]
\[
\omega^\nu \times n = 0 \quad \text{on } \Gamma \times (0, T),
\]
\[
u(0, x) = u_0 \quad \text{in } \Omega.
\]
and we will show how the convergence-rate can be improved.

We recall that the interest for these \textit{vorticity based} Navier’s boundary conditions is increasing, especially after the recent results by Xiao and Xin [23] and Beirão da Veiga and Crispo [6, 3, 4, 5], and [8] concerning strong solutions and strong convergence. See also the related work by Xin \textit{et al.} [22, 25]. The aim of our result is to show how the the initial data affect the convergence of the vanishing viscosity even in the energy norm. We point out that the approach in [24] is slightly different from ours and, even though the main idea is the same, we prove some additional results highlighting the role of the initial datum. The main result is Theorem 3.2, which shows improved convergence in the \( L^2 \)-norm when the initial datum has vorticity vanishing at the boundary.

2. Preliminaries. We consider a bounded domain \( \Omega \subset \mathbb{R}^3 \) with smooth boundary \( \Gamma \) and \( n \) denotes the exterior normal unit vector on \( \Gamma \). We will use the classical Lebesgue spaces \( (L^2(\Omega), \| \cdot \|) \) and \( (L^2(\Gamma), \| \cdot \|_{\Gamma}) \) and the Sobolev spaces \( (H^s(\Omega), \| \cdot \|) \) for \( s \in \mathbb{N} \) (we do not distinguish between scalar and vector valued functions). We will denote by \( (H^s(\Gamma), \| \cdot \|_{s, \Gamma}) \) the standard trace spaces on the boundary \( \Gamma \). We will also denote by \( C \) generic constants, which may change from line to line, but which are independent of the viscosity and of the solution of the equations we are considering.
The precise definition of weak solution for the Navier-Stokes equations that we use is the following.

**Definition 2.1.** The vector field \( u^\nu \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega)) \) is a (Leray-Hopf) weak solution of the Navier-Stokes equations (4) if the two following conditions hold:

\[
\begin{align*}
\nu \int_0^T \int_\Omega \left( - u^\nu \phi_t + \nabla u^\nu \nabla \phi + (u^\nu \cdot \nabla) u^\nu \phi \right) \, dx \, dt \\
+ \nu \int_0^T \int_\Gamma u^\nu \cdot (\nabla n)^T \cdot \phi \, dS \, dt - \int_\Omega u_0^\nu \phi(0) \, dx = 0,
\end{align*}
\]

for all \( \phi \in C^\infty_0([0, +\infty[ \times \overline{\Omega}) \) such that \( \nabla \cdot \phi = 0 \) in \( \Omega \times [0, T] \), and \( \phi \cdot n = 0 \) on \( \Gamma \times [0, T] \);

The following energy estimate

\[
\int_\Omega |u^\nu(t)|^2 \, dx + \nu \int_0^t \int_\Omega |\nabla u^\nu|^2 \, dx \, dt + \nu \int_0^t \int_\Gamma u^\nu \cdot (\nabla n)^T \cdot u^\nu \, dS \, dt - \int_\Omega |u_0|^2 \frac{1}{2} \, dx \leq 0,
\]

is satisfied for all \( t \in [0, T] \).

With this definition we have the following existence result.

**Theorem 2.2.** Let be given any positive \( T > 0 \) and \( u_0^\nu \in L^2(\Omega) \) which is divergence-free and such that \( u_0^\nu \cdot n = 0 \) on \( \Gamma \). Then, there exists at least a weak solution of the Navier-Stokes equations (4) \( u^\nu \) on \( [0, T] \).

The proof of global existence of weak solution can be found for instance in [23, §6]. We observe now that the definitions of weak solution in the above reference is slightly different, in particular the energy inequality in [23] is the following,

\[
\int_\Omega |u^\nu(t)|^2 \, dx + \nu \int_0^t \int_\Omega |\nabla u^\nu|^2 \, dx \, dt + \nu \int_0^t \int_\Gamma u^\nu \cdot (\nabla n)^T \cdot u^\nu \, dS \, dt - \int_\Omega |u_0|^2 \frac{1}{2} \, dx \leq 0.
\]

However, by using the next Lemma (see Ref. [2] for the proof), it is easy to prove the equivalence.

**Lemma 2.3.** Let \( u \) and \( \phi \) be two smooth enough vector fields, tangential to the boundary \( \Gamma \). Then it follows

\[
- \int_\Omega \Delta u \phi \, dx = \int_\Omega \nabla u \nabla \phi \, dx - \int_\Gamma (\omega \times n) \phi \, dS + \int_\Gamma u \cdot (\nabla n)^T \cdot \phi \, dS,
\]

where \( \omega = \text{curl} \, u \). Moreover, if \( \nabla \cdot u = 0 \), then \( -\Delta u = \text{curl} \, \text{curl} \, u \), and in addition

\[
- \int_\Omega \Delta u \phi \, dx = \int_\Omega \omega (\text{curl} \, \phi) \, dx + \int_\Gamma (\omega \times n) \phi \, dS.
\]

To conclude, we recall a well-known existence theorem for smooth solutions of the Euler equations (2) in Sobolev spaces.

**Theorem 2.4.** Let be given \( u_0^E \in H^3(\Omega) \) such that \( u_0^E \cdot n = 0 \) on \( \Gamma \), Then there exists a positive time \( T = T(||u_0^E||_3) > 0 \) such that there exists a unique solution of (2) in

\[
u^E \in C([0, T]; H^3(\Omega)).
The proof in the case of a bounded domain can be found in Ebin and Marsden [11] and Temam [19]. In particular \( T \geq \frac{C}{\|u_0\|_3} \), for some \( C > 0 \) independent of the solution and in the sequel \( T \) will be any positive time strictly smaller than the maximal time of existence \( T_{\text{max}} \).

3. Proof of the convergence results. We start by showing the basic convergence result, which is the counterpart of [12, Thm. 1] in our setting.

**Theorem 3.1.** Let \( \Omega \) be a bounded smooth open set in \( \mathbb{R}^3 \), and let \( u_0 \in H^3(\Omega) \), be a divergence-free vector-field tangential to the boundary. Let \( u^\nu \) be a weak solution of the Navier-Stokes equations (4) with initial datum \( u_0 \) and with the vorticity-based Navier’s conditions. Let also \( u^E \in C([0,T];H^3(\Omega)) \) be the unique solution of the Euler equations (2), with the same initial datum \( u_0 \) and defined in some interval \([0,T]\) for some \( T \) depending on \( u_0 \). Then, \( \sup_{t \in [0,T]} \| u^\nu(t) - u^E(t) \|^2 = O(\nu^2) \) and \( \int_0^T \| \nabla u^\nu(\tau) - \nabla u^E(\tau) \|^2 d\tau = O(\nu^2) \).

**Remark 1.** The proof can be easily adapted to handle the case of a sequence of initial data \( u_0^n \in L^2(\Omega) \) for the Navier-Stokes equations with \( \| u_0^n - u_0^E \| = O(\nu^2) \).

**Proof of Thm. 3.1.** The proof is simply obtained by taking the difference of the equation satisfied by \( u^\nu \) with that for \( u^E \), multiplying by \( u^E - u^E \), and integrating by parts over \( \Omega \times (0,T) \). However, since \( u^\nu \) is merely a weak solution and \( u^E \) enjoys the same regularity of \( u^\nu \), we cannot use \( u^\nu \) itself as a test function. We need to pass to an integral formulation and to use the energy inequality (6) to make the argument rigorous. Since \( u^E \) is a smooth solution of the (2) in \([0,T] \times \Omega \), we can multiply the Navier-Stokes equations by \( u^E \) and by integrating we get for all \( t \in [0,T] \)

\[
\int_\Omega u^\nu(t)u^E(t) \, dx + \int_0^t \int_\Omega \left( \nu \nabla u^\nu \cdot \nabla u^E - u^\nu u^E_t + (u^\nu \cdot \nabla) u^\nu u^E \right) \, dx \, d\tau
+ \nu \int_0^t \int_\Gamma (u^\nu \nabla n)^T u^E \, dS \, d\tau = \int_\Omega |u_0|^2 \, dx.
\]  (8)

Note that we have integrated by parts only the pressure and the Laplacian.

Another integral equality is obtained by multiplying the Euler equations by \( u^\nu \).

\[
\int_0^t \int_\Omega \left( u^E u^\nu + (u^E \cdot \nabla) u^E u^\nu \right) \, dx \, d\tau = 0.
\]  (9)

Next, by multiplying the Euler equations by \( u^E \) and by the usual integrations by parts we get the energy conservation

\[
\frac{1}{2} \int_\Omega |u^E(t)|^2 \, dx = \frac{1}{2} \int_\Omega |u_0|^2 \, dx.
\]  (10)

Then, by adding together (6)-(10) and subtracting (8)-(9), we get

\[
\frac{\|u(t)\|^2}{2} + \nu \int_0^t \int_\Omega \nabla u^\nu \cdot \nabla u \, dx \, d\tau + \int_0^t \int_\Omega (u \cdot \nabla) u^E u \, dx \, d\tau
+ \nu \int_0^t \int_\Gamma u^\nu (\nabla n)^T u \, dS \, d\tau \leq 0.
\]  (11)

Let us consider the second term in the left-hand side of (11). By using

\[
\int_0^t \int_\Omega \nabla u^\nu \cdot \nabla u \, dx \, d\tau = \int_0^t \int_\Omega |\nabla u|^2 \, dx \, d\tau + \int_0^t \int_\Omega \nabla u^E \cdot \nabla u \, dx \, d\tau
\]
we get
\[ \frac{\|u(t)\|^2}{2} + \nu \int_0^t \int_{\Omega} |\nabla u|^2 dx \, dt + \nu \int_0^t \int_{\Omega} (u \cdot \nabla) u^E \, dx \, dt \]
\[ + \nu \int_0^t \int_{\Gamma} u^E (\nabla n)^T u \, dS \, d\tau + \nu \int_0^t \int_{\Omega} \nabla u^E \nabla u \, dx \, dt \leq 0. \]
\[ (12) \]

By using the second identity from Lemma 2.3, we get
\[ \nu \int_0^t \int_{\Omega} \nabla u^E \nabla u \, dx \, dt = \nu \int_0^t \int_{\Gamma} (\omega^E \times n) u \, dS \]
\[ - \nu \int_0^t \int_{\Omega} u^E \cdot (\nabla n)^T u \, dx \, dt - \nu \int_0^t \int_{\Omega} \Delta u^E u \, dx \, dt. \]

Then, we get
\[ \frac{\|u(t)\|^2}{2} + \nu \int_0^t \int_{\Omega} |\nabla u|^2 dx \, dt \leq \nu \int_0^t \int_{\Gamma} u^E (\nabla n)^T u \, dS \]
\[ - \int_0^t \int_{\Omega} (u \cdot \nabla) u^E \, dx \, dt - \nu \int_0^t \int_{\Omega} u^E (\nabla n)^T u \, dS \]
\[ + \nu \int_0^t \int_{\Omega} \Delta u^E u \, dx \, dt - \nu \int_0^t \int_{\Gamma} (\omega^E \times n) u \, dS \]
\[ + \nu \int_0^t \int_{\Omega} \Delta u^E u \, dx \, dt \].

We estimate the space integral from the right-hand-side by using Schwartz inequality, trace inequalities and the regularity of $u^E$ as follows:
\[ \nu \left| \int_{\Gamma} (\omega^E \times n) u \, dS \right| \leq C \nu \|u\|_2 \|\nabla u\|_2 \leq C \nu \|u\|^2 + \frac{\nu}{2} \|\nabla u\|^2; \]
\[ \nu \left| \int_{\Gamma} u \cdot (\nabla n)^T u \, dS \right| \leq C \nu \|u\|_2 \|\nabla u\|_2 \leq C \nu \|u\|^2 \|\nabla u\|^2 \leq C \nu \|u\|^2 + \frac{\nu}{2} \|\nabla u\|^2; \]
\[ \nu \left| \int_{\Omega} \Delta u^E u \, dx \right| \leq C [||u||^2 + \nu^2]. \]

Then, by using also (6), we obtain the following differential inequality
\[ \|u(t)\|^2 + \nu \int_0^t \|\nabla u\|^2 \, dx \, dt \leq C \left[ \int_0^t \|u\|^2 \, dx \, dt + \nu^2 + \nu^2 \right]. \]

By using Gronwall-Lemma we have that
\[ \|u^E - u\|_{L_w(0,T;L^2)} = O(\nu^2) \quad \text{and} \quad \|\nabla u^E - \nabla u\|_{L_w(0,T;L^2)} = O(\nu^2). \]
\[ (13) \]

We want now to show better convergence, and this happens if the initial datum belongs to a particular sub-class.

**Theorem 3.2.** Let $\Omega$ be a bounded smooth open set in $\mathbb{R}^3$, and let $u_0 \in H^3(\Omega)$, be a divergence-free vector field tangential to the boundary, and such that
\[ \omega_0(x) = 0 \quad \forall x \in \Gamma. \]
Lemma 3.3. Let \( u^\nu \) be a weak solution of the Navier-Stokes equations (4) with initial datum \( u_0 \). Let \( u^E \in C([0,T]; H^3(\Omega)) \) the unique solution of the Euler equations (2), with the same initial datum \( u_0 \) and defined in some interval \([0,T]\). Then,
\[
\sup_{t \in [0,T]} \|u^\nu - u^E\|_{L^\infty(0,T; L^2)} = O(\nu^2), \quad \text{and} \quad \|\nabla u^\nu - \nabla u^E\|_{L^2(0,T; L^2)}^2 = O(\nu).
\]

A critical point in the proof is that in general the unique smooth solution of the Euler equations does not satisfy the boundary condition for the vorticity even if the initial datum is such that \( \omega^E_0 \times n = 0 \). Indeed, as observed in [4], by using the vorticity equation,
\[
\omega^E_t + (u^E \cdot \nabla) \omega^E = (\omega^E \cdot \nabla) u^E,
\]
by taking the exterior product with the normal unit vector on \( \Gamma \), and finally by using that \( \omega^E \times n = 0 \) implies that \( \omega^E_t \times n = 0 \) on \( \Gamma \), one has as a consequence that an extra-compatibility condition, generically false, should be satisfied by the initial velocity \( u^E_0 \). In particular, this implies that when the vanishing-viscosity limit is performed in topologies such that the vorticity \( \omega^E \) has Sobolev traces on the boundary the vorticity in general at least for small time \( \omega^E \) does not satisfy the Navier’s type conditions. On the other hand, by using the fact that for the Euler equations the vorticity is transported by the velocity \( u^E \) and stretched by \( \nabla u^E \), one has the well-known representation formula for classical solutions
\[
\omega^E(X(\alpha, t), t) = \nabla_\alpha X(\alpha, t) \omega^E(\alpha, 0),
\]
where the streamlines \( X : \Omega \times [0,T] \rightarrow \Omega \subset \mathbb{R}^3 \) solve the Cauchy problem
\[
\frac{d}{dt} X(\alpha, t) = u^E(X(\alpha, t), t),
\]
\[
X(\alpha, 0) = \alpha,
\]
for \( t \in [0,T] \) and \( \alpha \in \Omega \). Since \( u^E \cdot n = 0 \) on the boundary, the streamlines starting on the boundary remain on the boundary for all positive times. Let us consider \( \overline{\alpha} \in \Gamma \), then \( X(\overline{\alpha}, t) \in \Gamma \) for all \( t \in [0,T] \) and consequently
\[
\omega^E(X(\overline{\alpha}, t), t) \times n = \left[ \nabla_\alpha X(\overline{\alpha}, t) \omega^E(\overline{\alpha}, 0) \right] \times n, \quad \forall (\overline{\alpha}, t) \in \Gamma \times [0,T].
\]
Let us note that generically the matrix \( \nabla_\alpha X(\overline{\alpha}, t) \) not a multiple of the identity, this implies that generically \( \omega^E(X(\overline{\alpha}, t), t) \times n \neq 0 \). However, this observation makes clear that it is possible to get that the solutions of the Euler equations satisfy the boundary conditions for the vorticity by suitably restricting the class of initial data.

Lemma 3.3. Let \( u^E \in C([0,T]; H^3(\Omega)) \) be the unique solution of the Euler equations (2). If \( \omega^E_0(x) = 0 \) for all \( x \in \Gamma \), then \( \omega^E(x, t) = 0 \) for all \( (x, t) \in \Gamma \times [0,T] \).

Remark 2. The class of solutions with vanishing-vorticity of the boundary we will employ is in some sense optimal, since in [5] it is shown that if \( \omega_0 \cdot n \neq 0 \), then there exists a time \( T > 0 \) such that the \( \omega^E \times n \neq 0 \) for \( 0 < t \leq T \). Moreover this class of initial data in non-empty, since smooth and divergence-free functions with compact support satisfy the assumptions.

Remark 3. More detailed calculations and also applications to Boussinesq equations appear in [9].

Proof of Theorem 3.2. The proof follows by using the same techniques of that of Theorem 3.1 and obtaining better estimates for the boundary term by assuming that the initial datum is such that the vorticity vanishes as the boundary. In fact,
we have seen that by assuming (14) we have \( \omega^E(x,t) = 0 \) for all \( (x,t) \in \Gamma \times [0,T] \).

Writing again the same energy estimates we employed before to get (12), and by using

\[
\nu \int_0^t \int_\Omega \nabla u^E \nabla u \, dx \, dt = \nu \int_0^t \int_\Gamma \nabla u^E \cdot (\nabla n)^T \cdot u \, dS \, dt + \nu \int_0^t \int_\Omega \Delta u^E u \, dx \, dt,
\]

(notice that now the boundary integral \( \int_\Gamma (\omega^E \times n) \partial S \) vanishes) we can re-do the same calculations as before starting from (12) to obtain now

\[
\|u(t)\|^2 + \nu \int_0^t \|\nabla u\|^2 \, dt \leq C \left[ \int_0^t \|u(\tau)\|^2 \, d\tau + \nu^2 \right].
\]

Finally, by Gronwall’s inequality we get

\[
\sup_{t \in [0,T]} \|u'' - u^E\|_{L^\infty(0,T;L^2)}^2 = O(\nu^2), \quad \text{and} \quad \|\nabla u'' - \nabla u^E\|_{L^2(0,T;L^2)}^2 = O(\nu).
\]

\( \square \)

REFERENCES


E-mail address: stefano.spirito@unibas.ch
E-mail address: berselli@dma.unipi.it
METASTABLE DYNAMICS OF INTERFACES FOR THE HYPERBOLIC JIN-XIN SYSTEM

MARTA STRANI
Sapienza, Università di Roma
Dipartimento di Matematica
Piazzale A.Moro 2, 00185 Roma, Italy
and
Ecole Normale Supérieure de Paris
Département de Mathématiques et applications
45 rue d’Ulm, 75005 Paris, France

Abstract. We consider the slow motion of internal shock layer exhibited by the solution to the initial boundary-value problem for the Jin-Xin system in one space dimension. In this contest, the time-dependent solution drifts towards its steady state in an exponentially long time interval. The derivation of an ODE for the shock layer position is the main tool of our analysis.

1. Introduction. Let us consider the initial-boundary value problem for the quasi-linear Jin-Xin system in the bounded interval $I = (-\ell, \ell)$, with Dirichlet boundary conditions, that is

$$\begin{cases}
\partial_t u + \partial_x v = 0 & x \in I, \ t > 0 \\
\partial_t v + a^2 \partial_x u = \frac{1}{\varepsilon} (f(u) - v) \\
u(\pm \ell, t) = u_{\pm} \\
u(x, 0) = u_0(x), \ v(x, 0) = v_0(x) \equiv f(u_0(x)) & x \in I
\end{cases}$$

(1)

for some $\varepsilon, \ell, a > 0$, $u_{\pm} \in \mathbb{R}$ and flux function $f$ satisfying

$$f''(s) \geq c_0 > 0, \ f'(u_+) < 0 < f'(u_-), \ f(u_+) = f(u_-).$$

In compact form, system 1 can be rewritten as

$$\partial_t Z = \mathcal{F}[Z] , \ \left. Z \right|_{t=0} = Z_0,$$

where

$$Z = \begin{pmatrix} u \\ v \end{pmatrix} , \ \mathcal{F}[Z] := \begin{pmatrix} P_1[Z] \\ P_2[Z] \end{pmatrix} = \begin{pmatrix} -\partial_x v \\ -a^2 \partial_x u + \frac{1}{\varepsilon} (f(u) - v) \end{pmatrix}$$

This model was firstly introduced in [7] as a numerical scheme approximating solutions of the hyperbolic conservation law $\partial_t u + \partial_x f(u) = 0$, and it is a particular case of a more general class of hyperbolic relaxation systems, usually utilized to model a variety of non equilibrium processes in continuum mechanics: for example, non-thermal equilibrium gas dynamics ([10], [13]), traffic dynamics ([2], [11], [12]).

2000 Mathematics Subject Classification. Primary: 35L40, 35L04; Secondary: 35L65.
Key words and phrases. Metastability, slow motion, internal layers, relaxation systems.
and multiphase flows ([3], [4], [15]). Here $\varepsilon$ is a positive parameter, usually small, determining the relaxation time.

Our main goal is to show that, in the relaxation limit (i.e. $\varepsilon \to 0$), the solution to 1 exhibits a metastable behavior, that is there exists a first time scale in which such solution is close to some non-stationary state for an exponentially long time before converging to its asymptotic limit. Numerical computations show that, starting, for example, with a decreasing initial datum $u_0(x)$ (see Fig.1), a shock layer is formed in a $O(1)$ time scale. Once the shock layer is formed, it moves towards the equilibrium solution, and this motion is exponentially slow. As a consequence, two different time scales emerge: a first transient phase where an internal shock layer is formed in a $O(1)$ time scale, and a subsequent exponentially slow motion where the layer drifts toward its asymptotic limit. Concerning the function $v$, starting with the initial datum $v_0(x) = f(u_0(x))$, we can observe that the position of the shock of $u$ corresponds to the location the minimum value of the function $v$; so we have a first transient phase in which the profile of $v$ stabilizes, and an exponentially slow phase where the value of the minimum of such profile drifts towards the value $\xi$ that represents the location of the equilibrium solution for $u$.

![Figure 1](image1.png)

**Figure 1.** Profiles of $(u, v)$, solutions to 1, with $f(u) = u^2/2$, $a = 1$, $\varepsilon = 0.04$ and $u_\pm = \mp 1$. The initial datum is given by the couple $(u_0(x), f(u_0(x)))$, $u_0(x)$ being a decreasing function connecting $u_+$ and $u_-$. Profiles at times $t = 0, 0.2, 10, 10^5, 0.5 \times 10^6$ are depicted.

A large class of evolutive PDEs, concerning many different areas, exhibits this behavior. Among others, we include viscous shock problems (see [9], [18], [20]), phase
transition problems described by the Allen-Cahn equation, with the fundamental contributions [5], [6], and Cahn-Hilliard equation, studied in [1] and [17].

Usually such behavior is related to the presence of a first eigenvalue $\lambda_1^\varepsilon$ for the linearized operator around an equilibrium state that is small in $\varepsilon$, so that the large time behavior of solutions is described by terms of order $\varepsilon$, and the convergence to the asymptotically stable state is much slower as $\varepsilon \to 0$.

To begin with, let us consider the case when $\varepsilon = 0$, where system 1 can be approximated to leading order by

$$\partial_t u + \partial_x f(u) = 0, \quad v = f(u).$$  \hfill (2)

In this case, given $\xi \in (-\ell, \ell)$, equation 2 admits a large class of stationary solutions satisfying the boundary conditions, given by all that piecewise constant functions on the form

$$U_{hyp}(x; \xi) = u_+ \chi_{(-\ell, \xi)}(x) + u_- \chi_{(\xi, \ell)}(x)$$
$$V_{hyp}(x; \xi) = f(u_+) \chi_{(-\ell, \xi)}(x) + f(u_-) \chi_{(\xi, \ell)}(x),$$

so that we can construct a one-parameter family $\{U_{hyp}(\cdot; \xi), V_{hyp}(\cdot; \xi)\}$ of steady states, parametrized by $\xi$ that represents the location of the jump.

For $\varepsilon > 0$ the situation is very different: by differentiating with respect to $x$ the second equation to 1, we obtain

$$u_t = \varepsilon (a^2 \partial_{xx} u - \partial_t u) - \partial_x f(u).$$

Thus, stationary solutions to 1 solve

$$a^2 \varepsilon \partial_{xx} u = \partial_x f(u),$$  \hfill (3)

together with $\partial_x v = 0$. In this case, only a single stationary state, denoted here by $(\bar{U}^\varepsilon(x), \bar{V}^\varepsilon(x))$, is admitted (see [8]), and it is possible to prove that it is asymptotically stable, i.e. starting from an initial datum close to such equilibrium configuration, the time dependent solution approaches the steady state for $t \to +\infty$.

Moreover, the stationary solution $(\bar{U}^\varepsilon(x), \bar{V}^\varepsilon(x))$ converges as $\varepsilon \to 0$ pointwise to $(U_{hyp}(\cdot; \bar{\xi}), V_{hyp}(\cdot; \bar{\xi}))$, for some $\bar{\xi} \in I$.

The aim of this paper is to understand what happens to the dynamics generated by an initial datum localized far from the equilibrium solution and to determine a detailed description of the behavior of the solution in the relaxation limit. The strategy we use is the following: we build up a one parameter family of approximate steady states

$$\{\mathbf{W}^\varepsilon(x; \xi)\}_{\xi \in I} = \{U^\varepsilon(\cdot; \xi), V^\varepsilon(\cdot; \xi)\}_{\xi \in I},$$

where $\xi$ represents the location of the shock layer and such that

- $(U^\varepsilon(\cdot; \xi), V^\varepsilon(\cdot; \xi)) := (\bar{U}^\varepsilon, \bar{V}^\varepsilon)$ for some $\bar{\xi} \in I$
- $(U^\varepsilon(\cdot; \xi), V^\varepsilon(\cdot; \xi)) \to (U_{hyp}(\cdot; \bar{\xi}), V_{hyp}(\cdot; \bar{\xi}))$ as $\varepsilon \to 0$ in a sense to be specified.

When $\xi \neq \bar{\xi}$, an element of this family can be seen as an approximate stationary solution to the problem, in the sense that it satisfies the stationary equation up to an error that is small in $\varepsilon$. Indeed, following the idea firstly introduced in [14], we require that there exist two families of smooth positive functions $\Omega_1^\varepsilon = \Omega_1^\varepsilon(\xi)$ and $\Omega_2^\varepsilon = \Omega_2^\varepsilon(\xi)$, uniformly convergent to zero as $\varepsilon \to 0$, such that, for any $\xi \in I$, the following estimates hold

$$|\langle \psi(\cdot), \mathcal{P}_1[\mathbf{W}^\varepsilon(\cdot; \xi)]\rangle| \leq \Omega_1^\varepsilon(\xi) |\psi|_{L^\infty} \quad \forall \psi \in C(I)$$
$$|\langle \psi(\cdot), \mathcal{P}_2[\mathbf{W}^\varepsilon(\cdot; \xi)]\rangle| \leq \Omega_2^\varepsilon(\xi) |\psi|_{L^\infty} \quad \forall \psi \in C(I).$$
Hence we require the error

\[
\begin{pmatrix}
\mathcal{P}_1^\varepsilon [W^\varepsilon] \\
\mathcal{P}_2^\varepsilon [W^\varepsilon]
\end{pmatrix} :=
\begin{pmatrix}
- \partial_x V^\varepsilon \\
-a^2 \partial_x U^\varepsilon + \frac{1}{\varepsilon} (f(U^\varepsilon) - V^\varepsilon)
\end{pmatrix}
\] (4)

that describes how far is $W^\varepsilon(x; \xi)$ from being an exact stationary solution, to be small in $\varepsilon$ in the sense described above.

Finally, in order to derive an equation of motion for the shock layer position, we study the dynamics in a neighborhood of this family, by linearizing the original system around an element of $\{W^\varepsilon(x; \xi)\}_{\xi \in I}$.

2. Spectral analysis for the linearized operator. We look for a solution to 1 in the form

\[u(\cdot, t) = U^\varepsilon(\cdot; \xi(t)) + u^1(\cdot, t), \quad v(\cdot, t) = V^\varepsilon(\cdot; \xi(t)) + v^1(\cdot, t).\]

Substituting in 1, we obtain the linearized system around $W^\varepsilon$

\[
\begin{aligned}
\partial_t u^1 &= - \partial_x v^1 - \partial_\xi U^\varepsilon(\cdot; \xi) \frac{d\xi}{dt} + \mathcal{P}_1^\varepsilon [W^\varepsilon(\cdot; \xi)] \\
\partial_t v^1 &= - a^2 \partial_x u^1 + \frac{1}{\varepsilon} (f'(U^\varepsilon(\cdot; \xi))u^1 - v^1) - \partial_\xi V^\varepsilon(\cdot; \xi) \frac{d\xi}{dt} + \mathcal{P}_2^\varepsilon [W^\varepsilon(\cdot; \xi)] + Q^\varepsilon[u^1]
\end{aligned}
\] (5)

where $\mathcal{P}_1^\varepsilon [W^\varepsilon]$ and $\mathcal{P}_2^\varepsilon [W^\varepsilon]$ are defined in 4, and $Q^\varepsilon[u] := o(u)$. Moreover, we set $(u^1, v^1) = (u, v)$ and

\[Y = \begin{pmatrix} u \\ v \end{pmatrix}, \quad \mathcal{L}_\xi Y := \begin{pmatrix} - \partial_x v \\ -a^2 \partial_x u + \frac{1}{\varepsilon} (f'(U^\varepsilon)u - v) \end{pmatrix} \]

As already stressed before, the metastable behavior is usually related to the presence of a first small eigenvalue for the linearized operator around an equilibrium configuration. In this case, the eigenvalue problem for $\mathcal{L}_\xi$ reads

\[
\begin{aligned}
\lambda \varphi &= - \partial_x \psi \\
\lambda \psi &= - a^2 \partial_x \varphi + \frac{1}{\varepsilon} (f'(U^\varepsilon)\varphi - \psi)
\end{aligned}
\]

complemented with Dirichlet boundary conditions. Hence, by differentiating the second equation with respect to $x$, we obtain

\[\varepsilon a^2 \partial_{xx} \varphi - \partial_x (f'(U^\varepsilon)\varphi) = \lambda (1 + \varepsilon \lambda) \varphi.\]

We now use the spectral results obtained in [14] for the differential linear diffusion-transport operator

\[\mathcal{L}^{\varepsilon, \text{usc}} \varphi := a^2 \partial_{xx} \varphi - \partial_x (b^\varepsilon \varphi), \quad b^\varepsilon(x; \xi) := f'(U^\varepsilon(x; \xi)),\]

(for more details see [14, Section 4]); since $\lambda$ is an eigenvalue of $\mathcal{L}_\xi$ if and only if $\lambda^{\varepsilon, \text{usc}} := \lambda(1 + \varepsilon \lambda)$ is an eigenvalue for the operator $\mathcal{L}^{\varepsilon, \text{usc}}$, we can prove the following

**Proposition 1.** Let $b^\varepsilon$ be a family of functions satisfying the assumptions:

\[A0. \text{ There exists } C > 0, \text{ independent on } \varepsilon > 0, \text{ such that} \]

\[|b^\varepsilon|_{L^\infty} + \varepsilon \left| \frac{db^\varepsilon}{dx} \right|_{L^\infty} \leq C.\]
If there exist \( \xi \in C^0[-\ell, \ell] \), \( b^\varepsilon \) is twice differentiable at any \( x \neq \xi \) and
\[
\frac{db^\varepsilon}{dx}, \frac{d^2b^\varepsilon}{dx^2} < 0 < b^\varepsilon \quad \text{in} \quad (-\ell, \xi), \quad \text{and} \quad b^\varepsilon \frac{db^\varepsilon}{dx} < 0 < \frac{d^2b^\varepsilon}{dx^2} \quad \text{in} \quad (\xi, \ell)
\]

**A1.** \( b^\varepsilon \in C^0[-\ell, \ell] \), \( b^\varepsilon \) is twice differentiable at any \( x \neq \xi \) and
\[
\frac{db^\varepsilon}{dx}, \frac{d^2b^\varepsilon}{dx^2} < 0 < b^\varepsilon \quad \text{in} \quad (-\ell, \xi), \quad \text{and} \quad b^\varepsilon \frac{db^\varepsilon}{dx} < 0 < \frac{d^2b^\varepsilon}{dx^2} \quad \text{in} \quad (\xi, \ell)
\]

**A2.** For any \( C > 0 \) there exists \( c_0 > 0 \) such that, for any \( x \) satisfying \( |x-\xi| \geq c_0\varepsilon \), there holds
\[
|b^\varepsilon - b^0| \leq C\varepsilon \quad \text{and} \quad \varepsilon \left| \frac{db^\varepsilon}{dx} \right| \leq C.
\]

**A3.** There exist the left/right first order derivatives of \( b^\varepsilon \) at \( \xi \) and
\[
\liminf_{\varepsilon \to 0^+} \varepsilon \left| \frac{db^\varepsilon}{dx}(\xi \pm) \right| > 0.
\]

If there exist \( \xi \in (-\ell, \ell) \), \( b_+ < 0 < b_- \) and \( C > 0 \) for which \( |b^\varepsilon - b^0|_{L^1} \leq C\varepsilon \), where \( b^0 \) is the step function jumping from \( b^- \) to \( b^+ \), then the spectrum of the linearized operator \( L^\varepsilon_\xi \) can be decomposed as follow:

1. \( \lambda_{1X}^\varepsilon \in \mathbb{R} \) and \( -e^{-c/\varepsilon} \leq \lambda_{1X}^\varepsilon < 0 \).
2. All the remaining eigenvalues \( \lambda_{nX}^\varepsilon \) are such that
\[
\text{Re}[\lambda_{nX}^\varepsilon] \leq -C/\varepsilon, \quad \forall n \geq 2.
\]

For more details and for the proof of Proposition 1, we refer to [19].

**Remark 1.** In the special case of a Burgers flux, i.e. \( f(u) = u^2/2 \) and \( u_\pm = \mp u^* \), it is possible to obtain an explicit asymptotic expression for the first eigenvalue. More precisely
\[
|\lambda_{1X}^\varepsilon(\xi)| \sim \frac{\varepsilon \cdot u^2}{\pm \left[ e^{-u^*\varepsilon^{-1}(\ell-\xi)} + e^{-u^*\varepsilon^{-1}(\ell+\xi)} \right]}.
\]

This formula shows that the principal eigenvalue of the Jin-Xin system with \( f(u) = u^2/2 \) is exponentially small in \( \varepsilon \).

Now let us go back to system 5. Since we have proved that the first eigenvalue of the linearized operator tends to zero as \( \varepsilon \to 0 \), denoting by \( \psi_k^\varepsilon = \psi_k^\varepsilon(\xi, \cdot) \) the eigenfunctions of the corresponding adjoint operator \( L^\varepsilon_\xi^* \) and setting \( Y_k = Y_k(\xi; t) := \langle \psi_k^\varepsilon(\cdot; \xi), Y(\cdot, t) \rangle \), we impose that the component \( Y_1 = (u_1, v_1) \) is identically zero. More precisely, we set an algebraic condition ensuring orthogonality between \( \psi_1^\varepsilon \) and \( Y \), in order to remove the singular part of the operator \( L^\varepsilon_\xi^* \).

Thus, denoting by \( Y_0 \) the initial datum of the perturbation, we have
\[
\frac{d}{dt} \langle \psi_1^\varepsilon(\cdot; \xi(t)), Y(\cdot, t) \rangle = 0 \quad \text{and} \quad \langle \psi_1^\varepsilon(\cdot; \xi_0), Y_0(\cdot) \rangle = 0. \tag{6}
\]
Plugging 5 into 6, and by linearizing around \( Y \sim 0 \), we end up with the following coupled system for the shock layer location \( \xi(t) \) and the perturbation \( Y \)
\[
\begin{align*}
\frac{d\xi}{dt} &= \theta^\varepsilon(\xi) \left( 1 + \frac{\langle \partial_\xi \psi^\varepsilon_1, Y \rangle}{\langle \psi^\varepsilon_1(\cdot; \xi), \partial_\xi \psi^\varepsilon_1(\cdot; \xi) \rangle} \right) \tag{7}
\end{align*}
\]
\[
\partial_\xi Y = (L^\varepsilon_\xi + M^\varepsilon_\xi) + H^\varepsilon(x; \xi)
\]
where

\[ \theta^\varepsilon(\xi) := \frac{\langle \psi_1^\varepsilon, F^\varepsilon[W^\varepsilon] \rangle}{\langle \psi_1^\varepsilon(\cdot, \xi), \partial_\xi W^\varepsilon(\cdot, \xi) \rangle} \]

\[ \mathcal{M}^\varepsilon Y := \frac{1}{\alpha_0(\xi)} \begin{pmatrix} -\partial_\xi U^\varepsilon(\cdot, \xi) \theta^\varepsilon(\xi) \langle \partial_\xi \psi_1^\varepsilon(\cdot, \xi), Y \rangle \\ -\partial_\xi V^\varepsilon(\cdot, \xi) \theta^\varepsilon(\xi) \langle \partial_\xi \psi_1^\varepsilon(\cdot, \xi), Y \rangle \end{pmatrix} \]

\[ H(x; \xi) := \begin{pmatrix} P_1^\varepsilon[W^\varepsilon(\cdot, \xi)] - \partial_\xi U^\varepsilon(\cdot, \xi) \theta^\varepsilon(\xi) \\ P_2^\varepsilon[W^\varepsilon(\cdot, \xi)] - \partial_\xi V^\varepsilon(\cdot, \xi) \theta^\varepsilon(\xi) \end{pmatrix} \]

**Remark 2.** We stress that system 7 is obtained by linearizing with respect to \( Y \) so that all the nonlinear terms of order \( o(Y) \) are neglected. On the other hand, we keep the nonlinear dependence on \( \xi \), in order to describe the slow motion of the shock layer position far from the equilibrium location \( \bar{\xi} \).

3. The equations of motion. Before stating our main result, let us recall the assumptions. Let the family \( \{ W^\varepsilon(\cdot, \xi) \} \) be such that there exist two families of smooth positive functions \( \Omega_1^\varepsilon \) and \( \Omega_2^\varepsilon \) such that

\[ |\langle \psi(\cdot), P_1^\varepsilon[W^\varepsilon(\cdot, \xi)] \rangle| \leq \Omega_1^\varepsilon(\xi) |\psi|_{L^\infty} \quad \forall \psi \in C(I) \]

\[ |\langle \psi(\cdot), P_2^\varepsilon[W^\varepsilon(\cdot, \xi)] \rangle| \leq \Omega_2^\varepsilon(\xi) |\psi|_{L^\infty} \quad \forall \psi \in C(I). \]

We also assume that \( W^\varepsilon \) is asymptotically a stationary solution, i.e. we require that

\[ \lim_{\varepsilon \to 0} |\Omega_1^\varepsilon|_{L^\infty} = 0, \quad \lim_{\varepsilon \to 0} |\Omega_2^\varepsilon|_{L^\infty} = 0 \]

uniformly with respect to \( \xi \). For what concerns the eigenvalues of the linear operator \( L^\varepsilon_x \), we assume that there exists a constant \( C_3 > 0 \) such that

\[ \Omega_1^\varepsilon(\xi) + \Omega_2^\varepsilon(\xi) \leq C_3 |\lambda^X_{1+}(\xi)| \quad \forall \xi \in (-\ell, \ell). \]

Finally, concerning the solution \( Z = (z, w)^T \) to the linear problem \( \partial_t Z = L^\varepsilon_x Z \), we require that there exists \( \nu > 0 \) such that for all \( \xi \in (-\ell, \ell) \), there exist constants \( C_\xi \) and \( \bar{C} \) such that

\[ |(z, w)(t)|_{L^2} \leq C_\xi |(z_0, w_0)|_{L^2} e^{-\mu t}, \quad C_\xi \leq \bar{C} \quad \forall \xi \in (-\ell, \ell). \]

Under these hypotheses we can prove an estimate for the \( L^2 \)-norm of the perturbation \( Y \). The idea of the proof is based on the theory of stable families of generators (see [16]), and we refer to [19] for more details.

**Theorem 3.1** (Estimate for the perturbation). Let hypotheses 8, 9, 10, 11 be satisfied. Then, for \( \varepsilon \) sufficiently small, there exists a time \( T > 0 \) such that, for all \( t \leq T \), the solution \( Y \) to 7 is such that

\[ |Y|_{L^2}(t) \leq [C_1 |\Omega_1^\varepsilon|_{L^\infty} + C_2 |\Omega_2^\varepsilon|_{L^\infty}] \left( 1 - e^{-\mu^* t} \right) + e^{-\mu^* t} |Y_0|_{L^2} \]

for some positive constants \( C_1, C_2 \) and

\[ \mu^* := \sup_{\xi} \lambda^X_{1+}(\xi) - C(|\Omega_1^\varepsilon|_{L^\infty} + |\Omega_2^\varepsilon|_{L^\infty}) > 0. \]

As a consequence of the estimate 12, we can prove a result that shows the metastable behavior of the solution to 1 in terms of the slow motion of the shock layer location. Indeed, 12 can be used to decouple system 7 so that, by means of the standard method of separation of variables, one obtains an estimate for \( \xi(t) \) (for the proofs of the statements see [19, Section 5]).
Theorem 3.2 (Slow motion of the shock layer). Let hypotheses 8, 9, 10, 11 be satisfied. Assume also
\[ s \theta'(s) < 0 \quad \text{for any } s \in I, s \neq 0 \quad \text{and} \quad \theta''(\xi) < 0. \]
Then for \( \varepsilon \) and \( |Y_0|_{L^2} \) sufficiently small, the solution \( \xi(t) \) converges to \( \bar{\xi} \) as \( t \to +\infty \).

Corollary 1. Let the hypotheses of Theorem 3.2 be satisfied, then
\[ |\xi(t) - \bar{\xi}| \leq |\xi_0|e^{\beta t}, \quad \beta \sim \theta''(\bar{\xi}) \]
where \( \theta''(\bar{\xi}) \to 0 \) as \( \varepsilon \to 0 \).

Remark 3. Formula 13 shows the exponentially slow motion of the shock layer for small \( \varepsilon \). Precisely, the evolution of the collocation of the shock towards the equilibrium position is much slower as \( \varepsilon \) becomes smaller.

Example 3.3. To see how such formulas can be handled let us consider system 1 in the case of a Burgers flux, i.e. \( f(u) = u^2/2 \) with boundary conditions \( u(\pm \ell) = \mp u^* \), for some \( u^* > 0 \).

In this case, an approximate solution \( U^\varepsilon(\ell; \xi) \) to 3 is obtained by matching two different steady states satisfying, respectively, the left and the right boundary conditions together with the request \( U^\varepsilon|_{x=\ell} = 0 \). Hence, an explicit formula for an approximate stationary solution is the following

\[ U^\varepsilon(x; \xi) = \begin{cases} k_+ \tanh (k_+ (\ell - x)/2\varepsilon) & \text{in } (-\ell, \xi) \\ k_- \tanh (k_- (\ell - x)/2\varepsilon) & \text{in } (\xi, \ell) \end{cases} \]

where \( k_\pm \) are chosen so that the boundary conditions are satisfied and \( a = 1 \) to simplify notations. In this specific case it is also possible to compute \( \mathcal{P}_1^\varepsilon[W^\varepsilon(;\xi)] \) and \( \mathcal{P}_2^\varepsilon[W^\varepsilon(;\xi)] \). We have

\[ \mathcal{P}_1^\varepsilon[W^\varepsilon] = [\partial_x U^\varepsilon]_{x=\xi} \delta_{x=\xi}, \quad \mathcal{P}_2^\varepsilon[W^\varepsilon] \equiv 0. \]

Going further in details

\[ \|\partial_x U^\varepsilon\|_{x=\xi} \leq \frac{u^2}{\varepsilon} (e^{-u^*(\ell+\xi)/\varepsilon} - e^{-u^*(\ell-\xi)/\varepsilon}) + l.o.t. \]

showing that this term is exponentially small for \( \varepsilon \to 0 \) and it is null when \( \xi = 0 \), that corresponds to the equilibrium location of the shock when \( f(u) = u^2/2 \).

In this case, if we neglect the lower order terms, we can write an asymptotic formula for \( \Omega_1^\varepsilon \) and \( \Omega_2^\varepsilon \) that is

\[ \Omega_1^\varepsilon(\xi) \sim \frac{u^2}{\varepsilon} (e^{-u^*(\ell+\xi)/\varepsilon} - e^{-u^*(\ell-\xi)/\varepsilon}), \quad \Omega_2^\varepsilon(\xi) \equiv 0 \]

Finally, by asymptotic computing the functions \( \phi_1^\varepsilon \) and \( \partial_\xi W^\varepsilon \), it is possible to obtain an asymptotic expression for \( \theta^\varepsilon(\xi) \), that is

\[ \theta^\varepsilon(\xi) \sim \frac{u^*}{\varepsilon} (e^{-u^*(\ell+\xi)/\varepsilon} - e^{-u^*(\ell-\xi)/\varepsilon}). \]

This formula shows that the leading order term in the equation for \( \xi(t) \) is exponentially small in \( \varepsilon \), meaning that the speed of the shock layer position goes to zero exponentially as \( \varepsilon \) goes to zero. Moreover, the speed rate of the convergence of \( \xi(t) \) to the equilibrium position \( \xi = 0 \) is given by \( \theta^\varepsilon(0) \), that behaves like \( e^{-1/\varepsilon} \).
Finally, it is easy to check that the hypotheses of Theorem 3.1 and 3.2 are verified in the case of a quadratic flux function.

REFERENCES


E-mail address: strani@dma.ens.fr
ENTROPY FORMULATION FOR FORWARD–BACKWARD
PARABOLIC EQUATIONS

ANDREA TERRACINA
Dipartimento di Matematica “G. Castelnuovo”
Sapienza Università di Roma
Piazzale A. Moro 5, 00185 Roma, Italy

ABSTRACT. We give a brief overview of the results obtained for forward–backward parabolic equations of cubic type. In particular we summarize an approach that has some analogies with hyperbolic conservation laws. More precisely we give a concept of entropy solution and analyze in more details admissibility conditions satisfied by an interface that separates different phases. Finally we state uniqueness and existence results that appear in literature.

1. Introduction. Here we consider the equation

\[ u_t = [\phi(u)]_{xx} \quad \text{in} \quad Q_T := \Omega \times (0,T) \quad (1) \]

where \( T > 0, \Omega \subset \mathbb{R} \) is a bounded interval \( \Omega := (\omega_1, \omega_2), = -\infty \leq \omega_1 < 0 < \omega_2 \leq +\infty \), and \( \phi \) is a piecewise \( C^1 \) function satisfies the following assumptions:

\[
\begin{align*}
\phi'(u) > 0 & \quad \text{if} \quad u \in (-\infty, b) \cup (c, \infty), \\
\phi'(u) < 0 & \quad \text{if} \quad u \in (b, c), \\
B := \phi(b) > \phi(c) & =: A, \quad \phi(u) \to \pm \infty \quad \text{as} \quad u \to \pm \infty, \\
\end{align*}
\]

(H1)

Since the diffusion function \( \phi \) is nonmonotone, equation (1) is of forward-backward parabolic type then, in general, classical problems with initial and boundary conditions are ill-posed, see e.g. [10], [11] for the cubic case.

We denote by

\[ S_1 := \{(u, \phi(u)) \mid u \in (-\infty, b)\} \equiv \{(s_1(v), v) \mid v \in (-\infty, B)\} \quad (2) \]

and

\[ S_2 := \{(u, \phi(u)) \mid u \in [c, \infty)\} \equiv \{(s_2(v), v) \mid v \in [A, \infty)\} \quad (3) \]

the stable branches of the equation \( v = \phi(u) \), whereas

\[ S_0 := \{(u, \phi(u)) \mid u \in [b, c]\} \equiv \{(s_0(v), v) \mid v \in [A, B]\} \quad (4) \]

is referred to as the unstable branch.

Forward-backward parabolic equations arise in many applicative models, in particular equation (1) appears in the theory of phase transitions (see e.g. [4], [5], [8]). The function \( u \) represents the phase field and its values characterize the different phases; for diffusion functions as in (H1), the half-lines \((-\infty, b)\) and \((c, \infty)\) correspond to stable phases and the interval \((b, c)\) to an unstable phase (see e.g. [5]).

2000 Mathematics Subject Classification. Primary: 35K20, 35D99; Secondary: 35K70.

Key words and phrases. Entropy solutions, forward–backward equations, admissibility conditions, pseudoparabolic equations, singular limits.
In analogy with hyperbolic conservation laws we introduce a proper regularization. In fact, the idea is that ill-posedness of forward-backward equations derives from neglecting some relevant information in the modelling of the physical phenomenon. The idea is to obtain an entropy formulation for the original unperturbed problem that characterize the solution as that obtained by the limit of the solutions of the perturbed ones. In this general framework, different types of regularizations for equation (1) have been introduced and investigated (e.g., see [2], [3], [4]). Here, for any given $\varepsilon > 0$, we consider the pseudoparabolic regularization, described by the Sobolev equation

$$u_t = [\phi(u)]_{xx} + \varepsilon u_{xxt},$$

(5)

that takes nonequilibrium effects into account ([4], [8]).

Equation (5) has been studied in [15], while the singular limit as $\varepsilon \to 0^+$ has been analyzed in [16, 18]; see also [19] for other types of nonlinearities. In particular solutions of the Dirichlet or Neumann problems, obtained by the pseudoparabolic regularization, satisfy a class of suitable entropy inequalities (e.g., see [18], [7], [13]). Then [18] gives a notion of entropy solution in the context of the measure–valued solutions.

In this general framework, we consider the Neumann evolution problem

$$\begin{aligned}
\begin{cases}
u_t = [\phi(u)]_{xx} & \text{in } Q_T, \\
[\phi(u)]_x = 0 & \text{on } \partial\Omega \times (0, T), \\
u = u_0 & \text{in } \Omega \times \{0\},
\end{cases}
\end{aligned}$$

(6)

where $\phi$ is as in $(H_1)$, (Neumann condition in the unbounded case has to be understood as a limit condition) and the initial data function $u_0 \in L^\infty(\Omega)$ satisfies

$$\begin{aligned}
\phi(u_0) \in H^{1}_{loc}(\Omega).
\end{aligned}$$

(H2)

In view of the above assumption $(H_2)$, we look for a solution to problem (6) with a particular structure. More precisely, since the initial datum $u_0$ takes values only in the stable phases $S_1$ and $S_2$ of the graph of $\phi$, we consider solutions to problem (6) which describe evolution between such phases. In particular we require that the unstable phase $S_0$ does not influence the dynamics. Moreover, we assume that there exists a regular interface separating the rectangle $Q_T$ into two different regions $V_i$, $i = 1, 2$ and the solution takes values in each stable phase $S_i$ in $V_i$ ($i = 1, 2$). Taking the above requirements into account, it is possible to introduce a notion of solution to (6) that is in accordance with the general entropy formulation given in [18]. We call this kind of solution to problem (6) a two–phase entropy solution, whose formal definition is given below.

**Definition 1.1.** By a two–phase solution to problem (6) we mean a triple $(u,v,\xi)$ such that:

(i) $u \in L^\infty(Q_T)$, $v \in C(\overline{Q}_T) \cap L^2((0, T); H^1_{loc}(\overline{\Omega}))$, and $\xi : [0, T] \to \overline{\Omega}$. $\xi(0) = 0$. Moreover, there exists at most a finite number of disjoint intervals $(\tau', \tau'') \subseteq (0, T)$ such that $\xi'(t) = 0$ for every $t \in (\tau', \tau'')$;

(ii) we have:

$$u = s_i(v) \quad \text{in } V_i \quad (i = 1, 2),$$

(7)
where
\[ V_1 := \{(x,t) \in \overline{Q}_T \mid \omega_1 < x < \xi(t), \ t \in [0,T]\}, \]  
(8)
\[ V_2 := \{(x,t) \in \overline{Q}_T \mid \xi(t) < x < \omega_2, \ t \in [0,T]\}, \]  
(9)
and
\[ \gamma := \partial V_1 \cap \partial V_2 = \{(\xi(t),t) \mid t \in [0,T]\}; \]  
(10)
(iii) for every \( \psi \in C^1(\overline{Q}_T) \), \( \psi(\cdot, T) = 0 \) in \( \overline{\Omega} \), there holds
\[ \int \int_{Q_T} [u \dot{\psi} - v_x \psi_x] dx dt + \int_{\Omega} u_0(x) \psi(x,0) dx = 0; \]  
(11)
(iv) for every \( g \in C^1(\mathbb{R}) \), set
\[ G(\lambda) := \int_0^\lambda g(\phi(s)) ds + k \quad (k \in \mathbb{R}). \]  
(12)
Then, under the assumption \( g' \geq 0 \), the *entropy inequalities*
\[ \int \int_{Q_T} [G(u) \dot{\psi} - g(v) u_x \psi_x - g'(v) v_x^2 \psi] dx dt + \int_{\Omega} G(u_0(x)) \psi(x,0) dx \geq 0 \]  
(13)
hold for every \( \psi \in C^1(\overline{Q}_T) \), \( \psi \geq 0 \) and \( \psi(\cdot, T) = 0 \) in \( \overline{\Omega} \).

We call this *two–phase problem* (TP–problem).

In Section 2 we analyze the structure of the solutions of the TP–problem obtaining a “Rankine–Hugoniot” and an admissibility condition along the interface, here we give also a sketch of the proof of uniqueness proved in [22].

Finally in Section 3 we state the main results obtained for the TP–problem.

2. *Entropy Formulation.* Here we investigate the case in which \( \omega_1, \omega_2 \in \mathbb{R} \), the unbounded case could be handled analogously. Assume the following technical condition for the initial datum \( u_0 \):

\( (H_3) \) the equation \( \phi(u_0)(\cdot) = A \), respectively \( \phi(u_0)(\cdot) = B \), has a finite number of solutions in \( [\omega_1, 0] \), respectively in \( [0, \omega_2] \).

As observed in [13], there is an analogy between the two–phase entropy solutions given in Definition 1.1 and piecewise regular entropy solutions of scalar conservation laws. In some sense, the weak formulation (11) gives a Rankine–Hugoniot condition characterizing the evolution of the interface \( \xi \), whereas the weak entropy inequalities (13) assure local admissibility conditions along the interface. Since the parabolic equation degenerates when \( u = b \) or \( u = c \), in general we can not define the trace of the function \( v_x \) along the boundary of the domains \( V_1 \) and \( V_2 \). To overcome such a difficulty, we will establish a “Rankine–Hugoniot” and a local entropy condition along the interface in a weaker sense. These results could be obtained by the general theory of divergence–measure vector fields.

Let \( \Omega \) an open bounded set in \( \mathbb{R}^n \), we set
\[ DM(\Omega)_2 : \]
\[ = \left\{ h \in (L^2(\Omega))^N : \exists C > 0 : \left| \int_{\Omega} h \nabla \psi dx \right| \leq C \| \psi \|_{L^\infty(\Omega)} \quad \forall \psi \in C^1_c(\Omega) \right\}, \]  
this is the space of \( L^2(\Omega) \) vector fields whose divergence is a bounded Radon measure in \( \Omega \). Let us denote with \( \text{div} \ h \) such measure. Let us follow the outlines of [1] and
(see also [6]). For every $u \in H^1(\Omega) \cap C(\Omega) \cap L^\infty(\Omega)$ we define
\[
[h, u]_{\partial \Omega} := \int_{\Omega} u \, \text{div} \, h + \int_{\Omega} v \, \nabla u \, dx.
\]
Then there hold
\[
[h, u]_{\partial \Omega} = \int_{\partial \Omega} h \cdot \nu \, d\sigma \quad \text{if} \ h \in (C^1(\tilde{\Omega}))^N, \\
[h, u]_{\partial \Omega} = 0 \quad \text{if} \ u \in H^1_0(\Omega), \\
||[h, u]_{\partial \Omega}|| \leq C ||u||_{H^1(\Omega) \cap L^\infty(\Omega)}.
\]
It is possible to define a functional $\Lambda_{\delta}$ on $H^{\frac{1}{2}}(\partial \Omega) \cap L^\infty(\partial \Omega)$ which can be regarded as the normal trace $h \cdot \nu$ at the boundary. In order to determinate a more precise characterization of the boundary term $[h, u]_{\partial \Omega}$ it is necessary to consider a boundary layer sequence. More precisely

**Definition 2.1.** We will call $\{\zeta_\delta\}$ a boundary layer sequence if $\zeta_\delta$ is a sequence of $\text{Lip}(\Omega) \cap C_0(\tilde{\Omega})$ functions such that
\[
\lim_{\delta \to 0^+} \zeta_\delta = 1 \quad \text{pointwise in} \ \Omega, \quad 0 \leq \zeta_\delta \leq 1, \quad \zeta_\delta = 0 \quad \text{on} \ \partial \Omega.
\]
Observe that
\[
\lim_{\delta \to 0^+} \int_{\Omega} \varphi \nabla \zeta_\delta \, dx = -\lim_{\delta \to 0^+} \int_{\Omega} \text{div} \, (\varphi) \zeta_\delta \, dx = -\int_{\Omega} \text{div} \, (\varphi) \, dx = -\int_{\partial \Omega} \varphi \cdot \nu \, d\sigma,
\]
for any $\varphi \in (H^1(\Omega))^N$.

In [12] it is proved the following result

**Lemma 2.2.** Let $h \in DM(\Omega)_2$. Then for every $\xi \in H^1(\Omega) \cap C(\Omega) \cap L^\infty(\Omega)$ we have that
\[
\lim_{\delta \to 0^+} \int_{\Omega} h \nabla \zeta_\delta \xi \, dx = -[h, \xi]_{\partial \Omega},
\]
for any boundary layer sequence $\zeta_\delta$. In particular if $\xi \in H^1_0(\Omega)$ we have that
\[
\lim_{\delta \to 0^+} \int_{\Omega} h \nabla \zeta_\delta \xi \, dx = 0.
\]
Since we want to apply these results in the context of parabolic equations we will use the previous setting of divergence–measure vector fields in a parabolic domain $Q = \Omega \times (0, T)$. We set
\[
DM(Q)_2 := \{(w, h) \in L^2(Q) \times L^2(Q)^N : \\
\exists C > 0 : \left| \int_Q w \psi \, dx + \int_Q h \nabla \psi \, dx \right| \leq C \, ||\psi||_{L^\infty(Q)} \ \forall \psi \in C^\infty_c(Q) \}.
\]
Then we can consider as before $[(w, h), \psi]_{\partial Q}$ and supposing that $w \in C([0, T]; L^1(\Omega))$, we can define the trace on the lateral boundary by
\[
[(w, h), \psi]_{\partial \Omega \times (0, T)} = [(w, h), \psi]_{\partial Q} + \int_{\Omega} w(0) \psi(0) \, dx - \int_{\Omega} w(T) \psi(T) \, dx.
\]
We have the following result proved in [12]

**Lemma 2.3.** Let $(w, v) \in DM(Q)_2$, $w \in C([0, T]; L^1(\Omega))$. Then, for every $\psi \in H^1(Q) \cap L^\infty(Q) \cap C(\Omega)$, such that $\psi(\cdot, 0) \equiv \psi(\cdot, T) \equiv 0$,
\[
\lim_{\delta \to 0^+} \int_Q v \nabla \zeta_\delta \psi \, dx = -[(w, h), \psi]_{\partial \Omega \times (0, T)}
\]
for any boundary layer sequence $\zeta_\delta$ of $\Omega$.  

Let \((w, h)\) be a vector field in \(L^2((\omega_1, \omega_2) \times (0, T))^2\) such that
\[
\int_0^T \int_{\omega_1}^{\omega_2} w\psi_t + h\psi_x \, dx \, dt \geq 0 \quad (\leq 0)
\]  
(17)
for every \(\psi \in C^1_c((\omega_1, \omega_2) \times (0, T)), \psi \geq 0\). Then by a standard application of the Riesz theorem we have that \((w, h) \in DM(A)\) for every open set \(A\) such that \(\overline{A} \subset (\omega_1, \omega_2) \times (0, T)\). Moreover, the bounded Radon measure \(\text{div}(w, h)\) defined in \(A\) has a given sign. Using this results we can obtain informations on the interface \(\gamma\). Let \(A_1 = A \cap V_1\) and \(A_2 = A \cap V_2\), where \(V_1\) and \(V_2\) are the sets defined in (8), consider a boundary layer sequence \(\zeta_\delta\) of the set \(A_1 \cup A_2\), then for any function \(\psi \in C^1_c(A), \psi \geq 0\), define the sequence \(\psi_\delta = (1 - \zeta_\delta)\psi\).

We have for any \(\delta > 0\)
\[
(0 \geq 0) \leq \int_A w\psi_{t\delta} + h\psi_{x\delta} \, dx \, dt = \\
\int_{A_1} w\psi_{t\delta} + h\psi_{x\delta} \, dx \, dt + \int_{A_2} w\psi_{t\delta} + h\psi_{x\delta} \, dx \, dt = \\
[(w, h), \psi]_{\partial A_1} + [(w, h), \psi]_{\partial A_2} - \int_{A_1} \psi \text{div}_{t,x}(w, h) - \int_{A_2} \psi \text{div}_{t,x}(w, h)
\]  
(18)
Using the fact that \(\psi_\delta \equiv \psi\) on \(\partial A_1\) and \(\partial A_2\), for any \(\delta > 0\) and passing to the limit for \(\delta \to 0^+\) in (18) we obtain
\[
[(w, h), \psi]_{\partial A_1} + [(w, h), \psi]_{\partial A_2} \geq 0 \quad (\leq 0).
\]  
(19)
It is clear that if we choose \(A\) such that \(\gamma \cap A \neq \emptyset\) then inequality (19) gives an information for the jump of “the trace” of the vector field on the interface \(\gamma\).

Next, let us consider a two–phase solution \((u, v, \xi)\) of problem (6), hence the corresponding vector field \((w, h) = (u, -v)_x\). In this case, using (11) and choosing properly the boundary layer sequence in (19) we deduce the following “Rankine–Hugoniot” condition.

\[
\lim_{\delta \to 0^+} \frac{1}{\delta} \left[ \int_{t_1}^{t_2} \int_{-\delta}^{0} \{\xi'(t)u(y + \xi(t), t) + v_x(y + \xi(t), t)\} \alpha(t) \, dy \, dt \\
- \int_{t_1}^{t_2} \int_{-\delta}^{0} \{\xi'(t)u(y + \xi(t), t) + v_x(y + \xi(t), t)\} \alpha(t) \, dy \, dt \right] = 0.
\]  
(20)
for any test function \(\alpha \in C^1_c((t_1, t_2))\).

Regarding admissibility conditions for the evolution of the interface \(\xi(t)\), we have to consider the entropy inequalities (13) with the vector field \((w, h) = (G(u), -[F(v)]_x)\); here for every \(g \in C^1(\mathbb{R}), g' \geq 0\), \(G(\lambda)\) is defined in (12) and \(F(\lambda)\) is any primitive of \(g(\lambda)\). In this case, using again (19) and a proper boundary layer sequence, we have that, for any interval \((t_1, t_2) \subset (0, T)\) there holds
\[
\lim_{\delta \to 0^+} \frac{1}{\delta} \left[ \int_{t_1}^{t_2} \int_{-\delta}^{0} \{[F(v)]_x(y + \xi(t), t) + \xi'(t) G(u + \xi(t), t)\} \alpha(t) \, dy \, dt \\
- \frac{1}{\delta} \int_{t_1}^{t_2} \int_{-\delta}^{0} \{[F(v)]_x(y + \xi(t), t) + \xi'(t) G(u + \xi(t), t)\} \alpha(t) \, dy \, dt \right] \leq 0.
\]  
(21)
for any test function \(\alpha \in C^1_c((t_1, t_2)), \alpha \geq 0\).
In some sense we can think that (21) is an entropy condition on the interface $\gamma$. By (21) we can prove the following result, that is analogous to that obtained in [13], [14], [7] when the trace of $v_x$ exists in the classical sense.

**Proposition 1.** Let $(u, v, \xi)$ be a two-phase solution of problem (6). Then

\[
\begin{align*}
\xi'(t) & \leq 0 \quad \text{if } v(\xi(t), t) = B \\
= 0 & \quad \text{if } v(\xi(t), t) \in (A, B) \\
\geq 0 & \quad \text{if } v(\xi(t), t) = A.
\end{align*}
\]

(22)

**Proof.** Sketch of the proof: if $v(\xi(t), t) = B$, consider a nondecreasing function $g_\rho$ such that

\[
g_\rho \left\{ \begin{array}{ll}
= 0 & \text{in } (-\infty, B^* + \rho] \\
> 0 & \text{in } (B^* + \rho, +\infty).
\end{array} \right.
\]

Then, we choose in (21) the following functions

\[
G_\rho(\lambda) = \int_0^\lambda g_\rho(\phi(s)) \, ds, \quad F_\rho(\lambda) = \int_0^\lambda g_\rho(s) \, ds.
\]

then using (21) we obtain $\xi'(t) \geq 0$. The case $v(\xi(t), t) > A$ is similar.
\]

Now we can state the following uniqueness theorem.

**Theorem 2.4.** Suppose that conditions $(H_1) - (H_3)$ are satisfied. Then, there exists at most one two-phase solution of problem (6).

**Proof.** Let us give again a sketch of the proof. Let $(u^1, v^1, \xi^1)$ and $(u^2, v^2, \xi^2)$ be two two-phase solutions of problem (6). Choosing

\[
\psi(x, t) := \int_T^t \{v^1(x, s) - v^2(x, s)\} \, ds,
\]

(23)

by (11) we have

\[
\iint_{Q_T} (u^1 - u^2)(v^1 - v^2) \, dxdt
= \int_T (v^1_x - v^2_x) \left( \int_T^t (v^1_x(x, s) - v^2_x(x, s)) \, ds \right) \, dxdt \leq 0.
\]

(24)

Using the entropy admissibility conditions (1) we obtain that $u_1 \equiv u_2$ (see [22] for details).

3. Existence and qualitative results. Entropy conditions (1) suggest the introduction of two auxiliary problems: a moving boundary problem (MB–problem) and a steady boundary problem (SB–problem). The former appears when the interface evolves in time and it can be regarded as a free boundary problem. The SB–problem arises when the interface does not move, this is a parabolic problem with compatibility conditions along the interface $x = 0$.

Local existence of the TP–problem for piecewise linear functions $\phi$ is investigated in [14]. The unbounded case ($\omega_1 = -\infty, \omega_2 = +\infty$) is considered and existence is obtained using the two auxiliary problems. An improvement of such results is done in [23], where it is shown that the solution to the TP–problem can be extended in time by a suitable alternation of MB–problems and SB–problems. Here it was analyzed also qualitative behaviour of the solutions proving that, even if the
structure could be complex, it is possible to estimate for every $t > 0$ the number of disjoint regions of convexity of the solutions $u(\cdot, t)$ respect to that of initial data.

Nonlinear case was considered in [22], existence was proved using directly the pseudoparabolic regularization (5). The investigation of the singular limit leads to prove local existence also in the nonlinear case.

More precisely for every initial datum $u_0$ satisfying $(H_2)$ and one between

\[
\begin{cases}
A < \{\phi(u_0)\}(0) \leq B, \\
\text{there exists } \delta > 0 \text{ s.t. } \phi(u_0) \leq B \text{ in } (-\delta, \delta),
\end{cases}
\]

or

\[
\begin{cases}
A \leq \{\phi(u_0)\}(0) < B, \\
\text{there exists } \delta > 0 \text{ s.t. } \phi(u_0) \geq A \text{ in } (-\delta, \delta),
\end{cases}
\]

there exists a local (steady–boundary) two–phase solution to problem (6) obtained by taking the vanishing viscosity limit in (5).

Obviously it is necessary to obtain a priori estimate for the regularized problems that allows to prove that at the limit we obtain a two–phase solution. In fact when initial datum satisfies $(A_1)$ or $(A_2)$ the solutions to the pseudoparabolic problems associated with (5) have the same structure of the corresponding two-phase solutions to the original unperturbed problem (6). In this case we obtain a solution of the two–phase problem with a steady interface.

In [22] it was considered also the singular limit for more general initial data giving sufficient conditions which imply the existence of domains where the limiting function $u$ takes values only in one of the two stable phases $S_1$ and $S_2$. Obviously, it would be very interesting to prove that $u$ is a two–phase solution to the TP–problem (6) with a moving interface $\xi$. This is an open problem although the results obtained in [22] go in this direction.

Recently it was studied the case in which the initial data takes values in the unstable region. In [24] it was considered the TP–problem with an unstable phase. Finally in [25] it was proved that uniqueness does not hold in the class of entropy measure-valued solutions introduced by Plotnikov.

REFERENCES


990 ANDREA TERRACINA


E-mail address: terracin@mat.uniroma1.it
STABILITY OF A SUPERSONIC FLOW ABOUT A WEDGE
WITH A WEAK SHOCK WAVE SATISFYING THE
LOPATINSKI CONDITION

Dmitry Tkachev and Aleksander Blokhin
Sobolev Institute of Mathematics
Novosibirsk State University
Novosibirsk, 630090, Russia

Abstract. We are studying the problem of a stationary supersonic flow of
an inviscid non-heat-conducting gas in thermodynamical equilibrium onto a
planar infinite wedge. It is known that theoretically this problem has two
solutions: the solution with a strong shock wave (when the velocity behind the
front of the shock wave is subsonic) and the solution with a weak shock wave
(when, generally speaking, the velocity behind the front of the shock wave is
supersonic). In the present paper, the case of a weak shock wave is studied. It
is proved that if the Lopatinski condition for the shock wave is satisfied (in a
weak sense), then the corresponding linearized initial boundary-value problem
is well-posed, and its classical solution is found. In this case, unlike the case
when the uniform Lopatinski condition holds, additional plane waves appear. It
is shown that for compactly supported initial data the solution of the linearized
problem converges in finite time to the zero solution. Therefore, for the case
of a weak shock wave and when the Lopatinski condition holds in a weak sense
these results complete the verification of the well-known Courant-Friedrichs’
conjecture that the strong shock wave solution is unstable whereas the weak
shock wave solution is stable.

Bibliography: 15 titles.

1. Introduction. It is well-known that the classical problem of a stationary supersonic flow of an inviscid non-heat-conducting gas onto an infinite plane wedge
(with a sufficiently small angle \(\sigma\) at the vertex) has two solutions: one of them corre-

\[\mathbf{U}_0 : \mathbf{u}_0, \mathbf{v}_0\] behind the shock satisfies the inequality \(u_0^2 + v_0^2 < c_0^2\) (\(c_0\) is the
sound speed of the gas behind the shock wave), and the second corresponds to the
case of a weak shock wave, that is \(u_0^2 + v_0^2 > c_0^2\) [1]-[3].

In Figure 1 the angular coordinates \(\theta_s, \theta_w\) determine strong and weak shocks
respectively. The vector of flow velocity \(\mathbf{U}_\infty\) is parallel to the Ox axis.
Paradoxically, but in practice, in physical or numerical experiments, only the
weak shock wave solution is actually realized. One possible explanation was sug-
gested by Courant and Friedrichs [1]. They conjectured that the solution corre-

2000 Mathematics Subject Classification. Primary: 76N15, 35L67, 35B35; Secondary: 35L45, 65C15.

Key words and phrases. Weak shock wave, asymptotic stability (in the sense of Lyapunov).

This research was supported by the Russian Foundation for Basic Research (grant no. 10-01-
Figure 1.

The Courant-Friedrichs’ hypothesis was verified in [4]-[7] (but this conclusion was based only on some qualitative reasons). A strict mathematical justification (and this is very important) of this statement for the linearized problem appeared in recent years in [8]-[12]. Briefly speaking, it was shown in [8]-[12] that in the case of a strong shock wave (for compactly supported initial data!) the perturbation arrives to the wedge’s vertex as time increases having the growth $r^\alpha (\alpha < 0)$ or a logarithmic growth in space variables, and this causes instability of the steady-state solution under consideration.

The situation with the weak shock wave is totally different. In this case the perturbation decays with time. Moreover, this solution is asymptotically stable by Lyapunov. It is assumed that in the both cases the shock satisfies the well-known uniform Lopatinski condition [13].

The present work is a continuation of [11]. We consider the more general case when the Lopatinski condition on the shock wave is satisfied only in a weak sense [13], i.e. the uniform Lopatinski condition can be violated. In particular, this makes finding the classical solution of this problem more difficult. The key point of the work is the analysis of an explicit form of the solution and we extensively use the technique developed in [11].

2. Statement of the original and the auxiliary problems and formulation of main results. The linear problem of finding a supersonic stationary gas flow onto a planar infinite wedge can be stated as follows [14]. We seek for a solution of the system of acoustic equations

$$AU_t + BU_x + C_\sigma U_y = 0,$$

in the domain $t, x > 0, y > x \tan \sigma$ that satisfies the following boundary conditions at the shock wave $(x = 0)$ and on the wedge $(y = x \tan \sigma)$:

$$u_1 + du_3 = 0, \quad u_3 + u_4 = 0, \quad u_2 = \frac{\lambda}{\mu} F_y, \quad F_1 + F_y \tan \sigma = \mu u_3;$$

$$u_2 = u_1 \tan \sigma,$$

and at $t = 0$ it also satisfies the initial data

$$U(0, x, y) = U_0(x, y), \quad F(0, y) = F_0(y).$$

Here $U(t, x, y) = (u_1, u_2, u_3, u_4)^T$: $u_1, u_2, u_3, u_4$ are smooth perturbations of the components of the velocity, the pressure and the entropy respectively; $x = F(t, y)$ is a small displacement of the discontinuity front, and

$$F(t, 0) = F_0(0) = 0.$$
that means that we consider the case of a shock wave attached to the wedge’s vertex. It is also assumed that the components of the vector $U_0(x,y)$ of initial data are compactly supported, i.e.

$$\text{supp } u_{0i} \subset R_+^2 = \{(x,y) | x, y > 0\}, \ i = 1, 2, 3, 4.$$ 

The matrices $A, B, C, C_\sigma$ read as follows:

$$A = \text{diag}(M^2, M^2, 1, 1), \ B = \begin{pmatrix} M^2 & 0 & 1 & 0 \\ 0 & M^2 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \ C = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$C_\sigma = C + \tan \sigma A;$$

$M = \frac{u_0}{c_0}, \ M < 1$ is the Mach number behind of the shock wave ($u_0, v_0 = u_0 \tan \sigma$ are the components of the velocity vector of the stationary solution, $c_0$ is the downstream sound speed), and the physical constants $d, \lambda, \mu$ were described in detail in [14]. These constants depend on the components of the piecewise constant solution corresponding to the step shock as well as on the state equation of gas $p = p(\rho, s)$ ($\rho$ is the density of the gas and $s$ is the entropy) and the equation of the Hugoniot adiabat.

If the solution of problem (1)-(4) is continuous up to the boundary $x = 0, y = x \tan \sigma$, then, in view of (5), it follows from the boundary conditions (2), (3) that the following compatibility condition should be fulfilled at the edge $t \geq 0, x = y = 0$:

$$(\lambda + d \tan^2 \sigma)u_3(t,0,0) = 0, \ t \geq 0,$$

i.e. if $D_1 = \lambda + d \tan^2 \sigma \neq 0$, then

$$U(t,0,0) = 0, \ t \geq 0.$$ 

(6)

**Remark 1.** We have formulated the initial boundary-value problem (1)-(5) for the case when the gas flow in a neighborhood of the wedge with shock wave directed along the $Oy$-axis is fixed as the main solution (Fig. 2).

![Figure 2](image-url)

In this paper we consider the case of a weak shock wave, i.e. the case when

$$M_0 = \sqrt{\frac{u_0^2 + v_0^2}{c_0^2}} = \frac{M}{\cos \sigma} > 1.$$ 

(7)
Further we will also assume that the state equation of gas is such that
\[
\lambda < 0, \ d > -\frac{1}{M}, \ d(1 - M^2) + \lambda M^2 < 0,
\]  
(8)
The coefficient \(D_1 \neq 0\) and, moreover, \(D_1 > 0\) (for example, this is true for the polytropic gas [11]; some examples of state equations satisfying (8) are considered in [15]).

Assume that the solution of problem (1)–(5) is not just continuous but also has second derivatives which are continuous up to the boundary of the domain. Taking mixed derivatives we can reduce problem (1)–(5) with condition (6) to the following initial boundary-value problem for the component \(u_3\) (the pressure). In the domain \(t > 0, \ x > 0, \ y > x \tan \sigma\) we seek for a smooth solution of the wave equation
\[
\left\{ M^2 L_1^2 - L_2^2 - \left( \frac{\partial}{\partial y} \right)^2 \right\} u_3 = 0,
\]
(9)
that satisfies the following boundary conditions at the shock wave \((x = 0)\) and on the wedge \((y = x \tan \sigma)\):
\[
\left\{ mL_1^2 + nL_2^2 - \beta \frac{\partial}{\partial y} \right\} u_3 = 0;
\]
(10)
\[
\left\{ \cos \sigma \left( \frac{\partial}{\partial y} \right) - \sin \sigma \left( \frac{\partial}{\partial x} \right) \right\} u_3 = 0;
\]
(11)
\[ u_3(t, 0, 0) = 0, \]
(12)
and it also satisfies the initial data for \(t = 0\):
\[ u_3|_{t=0} = u_0(x, y), \ (u_3)_{t=0} = u_1(x, y) \]
(13)
(note that the derivative \((u_3)_{t=0}\) can be found from the third equation of system (1)).

In equations (9), (10) we have used the following notations:
\[
L_1 = \frac{1}{\beta} l_1, \quad l_1 = \frac{\partial}{\partial t} + \tan \sigma \frac{\partial}{\partial y}, \quad L_2 = \beta \frac{\partial}{\partial x} - \frac{M^2}{\beta} l_1,
\]
\[ \beta^2 = 1 - M^2, \quad n = -\frac{\lambda}{\beta}, \quad m = \beta d + \frac{\lambda M^2}{\beta}. \]

The converse also holds, i.e. for each solution \(u_3\) of problem (9)-(13) we can uniquely find a corresponding smooth solution \(U(t, x, y), \ F(t, y)\) of problem (1)-(5), (6). This fact can be proved by the same way as for the case of a half-plane \(x > 0\) (see [14]).

Unlike [11], we assume that problem (9)-(13) satisfies the Lopatinski condition [13] at the boundary \(x = 0\) only in a weak sense, i.e. the uniform Lopatinski condition \(m > 0, \ n > 0\) (see [14]) or equivalently the condition
\[ \lambda < 0, \ d > -\frac{\lambda M^2}{\beta^2}. \]
(14)
can be violated. In Fig. 3 we shade the domain where the Lopatinski condition is satisfied (in a weak sense).

This domain is described by the two systems of inequalities
\[
\left\{ \begin{array}{l}
 d > -\frac{1}{M}, \ d < -\frac{\lambda M^2}{\beta^2}, \\
 \lambda < 0
\end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{l}
 d < -\frac{1}{M}, \\
 d > -\frac{\lambda M^2}{\beta^2}
\end{array} \right.
\]
(15)
The straight line \( d = \frac{1}{M^2} + \frac{\rho_\infty \rho_0}{\rho_0} \lambda \) corresponds to the gas dynamic case, and \( \rho_\infty \) and \( \rho_0 \) are the densities ahead and behind of the shock front respectively.

**Remark 2.** For the state equations from [15] the point \((\lambda, d)\) lies in the second quadrant and \( d < -\frac{\lambda M^2}{\beta^2} \).

We point out that the uniform Lopatinski condition holds at the boundary \( y = x \tan \sigma \).

Now we make a convenient change of coordinates by setting
\[
 x' = x, \quad y' = y - x \tan \sigma. \tag{16}
\]
We shall drop the primes in what follows.

Then problem (9)-(13) takes the following form:
\[
 \begin{aligned}
 M^2 \left( \frac{\partial}{\partial t} + \frac{\partial}{\partial x} \right)^2 - \left( \frac{\partial}{\partial x} - \tan \sigma \frac{\partial}{\partial y} \right)^2 - \left( \frac{\partial}{\partial y} \right)^2 \right) u = 0, \quad t, x, y > 0; \tag{17}
 \end{aligned}
\]
\[
 \begin{aligned}
 \left\{ \left( \frac{\partial}{\partial t} + \tan \sigma \frac{\partial}{\partial y} \right) \left[ \frac{\partial}{\partial t} + \frac{\partial}{\partial x} + d \left( \frac{\partial}{\partial t} + \tan \sigma \frac{\partial}{\partial y} \right) \right] - \frac{1}{M^2} \left( \frac{\partial}{\partial x} - \frac{\partial}{\partial y} \right) \right\} u = 0, \quad x = 0; \tag{18}
 \end{aligned}
\]
\[
 \begin{aligned}
 \left( \frac{\partial}{\partial y} - \sin \sigma \cos \sigma \frac{\partial}{\partial x} \right) u = 0, \quad y = 0; \tag{19}
 \end{aligned}
\]
\[
 \begin{aligned}
 u(t, 0, 0) = 0; \tag{20}
 u_{|t=0} = u_0(x, y), \quad u_{|t=0} = u_1(x, y) \tag{21}
 \end{aligned}
\]
(we also drop the subscript 3 by the unknown function \( u_3 \)).

Since problems (1)-(5) and (17)-(21) are equivalent, it is sufficient to state our main results for the solution \( u(t, x, y) \) of problem (17)-(21).

In addition, we also assume that the following property characterizing the behavior of the solution as \( t, x \to +\infty \) holds: there exist parameters \( s_0 \) and \( p_0 \) such that the function \( e^{-s_0 t}e^{-p_0 y}u(t, x, y) \) is bounded as \( t, x \to +\infty \) for each fixed \( y > 0 \), i.e.
\[
 u(t, x, y) = O(e^{s_0 t + p_0 y}), \quad t, x \to +\infty, \quad y > 0 \text{ is fixed.} \tag{22}
\]

Let us introduce the notations \( \Delta = M^2_0 - 1 \), \( B_0 = \frac{\sqrt{\Delta}}{\tan \sigma} \) \((B_0 < 1)\), \( D_1 = \lambda + d \tan^2 \sigma \), \( D_2 = \frac{\sqrt{\Delta}}{M^2} \tan \sigma \), \( L = \frac{1-B_0}{1+B_0} \) and assume that the problem’s parameters are linked by the relation
\[
 |D_2 - D_1| < L. \tag{23}
\]
The following results hold.

**Theorem 2.1.** If the initial data for problem (17)-(21) are compactly supported (i.e. supp \(u_0, u_1 \subset R^2_b\)) and inequality (23) holds, then the classical solution of the problem exists, is unique and can be defined by formula

\[
\begin{align*}
&u(t, x, y) = \frac{8\Delta}{\beta^2M^2} \left( \frac{\partial}{\partial y''} + B_0 \frac{\partial}{\partial x''} \right) \int_0^\infty \left( \frac{\psi'' + y''}{x'' + y''} \right) \bar{g}(t, \xi) \ast_t \\
&\ast_t E \left( t - M^2 \frac{2\Delta}{\sqrt{\Delta}} \bar{y} (y'' - B\xi) - (x'' - \xi) \tan \sigma \right) \, d\xi - \frac{8\Delta}{\beta^2M^2} \int_0^\infty \bar{l}(t, \xi) \ast_t E \left( t - M^2 \frac{2\Delta}{\sqrt{\Delta}} \bar{y} (y'' - (x'' - \xi) \tan \sigma) \right) \, d\xi - \\
&- \frac{8\Delta}{\beta^2M^2} \frac{\partial}{\partial y''} \int_0^\infty \bar{L}(t, \xi) \ast_t E \left( t - M^2 \frac{2\Delta}{\sqrt{\Delta}} \bar{y} (y'' - (x'' - \xi) \tan \sigma) \right) \, d\xi - \\
&\ast_t E \left( t - M^2 \frac{2\Delta}{\sqrt{\Delta}} \bar{y} (y'' - (x'' - \xi) \tan \sigma) \right) \, d\xi + \\
&+ 8 \int_{\square OQM_0 P} (\tan \sigma u_0 \xi + \sqrt{\Delta} u_0 \eta) \times \\
&\times E \left( t - M^2 \frac{2\Delta}{\sqrt{\Delta}} \bar{y} (y'' - \eta) - (x'' - \xi) \tan \sigma \right) \, d\xi d\eta + \\
&+ 2 \frac{\partial}{\partial t} \int_{\square OQM_0 P} u_0 (\xi, \eta) \times \\
&\times E \left( t - M^2 \frac{2\Delta}{\sqrt{\Delta}} \bar{y} (y'' - \eta) - (x'' - \xi) \tan \sigma \right) \, d\xi d\eta + \\
&+ 2 \int_{\square OQM_0 P} u_1 (\xi, \eta) \times \\
&\times E \left( t - M^2 \frac{2\Delta}{\sqrt{\Delta}} \bar{y} (y'' - \eta) - (x'' - \xi) \tan \sigma \right) \, d\xi d\eta; \\
x'' &= 2 \left( y + x \frac{\tan \sigma}{\beta^2} \right), \quad y'' = 2 \frac{\sqrt{\Delta}}{\beta^2} x,
\end{align*}
\]

the first integral is over the line \(y'' = B_0 x''\), and the next two integrals are over the abscissa axis \(y'' = 0\); in the last two integrals over the quadrangle \(\square OQM_0 P\) we have the following coordinates: the point \(Q(x'' - y'', 0)\), the point \(P \left( x'' + y'', \frac{B_0(x'' + y'')}{2} \right)\), and the point \(M_0(x'', y'')\); the function \(E(t, x, y)\) is the fundamental solution of the operator of equation (1.14); \(u_0(x, y)\) and \(u_1(x, y)\) are the initial data (where the coordinates \(x, y\) are expressed through the variables \(x'', y''\)); the functions \(\bar{g}(t, x''), \bar{l}(t, x''), \bar{L}(t, x'')\) are known, in particular, \(\bar{f}(t, x'')\) is determined as follows:

\[
\bar{f}(t, x'') = \sum_{n=0}^{N(t,x'')} L_{p=x''+s-t}^{-1} H_n(p, s),
\]

where \(N(t, x'')\) is a certain integer number, \(N(t, x'') \geq 0\).
Remark 3. We do not write down here formulae defining the functions $\bar{g}(t,x'')$, $\bar{l}(t,x'')$, $\bar{f}(t,x'')$ and $H_n(p,s)$ because of their awkwardness.

Theorem 2.2. For problem (17) - (21) the trace $f(t,x'')|_{x''=2y} = u(t,x,y)|_{x=0}$ at the shock wave is a superposition of a finite number of cylindrical and plane waves. Namely, representation

$$u(t,0,y) = \sum_{n=0}^{N(t,y)} L_{p\to x'',s\to t}^{-1} H_n(p,s)|_{x''=y}.$$  

takes place.

Theorem 2.3. If $y \in K$ ($K$ is a compact subset of the real half-axis), then there exists $t^*(K,\Omega)$ ($\Omega$ is the support of the initial data $u_0$ and $u_1$) such that $f(t,y) \equiv 0$ when $t \geq t^*(K,\Omega)$, $y \in K$.

3. Conclusion. So,

1) we prove on the linearized level that the solution with a weak shock is asymptotically stable (by Lyapunov) in the case when on the shock front the Lopatinski conditions is satisfied in a weak sense, i.e., the uniform Lopatinski condition is violated.

2) Moreover, for compactly supported initial data any solution of the linearized initial boundary value problem becomes stationary for a finite time.

Our result justifies the classical Courant-Friedrichs’ hypothesis that the strong shock wave solution is unstable whereas the weak shock wave solution is stable (when the Lopatinski condition holds at least in a weak sense).

Acknowledgments. The authors are also grateful to A.V. Yegitov for the assistance in preparing the manuscript.

REFERENCES


E-mail address: tkachev@math.nsc.ru
E-mail address: blokhin@math.nsc.ru
EXISTENCE AND STABILITY OF RELATIVISTIC PLASMA-VACUUM INTERFACES

YURI TRAKHININ
Sobolev Institute of Mathematics
Koptyug av. 4, 630090, Novosibirsk, Russia

Abstract. We survey recent results [15, 16] for the plasma-vacuum interface problem in relativistic magnetohydrodynamics for the case when the plasma density does not go to zero continuously, but jumps. In the vacuum region we consider the Maxwell equations for electric and magnetic fields. We show that a sufficiently large vacuum electric field can make the planar interface violently unstable. By using a suitable secondary symmetrization of the vacuum Maxwell equations, we find a sufficient condition that precludes violent instabilities. Under this condition satisfied at each point of the initial nonplanar interface we prove the well-posedness of the free boundary problem in the anisotropic weighted Sobolev spaces $H^{m+}$. 

1. Statement of the interface problem. The equations of relativistic magnetohydrodynamics (RMHD) for an ideal fluid written in its usual covariant form (see, e.g., [5]) can rewritten in the form of the following system of conservation laws:

\begin{align}
&\partial_t (\rho \Gamma) + \text{div} (\rho u) = 0, \\
&\partial_t (\rho h \Gamma u + |H|^2 v - (v, H)H) + \text{div} ((\rho h + B^2)u \otimes u - b \otimes b) + \nabla q = 0, \\
&\partial_t (\rho h \Gamma^2 + |H|^2 - q) + \text{div} (\rho h \Gamma u + |H|^2 v - (v, H)H) = 0, \\
&\partial_t H - \nabla \times (v \times H) = 0,
\end{align}

where $v$ and $H$ are the 3-vectors of velocity and magnetic field, $u = \Gamma v$, $\Gamma = (1 - |v|^2)^{-1/2}$, $\rho$ is the proper rest-mass density, $h = 1 + e + (p/\rho)$, $p$ the pressure, $e = e(\rho, S)$ the specific internal energy, $S$ the entropy, $q = p + \frac{1}{2}B^2$ the total pressure,

$$b = \frac{H}{\Gamma} + (u, H) v, \quad B^2 = \frac{|H|^2}{\Gamma^2} + (v, H)^2 = |H|^2 - |E|^2, \quad \text{with} \quad E = -v \times H.$$ 

Note that we restrict ourselves to the case of special relativity, but in the future we hope to extend our results to general relativity. Moreover, as in the the nonrelativistic case, the equation

$$\text{div} H = 0,$$

2000 Mathematics Subject Classification. Primary: 76W05; Secondary: 35L50, 35L65, 83A05.

Key words and phrases. Relativistic magnetohydrodynamics, plasma-vacuum interface, free boundary problem.

This work was supported by RFBR (Russian Foundation for Basic Research) grant No. 10-01-00320-a.
should be now considered as the divergence constraint on the initial data \( U(0, x) = U_0(x) \) for the unknown \( U = (p, u, H, S) \).

Using (5) and the additional conservation law (entropy conservation)

\[
\partial_t (\rho S) + \text{div} (\rho S u) = 0, \tag{6}
\]

which holds on smooth solutions of system (1)–(4), and following the Godunov’s symmetrization procedure, we can symmetrize the conservation laws (1)–(4) in terms of a vector of canonical variables \( Q = Q(U) \). This was done in [8, 2]. However, due to serious technical difficulties a concrete form of symmetric matrices was not found in [8, 2]. Moreover, if we deal with an initial-boundary value problem it is very inconvenient to work in terms of the vector \( Q \). Serious technical difficulties towards finding a concrete form of symmetric matrices were overcome in [4] by using the Lorentz transform. Namely, the symmetric form

\[
A_0(U) \partial_t U + \sum_{j=1}^{3} A_j(U) \partial_j U = 0, \tag{7}
\]

of the RMHD equations in terms of the primitive (physical) variable \( U \) was found in [4], but here we just refer the reader to [4] for more details and the concrete form of the symmetric matrices \( A_\alpha \).

As is known [2], natural physical restrictions guaranteeing the hyperbolicity of the RMHD system do not depend on the magnetic field and coincide with corresponding ones in relativistic gas dynamics. In our case, by direct calculations one can show that the hyperbolicity condition \( A_0 > 0 \) holds provided that

\[
\rho > 0, \quad p_\rho > 0, \quad 0 < c_s^2 < 1 \tag{8}
\]

(of course, by default we also assume that \( |v| < 1 \)), where \( c_s \) is the relativistic speed of sound, \( c_s^2 = a^2/h = p_\rho/h \). The last inequality in (8) is the relativistic causality condition.

Plasma-vacuum interface problems for the nonrelativistic MHD system usually appear in the mathematical modeling of plasma confinement by magnetic fields. Most of theoretical studies were devoted to finding stability criteria of equilibrium states, and the typical work in this direction is the classical paper [3]. According to our knowledge the first mathematical study of the well-posedness of the full (nonstationary) plasma-vacuum model is [14] and its recent continuations in [10, 11] (see also [7] for the case of incompressible plasma). This model consists in the MHD system in the plasma domain, the elliptic system \( \nabla \times H = 0, \text{div} H = 0 \) (so-called pre-Maxwell dynamics) in the vacuum region, and corresponding interface boundary conditions [3, 14], where by \( H \) we denote the vacuum magnetic field.

The relativistic version of the classical plasma-vacuum interface problem [3, 14] is of great interest in connection with various astrophysical applications such as, for example, neutron stars. Unlike the nonrelativistic case, we cannot now neglect the displacement current \( \partial_t E \) and must consider the full Maxwell equations in vacuum:

\[
\partial_t H + \nabla \times E = 0, \quad \partial_t E - \nabla \times H = 0, \tag{9}
\]

where \( E \) is the electric field in the vacuum and we recall that the speed of the light is equal to unity. We do not include the equations

\[
\text{div} H = 0, \quad \text{div} E = 0 \tag{10}
\]

into the main system (9) because they are just divergence constraints on the initial data \( V(0, x) = V_0(x) \) for the “vacuum” unknown \( V = (H, E) \).
For technical simplicity we assume that the interface (free boundary) between plasma and vacuum has the form of a graph

$$
\Sigma(t) = \{x^1 = \varphi(t, x')\}, \quad \text{with} \quad x' = (x^2, x^3),
$$

that is not a strong restriction for special relativity. Let

$$
\Omega^\pm(t) = \{\pm(x^1 - \varphi(t, x')) > 0\}
$$

be space-time domains occupied by the plasma and the vacuum respectively. That is, in the domain \(\Omega^+(t)\) we consider system (1)–(4) (or (7)) governing the motion of an ideal relativistic plasma and in the domain \(\Omega^-(t)\) we have the Maxwell equations (9) in vacuum. The interface \(\Sigma(t)\) is a free boundary and moves with the velocity of plasma particles at the boundary:

$$
\partial_t \varphi = v_N \quad \text{on} \quad \Sigma(t),
$$

where \(v_N = (v, N)\) and \(N = (1, -\partial_2 \varphi, -\partial_3 \varphi)\). Moreover, the following boundary conditions

$$
q = \frac{|\mathcal{H}|^2 - |E|^2}{2} \quad \text{on} \quad \Sigma(t),
$$

$$
E_2 = \mathcal{H}_3 \partial_1 \varphi - E_1 \partial_2 \varphi, \quad E_3 = -\mathcal{H}_2 \partial_1 \varphi - E_1 \partial_3 \varphi \quad \text{on} \quad \Sigma(t)
$$

should be satisfied for the total pressure and the vacuum electric and magnetic fields (see [15] for a discussion of conditions (12) and (13)). Note also that the conditions

$$
H_N = 0, \quad \mathcal{H}_N = 0 \quad \text{on} \quad \Sigma(t),
$$

with \(\mathcal{H}_N = (\mathcal{H}, N)\) and \(H_N = (H, N)\), should be considered as restrictions on the initial data.

Thus, we have the free boundary value problem for system (7) in \(\Omega^+(t)\) and system (9) in \(\Omega^-(t)\) with the boundary conditions (11)–(13) on \(\Sigma(t)\) and the initial data

$$
U(0, x) = U_0(x), \quad V(0, x) = V_0(x), \quad x \in \Omega^+(0), \quad x \in \Omega^-(0),
$$

$$
\varphi(0, x') = \varphi_0(x'), \quad x \in \mathbb{R}^2.
$$

Moreover, we can prove (see [15]) that (5), (10) and (14) are restrictions on the initial data (15), i.e., if they are satisfied at the first moment \(t = 0\), then they hold for all \(t > 0\). System (9) is always hyperbolic and we assume that the hyperbolicity condition (8) is satisfied up to the boundary of the domain \(\Omega^+(t)\). That is, as in the fluid-vacuum problem in [6, 13], we consider the case when the density does not go to zero continuously, but jumps. However, unlike the problem in [6, 13], we do not have the boundary condition \(\rho|_\Sigma = 0\) and, therefore, the condition \(\rho|_\Sigma > 0\) does not contradict, for example, the equation of state of a polytropic gas. We may suppose that the jump \([\rho] = \rho|_\Sigma\) is very small that corresponds to a small but non-zero pressure on the interface.

Our final goal is to find (at least sufficient) conditions on the initial data (15) that guarantee the local-in-time existence and uniqueness of a smooth solution \((U, V, \varphi)\) of problem (7), (9), (11)–(13), (15) in Sobolev spaces.
2. Secondary symmetrization of the vacuum Maxwell equations. One of the first steps towards the proof of the local-in-time existence and uniqueness theorem is a careful study of a corresponding linearized problem. The main result of [15] is finding a sufficient stability condition which gives a (basic) energy a priori estimate. The crucial role in deriving the basic a priori estimate for the linearized problem is played in [15] by a secondary symmetrization of the Maxwell equations (9).

The Maxwell equations (9) form the symmetric hyperbolic system

$$\partial_t V + \sum_{j=1}^3 B_j \partial_j V = 0, \quad (16)$$

with

$$B_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. $$

At the same time, they imply the wave equation for \( \mathcal{H} \) (and \( E \)) which can be, in turn, symmetrized, i.e., it can be rewritten as a symmetric hyperbolic system for derivatives of \( \mathcal{H} \) (and \( E \)). There are different ways to symmetrize the wave equation, and symmetrizations may depend on arbitrary constants or functions. However, for our goals it is better to use a counterpart of one of these symmetrizations for the original system (16). Since the original system was already symmetric, it is natural to call its new symmetric form secondary symmetrization.

Consider (16) in the whole space \( \mathbb{R}^3 \). Clearly,

$$\frac{d}{dt} \int_{\mathbb{R}^3} |V|^2 \, dx = 0,$$

but, using constraints (10), by simple manipulations from (9) we can deduce the following additional conserved integrals:

$$\frac{d}{dt} \int_{\mathbb{R}^3} (\mathcal{H}_2 E_3 - \mathcal{H}_3 E_2) \, dx = 0,$$

$$\frac{d}{dt} \int_{\mathbb{R}^3} (\mathcal{H}_3 E_1 - \mathcal{H}_1 E_3) \, dx = 0,$$

$$\frac{d}{dt} \int_{\mathbb{R}^3} (\mathcal{H}_1 E_2 - \mathcal{H}_2 E_1) \, dx = 0.$$

Hence, we have

$$\frac{d}{dt} \int_{\mathbb{R}^3} \left\{ |V|^2 + \nu_1 (\mathcal{H}_2 E_3 - \mathcal{H}_3 E_2) + \nu_2 (\mathcal{H}_3 E_1 - \mathcal{H}_1 E_3) + \nu_3 (\mathcal{H}_1 E_2 - \mathcal{H}_2 E_1) \right\} \, dx = 0,$$

where \( \nu_i \) are yet arbitrary constants, but below they will be arbitrary functions satisfying a certain hyperbolicity condition.
The energy identity (17) also reads
\[
\frac{d}{dt} \int_{\mathbb{R}^3} (B_0 V, V) \, dx = 0,
\]
where
\[
B_0 = \begin{pmatrix}
1 & 0 & 0 & 0 & \nu_3 & -\nu_2 \\
0 & 1 & 0 & -\nu_3 & 0 & \nu_1 \\
0 & 0 & 1 & \nu_2 & -\nu_1 & 0 \\
\nu_3 & 0 & -\nu_1 & 0 & 1 & 0 \\
-\nu_2 & \nu_1 & 0 & 0 & 0 & 1
\end{pmatrix}.
\]
We now assume that \( \nu_i \) in \( B_0 \) are arbitrary functions \( \nu_i(t, x) \). Then, taking into account the divergence constraints (10), it follows from (16) that
\[
B_0 \partial_t V + \sum_{j=1}^{3} B_0 B_j \partial_j V + R_1 \text{div} H + R_2 \text{div} E = B_0 \partial_t V + \sum_{j=1}^{3} B_j \partial_j V = 0,
\]
where
\[
B_1 = \begin{pmatrix}
\nu_1 & \nu_2 & \nu_3 & 0 & 0 & 0 \\
\nu_2 & -\nu_1 & 0 & 0 & 0 & -1 \\
\nu_3 & 0 & -\nu_1 & 0 & 1 & 0 \\
0 & 0 & 0 & \nu_1 & \nu_2 & \nu_3 \\
0 & 0 & 1 & \nu_2 & -\nu_1 & 0 \\
0 & -1 & 0 & \nu_3 & 0 & -\nu_1
\end{pmatrix},
\]
\[
B_2 = \begin{pmatrix}
-\nu_2 & \nu_1 & 0 & 0 & 0 & 1 \\
\nu_1 & \nu_2 & \nu_3 & 0 & 0 & 0 \\
0 & \nu_3 & -\nu_2 & -1 & 0 & 0 \\
0 & 0 & -1 & -\nu_2 & \nu_1 & 0 \\
0 & 0 & 0 & \nu_1 & \nu_2 & \nu_3 \\
1 & 0 & 0 & 0 & \nu_3 & -\nu_2
\end{pmatrix},
\]
\[
B_3 = \begin{pmatrix}
-\nu_3 & 0 & \nu_1 & 0 & -1 & 0 \\
0 & -\nu_3 & \nu_2 & 1 & 0 & 0 \\
\nu_1 & \nu_2 & \nu_3 & 0 & 0 & 0 \\
0 & 1 & 0 & -\nu_3 & 0 & \nu_1 \\
-1 & 0 & 0 & 0 & -\nu_3 & \nu_2 \\
0 & 0 & 0 & \nu_1 & \nu_2 & \nu_3
\end{pmatrix},
\]
\[
R_1 = \begin{pmatrix}
\nu_1 \\
\nu_2 \\
\nu_3
\end{pmatrix},
\]
\[
R_2 = \begin{pmatrix}
0 \\
0 \\
\nu_1 \\
\nu_2 \\
\nu_3
\end{pmatrix}.
\]
The symmetric system (18) is hyperbolic if \( B_0 > 0 \), i.e.,
\[
|\nu| < 1,
\]
with the vector-function \( \nu = (\nu_1, \nu_2, \nu_3) \).

3. **Main results and discussion.** The main difficulty in the study of the relativistic plasma-vacuum problem is the same as for compressible current-vortex sheets [13] for the MHD system, i.e., we cannot test the Kreiss-Lopatinski condition analytically. On the other hand, since, as for current-vortex sheets, the number of dimensionless parameters for the constant coefficients linearized problem is big, a complete numerical test of the Kreiss-Lopatinski condition seems unrealizable in practice. We overcome this principal difficulty by using the energy method and the secondary symmetrization (18).
Another principal difficulties are again the same as for current-vortex sheets. First, the Kreiss-Lopatinski condition can be satisfied only in a weak sense and hence we have a loss of derivatives in a priori estimates. Therefore, in the nonlinear analysis we have to use a suitable Nash-Moser-type iteration scheme. Second, the plasma-vacuum interface is a characteristic free boundary. This implies a natural loss of control on derivatives in the normal direction that cannot be compensated in RMHD unlike the situation in gas dynamics [13]. In gas dynamics, one can estimate missing normal derivatives of solutions of the linearized problem through a vorticity-type linearized system and finally get a priori estimates in usual Sobolev spaces. But in our case, exactly as in nonrelativistic compressible MHD [12], the natural functional setting is provided by the anisotropic weighted Sobolev spaces $H^m_*$ (see below and, e.g., [9] for their definition).

Main steps towards the proof of the local-in-time existence and uniqueness theorem for the free boundary problem (7), (9), (11)–(13), (15) are the following:

1) Reduction of the free boundary problem to an initial-boundary value problem in a fixed space domain,
2) Linearization of the reduced problem about a given basic state,
3) Passage to a "good unknown" $(\tilde{U}, \tilde{V})$ (see [1]) for avoiding the appearance of the derivatives of the interface perturbation in the linearized interior equations,
4) Construction of a dissipative energy integral and a basic energy a priori estimate for the prolonged linearized problem for the first-order space-time derivatives of $(\tilde{U}, \tilde{V})$ by using the secondary symmetrization (18) for a suitable choice of the vector-function $\nu$ (for this choice the hyperbolicity condition (19) gives us a sufficient stability condition for planar interfaces),
5) Derivation of a tame estimate in higher order Sobolev spaces for the linearized problem (an estimate "linear in high norms", see [1] and, e.g., [13]),
6) The proof of the existence of solutions of the nonlinear problem by Nash-Moser iterations whose convergence is justified by using the tame estimate,
7) Uniqueness follows from the basic a priori estimate for the linearized problem.

Moreover, by the spectral method (by constructing Hadamard-type ill-posedness examples) for particular cases it was shown in [15] that a sufficiently large unperturbed vacuum electric field can make the planar interface violently unstable.

The reduction of the free boundary problem to that in the the fixed space domain $\mathbb{R}^3_+ = \{x^1 > 0, x^\prime \in \mathbb{R}^2\}$ was made by the simple change of variables $\tilde{x}^1 = x^1 - \varphi(t, x^\prime)$ for the RMHD system (7) and $\tilde{x}^1 = -x^1 + \varphi(t, x^\prime)$ for the Maxwell equations (9). After that the reduced problem with the straightened boundary $x^1 = 0$ (we drop the tilde) is linearized with respect to a sufficiently smooth basic state $(\tilde{U}, \tilde{V}, \tilde{\varphi})$.

The main assumptions on the basic state are the following:

1) It satisfies the hyperbolicity conditions (8),
2) It satisfies the boundary conditions (11)–(13),
3) Either the plasma expands into the vacuum ($\partial_t \tilde{\varphi} < 0$) or alternatively the vacuum expands into the plasma ($\partial_t \tilde{\varphi} > 0$).

The mixed case when for some portions of the unperturbed interface $\partial_t \tilde{\varphi} < 0$ and for other ones $\partial_t \tilde{\varphi} > 0$ corresponds to a hyperbolic problem with characteristic boundary of variable multiplicity and we do not yet study this difficult problem.

Before formulating the main result for the linearized problem obtained in [15] we give the definition of the anisotropic weighted Sobolev spaces $H^m_*$. The functional
space $H^m_\sigma$ is defined as follows:

$$H^m_\sigma(\mathbb{R}^3_+) = \{ u \in L_2(\mathbb{R}^3_+) \mid \partial^\sigma \partial_t^k u \in L_2(\mathbb{R}^3_+) \text{ if } |\sigma| + 2k \leq m \},$$

where $m \in \mathbb{N}$, $\partial^\sigma = (\sigma \partial_1)^{\alpha_1} \partial_2^{\alpha_2} \partial_3^{\alpha_3}$, and $\sigma(x^1) \in C^\infty(\mathbb{R}_+)$ is a monotone increasing function such that $\sigma(x^1) = x^1$ in a neighborhood of the origin and $\sigma(x^1) = 1$ for $x^1$ large enough. The space $H^m_\sigma(\mathbb{R}^3_+)$ is normed by

$$\| u \|_{m,\sigma}^2 = \sum_{|\sigma| + 2k \leq m} \| \partial^\sigma \partial_t^k u \|_{L_2(\mathbb{R}^3_+)}^2.$$  

We also define the space

$$H^m_{\sigma}(\Omega_T) = \bigcap_{k=0}^m H^k((-\infty, T], H^{m-k}_\sigma(\mathbb{R}^3_+))$$

equipped with the norm

$$[u]_{m,\sigma,T}^2 = \int_{-\infty}^T \| u(t) \|_{m,\sigma}^2 dt, \text{ where } \| u(t) \|_{m,\sigma}^2 = \sum_{j=0}^m \| \partial_t^j u(t) \|_{m-j,\sigma}^2.$$  

**Theorem 3.1** (Basic a priori estimate for the linearized problem). Let the basic state $(\hat{U}, \hat{V}, \hat{\varphi})$ satisfies all the assumptions above. Let also

$$|\hat{H}_2 \hat{H}_3 - \hat{H}_3 \hat{H}_2|_{x^1=0} \geq \epsilon > 0,$$

where $\epsilon$ is a fixed constant. Then there exists a positive constant $\mu^*$ such that for all the source terms $(f_1, f_{11}, g) \in H^3(\Omega_T) \times H^3(\Omega_T) \times H^3(\partial \Omega_T)$ which vanish in the past the linearized problem with the zero initial data has a unique solution $(\hat{U}, \hat{V}, \hat{\varphi}) \in H^3(\Omega_T) \times H^1(\Omega_T) \times H^1(\partial \Omega_T)$ for all

$$\mu < \mu^*,$$

where

$$\mu = |\hat{E}_1 + \hat{v}_2 \hat{H}_3 - \hat{v}_3 \hat{H}_2|_{x^1=0}.$$

Here $\Omega_T = (-\infty, T] \times \mathbb{R}^3_+$ and $f_1$, $f_{11}$ and $g$ are given source terms in the linearized RMHD equations, the linearized Maxwell equations (after straightening the interface they become nonlinear) and the linearized boundary conditions respectively. Moreover, this solution obeys the a priori estimate

$$[\hat{U}]_{1,\sigma,T} + \| \hat{V} \|_{H^1(\Omega_T)} + \| \hat{\varphi} \|_{H^1(\partial \Omega_T)} \leq C \{ [f_1]_{3,\sigma,T} + \| f_{11} \|_{H^3(\Omega_T)} + \| g \|_{H^3(\partial \Omega_T)} \},$$

where $C = C(T) > 0$ is a constant independent of the data $(f_1, f_{11}, g)$.

Starting from the basic a priori estimate (22) we can also derive the mentioned tame estimate in $H^m_\sigma$. Such an estimate is crucial for proving the convergence of Nash-Moser iterations for the nonlinear problem. We do not present here this estimate and just announce the final result from [16] which is the local-in-time existence and uniqueness theorem for the free boundary problem (7), (9), (11)–(13), (15).

**Theorem 3.2** (Local-in-time existence and uniqueness of a smooth stable (nonplanar) interface). Let $m \geq 12$ and the initial data

$$(U_0 - \hat{U}, V_0 - \hat{V}, \varphi_0) \in H^{2m+19}_\sigma(\mathbb{R}^3_+) \times H^{2m+19}(\mathbb{R}^3_+) \times H^{2m+19}(\mathbb{R}^2),$$
where \( \bar{U} = (\bar{p}, 0, \bar{H}, 0) \) and \( \bar{V} = (\bar{\mathcal{H}}, \bar{E}) \) are constant vectors, satisfy compatibility conditions of order \( m + 9 \) (in a suitable sense [16]) and the same requirements as the basic state in the linearized problem (in particular, the sufficient stability condition (20), (21) for all \( x' \in \mathbb{R}^2 \)). Then, there exists a sufficiently short time \( T > 0 \) such that the free boundary problem (7), (9), (11)–(13), (15) (reduced to that in the fixed domain \( \mathbb{R}^3_+ \)) has a unique solution

\[
(U - \bar{U}, V - \bar{V}, \varphi) \in H^m([0, T] \times \mathbb{R}^3_+) \times H^m([0, T] \times \mathbb{R}^3_+) \times H^m([0, T] \times \mathbb{R}^2).
\]

The Nash-Moser iteration scheme used to prove the existence of solutions is very close to that from [12].

REFERENCES


E-mail address: trakhin@math.nsc.ru
THE LINEARIZED PLASMA-VACUUM INTERFACE PROBLEM
IN IDEAL INCOMPRESSIBLE MHD

ALESSANDRO MORANDO AND PAOLA TREBESCHI
DICATAM, Sezione di Matematica
Università di Brescia
Via Valotti 9, 25133 Brescia, Italy

YURI TRAKHININ
Sobolev Institute of Mathematics
Koptyug av. 4, 630090, Novosibirsk, Russia

Abstract. We study the free boundary problem for the plasma-vacuum interface in ideal incompressible magnetohydrodynamics (MHD). In the vacuum region we consider the system of pre-Maxwell dynamics, while at the interface the total pressure is continuous and the magnetic field is tangent to the boundary. Under a suitable stability condition satisfied at the plasma-vacuum interface, we prove the well-posedness of the linearized problem in Sobolev spaces.

1. Introduction. We consider the equations of ideal incompressible magnetohydrodynamics (MHD) governing the motion of a perfectly conducting inviscid incompressible plasma. In the case of homogeneous plasma (the density $\rho(t,x) \equiv \text{const} > 0$) these equations in a dimensionless form are

\begin{align}
\partial_t v + (v, \nabla) v - (H, \nabla) H + \nabla q &= 0, \tag{1a} \\
\partial_t H + (v, \nabla) H - (H, \nabla) v &= 0, \tag{1b} \\
\text{div} \, v &= 0, \tag{1c}
\end{align}

where $v = v(t,x) = (v_1, v_2, v_3)$ denotes the plasma velocity, $H = H(t,x) = (H_1, H_2, H_3)$ the magnetic field (in Alfvén velocity units), $q = p + |H|^2/2$ the total pressure, and $p = p(t,x)$ the pressure (divided by $\rho$). As the unknown we fix the vector $U = (q, W)$ with $W = (v, H)$. System (1) is supplemented by the divergence constraint, on the initial data $W|_{t=0} = W_0$,

\begin{equation}
\text{div} \, H = 0. \tag{2}
\end{equation}

In the classical plasma-vacuum interface problem the plasma is confined inside a perfectly conducting rigid wall and isolated from it by a vacuum region. The linearized plasma-vacuum problem in ideal compressible MHD was studied in [6, 9], and the well-posedness of the original nonlinear free boundary problem was recently proved in [7] by the Nash-Moser method. Our main goal is to obtain an analogous result for the plasma-vacuum interface problem for the model of incompressible MHD which can be used when the characteristic plasma velocity is very small compared to the speed of sound. In this paper we concentrate on the corresponding

2000 Mathematics Subject Classification. Primary: 76W05; Secondary: 35Q35, 35L50, 76E17. 

Key words and phrases. Ideal incompressible Magneto-Hydrodynamics, plasma-vacuum interface, free boundary.
linearized problem. For our plasma-vacuum problem we consider the case of an unboumed plasma domain and we neglect the influence of gravity because it just contributes with a lower-order term in (1a).

Let $\Omega^+(t)$ and $\Omega^-(t)$ be space-time domains occupied by the plasma and the vacuum respectively. In $\Omega^+(t)$ we consider system (1) and in $\Omega^-(t)$ we have the elliptic (div-curl) system

$$\nabla \times H = 0, \quad \text{div } H = 0,$$

(3)
describing the vacuum magnetic field $H = H(t, x) = (H_1, H_2, H_3) \in \mathbb{R}^3$. Here, we consider so-called *pre-Maxwell dynamics* (usual in nonrelativistic MHD). The boundary $\Gamma(t)$ of $\Omega^+\Omega^+$ is the interface between plasma and vacuum. It is to be determined and moves with the velocity of plasma particles at the boundary. We assume that $\Omega^\pm(t)$ are unbounded and $\Gamma(t)$ has the form of a graph: $x_1 = \varphi(t, x')$, $x' = (x_2, x_3)$; that is $\Omega^\pm(t) = \{ \pm(x_1 - \varphi(t, x')) > 0 \}$. The function $\varphi(t, x')$ is to be determined and must satisfy the transport equation (where $N = (1, -\partial_2 \varphi, -\partial_3 \varphi)$)

$$\partial_t \varphi = (v, N) \quad \text{on } \Gamma(t).$$

(4)

The plasma variable $U$ and the vacuum magnetic field $H$ are connected through

$$[q] = 0, \quad (H, N) = 0 \quad (H, N) = 0, \quad \text{on } \Gamma(t),$$

(5)

where $[q] = q|_\Gamma - \frac{1}{2}|H|^2|_\Gamma$ denotes the jump of the total pressure across the interface. These relations together with (4) are the boundary conditions at the interface $\Gamma(t)$.

A natural wish is to find conditions on the initial data

$$W(0, x) = W_0(x), \quad x \in \Omega^+(0), \quad \varphi(0, x') = \varphi_0(x'), \quad x' \in \mathbb{R}^2,$$

(6)

$$\mathcal{H}(0, x) = \mathcal{H}_0(x), \quad x \in \Omega^-(0),$$

(7)

providing the local-in-time existence and uniqueness of a solution $(U, \mathcal{H}, \varphi)$ of problem (1), (3)–(7) in Sobolev spaces.

We first reduce the free boundary problem (1)–(5), (6), (7) to that in a fixed domain by a suitable straightening of the unknown interface; then we linearize the resulting problem around a basic state (“unperturbed flow”). Under a suitable stability condition satisfied at each point of the unperturbed interface, we prove the well-posedness of the linearized problem in the Sobolev space $H^3$.

2. **Reduction to a fixed domain.** We straighten the interface $\Gamma(t)$ by using the same change of independent variables as in [6], that is inspired by Lannes [3] (see also [2]). We set $\Omega^\pm := \mathbb{R}^3 \cap \{ \pm x_1 > 0 \}$ and $\Gamma := \mathbb{R}^3 \cap \{ x_1 = 0 \}$. We reduce the free boundary problem (1)–(5), (6), (7) to the fixed domains $\Omega^\pm$, by constructing a global diffeomorphism of $\mathbb{R}^3$, mapping $\Omega^\pm(t)$ onto $\Omega^\pm$ and $\Gamma(t)$ onto $\Gamma$ at each time $t \in [0, T]$. The diffeomorphism is given by the following lemma (cf. [5, Lemma 2.1]).

**Lemma 2.1.** Let $m \geq 3$ be an integer. For all $T > 0$ and for all $\varphi \in \cap_{j=0}^{m-1} C^j([0, T]; H^{m-j-0.5} (\mathbb{R}^2))$ satisfying without loss of generality $\| \varphi \|_{C([0, T]; H^2(\mathbb{R}^2))} \leq 1$, there exists a function $\Psi \in \cap_{j=0}^{m-1} C^j([0, T]; H^{m-j}(\mathbb{R}^3))$ such that

$$\Phi(t, x) := (x_1 + \Psi(t, x), x') = (\Phi_1(t, x), x'), \quad (t, x) \in [0, T] \times \mathbb{R}^3$$

(8)
is a $H^m$–diffeomorphism of $\mathbb{R}^3$ for all $t \in [0, T]$. Moreover, for $j = 0, \ldots, m-1$, $\partial^j_t (\Phi - Id) \in C([0, T]; H^{m-j}(\mathbb{R}^3))$, $\Phi(t, 0, x') = (\varphi(t, x'), x')$, $\partial_3 \Phi(t, 0, x') = (1, 0, 0)$. 
We introduce the change of unknown functions induced by (8), by setting
\[ \bar{U}(t, x) := U(t, \Phi(t, x)), \quad \bar{H}(t, x) := H(t, \Phi(t, x)). \]
The functions \( \bar{U} = (\bar{q}, \bar{v}, \bar{H}) \) and \( \bar{H} \) are smooth in the half-spaces \( \Omega^+ \) and \( \Omega^- \) respectively. Dropping the tildes for convenience, the problem (1)-(5), (6), (7) can be restated in \( \Omega^\pm \) as the following initial-boundary value problem
\[
\begin{align*}
\mathbb{P}(U, \Psi) &= 0, \quad \text{in } [0, T] \times \Omega^+, \quad (9a) \\
\mathbb{V}(H, \Psi) &= 0, \quad \text{in } [0, T] \times \Omega^-, \quad (9b) \\
\mathbb{B}(U, H, \varphi) &= 0, \quad \text{on } [0, T] \times \Gamma, \quad (9c)
\end{align*}
\]
where \( \mathbb{P}(U, \Psi), \mathbb{V}(H, \Psi), \mathbb{B}(U, H, \varphi) \) are the operators defined as follows.

2.1. Operator \( \mathbb{P}(U, \Psi) \).
\[
\mathbb{P}(U, \Psi) := \left( \frac{L(U, \Psi)}{\text{div} u} \right),
\]
where \( u := (v_n, v_2 \partial_1 \Phi_1, v_3 \partial_1 \Phi_1), \quad v_n := (v, n) := (1, -\partial_2 \Psi, -\partial_3 \Psi), \) with \( L(U, \Psi) := L(W, \Psi)U = L_1(W, \Psi)W + \left( \begin{array}{c} \nabla \phi q \\ 0 \end{array} \right), \) where
\[
L_1(W, \Psi) := \partial_1 + \bar{A}_1(W, \Psi) \partial_1 + A_2(W) \partial_2 + A_3(W) \partial_3
\]
\[\bar{A}_1(W, \Psi) := \frac{1}{\partial_1 \Phi_1} \left( A_1(W) - \sum_{k=2}^3 A_k(W) \partial_k \Psi - I_6 \partial_1 \Psi \right), \]
\[A_k(W) := \begin{pmatrix} v_k I_3 & -H_k I_3 \\ -H_k I_3 & v_k I_3 \end{pmatrix}, \]
\[\nabla \phi q := \begin{pmatrix} \partial_1 q \\ \partial_1 \Phi_1 \end{pmatrix}, \quad \frac{\partial_2 \Psi}{\partial_1 \Phi_1} \partial_1 q + \partial_2 q, \quad -\frac{\partial_3 \Psi}{\partial_1 \Phi_1} \partial_1 q + \partial_3 q \right). \]

2.2. Operator \( \mathbb{V}(H, \Psi) \).
\[
\mathbb{V}(H, \Psi) := \left( \begin{array}{c} \nabla \times \delta \Phi \\ \text{div} \delta \Phi \end{array} \right)
\]
where \( \delta \Phi = (H_1 \partial_1 \Phi_1, H_{r2}, H_{r3}), \quad \delta \Phi = (H_n, H_2 \partial_1 \Phi_1, H_3 \partial_1 \Phi_1), \quad H_{r_k} = H_1 \partial_k \Psi + H_k, \quad k = 2, 3, \quad H_n = (H, n). \)

2.3. Operator \( \mathbb{B}(U, H, \varphi) \).
\[
\mathbb{B}(U, H, \varphi) := \left( \begin{array}{c} \partial_1 \varphi - v_N \\ [q] \\ H_N \end{array} \right)
\]
where \( [q] := q_{\Gamma} - \frac{1}{2} |H|^2_{E}, \quad v_N := (v, N), \quad H_N := (H, N), \quad N := (1, -\partial_2 \varphi, -\partial_3 \varphi). \)

We also did not include in our problem the equations
\[\text{div} h = 0 \quad \text{in } [0, T] \times \Omega^+, \quad H_N = 0 \quad \text{on } [0, T] \times \Gamma \]
because they are just restrictions on the initial data (9d) (see [8], [4] for details).
3. Linearized problem.

3.1. Basic state. For $T > 0$, let us set $Q^\pm_T := (-\infty, T] \times \Omega^\pm$ and $\omega_T := (-\infty, T] \times \Gamma$. Let

$$\begin{align*}
(\hat{U}(t, x), \hat{H}(t, x), \hat{\varphi}(t, x')) & \quad (11)
\end{align*}$$

be a given sufficiently smooth vector-function, respectively defined on $Q^+_T$, $Q^-_T$, $\omega_T$, with $\hat{U} = (\hat{q}, \hat{v}, \hat{H})$, such that

$$\begin{align*}
\|\hat{U}\|_{W^{2,\infty}(Q^+_T)} + \|\partial_1 \hat{U}\|_{W^{2,\infty}(Q^+_T)} + \|\hat{H}\|_{W^{2,\infty}(Q^-_T)} + \|\hat{\varphi}\|_{W^{3,\infty}(\partial Q^+_T \times \mathbb{R}^2)} & \leq K, \\
\|\hat{\varphi}\|_{C([0,T]; H^2(\mathbb{R}^2))} & \leq 1,
\end{align*}$$

(12)

where $K > 0$ is a constant. Corresponding to $\hat{\varphi}$, let the function $\hat{\Psi}$ and the diffeomorphism $\hat{\Phi}$ be constructed as in Lemma (2.1) such that $\partial_1 \hat{\Phi} \geq 1/2$. We assume that the basic state (11) satisfies

$$\begin{align*}
\partial_t \hat{H} + \frac{1}{\partial_1 \hat{\Phi}_1} \left\{ (\hat{w}, \nabla) \hat{H} - (\hat{h}, \nabla) \hat{v} \right\} & = 0, \quad \text{div} \hat{u} = 0 \quad \text{in} \ Q^+_T, \\
\text{div} \hat{h} & = 0 \quad \text{in} \ Q^-_T, \\
\partial_t \hat{\varphi} - \hat{v}_N & = 0, \quad [\hat{q}] = 0, \quad \hat{H}_N = 0 \quad \text{on} \ \omega_T,
\end{align*}$$

(13a)

(13b)

(13c)

where all the “hat” values are determined like corresponding values for $(U, H, \varphi)$. It follows from (13a) that the constraints

$$\begin{align*}
\text{div} \hat{h} & = 0 \quad \text{in} \ Q^+_T, \\
\hat{H}_N & = 0 \quad \text{on} \ \omega_T
\end{align*}$$

(14)

are satisfied for the basic state (11), if they hold at $t = 0$. Thus, for the basic state we also require the fulfillment of conditions (14) at $t = 0$.

3.2. Linearized problem. The linearized equations for (9a)-(9c) read:

$$\begin{align*}
P'(\hat{U}, \hat{\Psi}) & \left( \delta U, \delta \Psi \right) := \frac{d}{d\varepsilon} P(U_\varepsilon, \Psi_\varepsilon)|_{\varepsilon = 0} = f \quad \text{in} \ Q^+_T, \\
V'(\hat{H}, \hat{\Psi}) & \left( \delta H, \delta \Psi \right) := \frac{d}{d\varepsilon} V(H_\varepsilon, \Psi_\varepsilon)|_{\varepsilon = 0} = F \quad \text{in} \ Q^-_T, \\
B'(\hat{U}, \hat{\Psi}) & \left( \delta U, \delta \varphi \right) := \frac{d}{d\varepsilon} B(U_\varepsilon, H_\varepsilon, \varphi_\varepsilon)|_{\varepsilon = 0} = g \quad \text{on} \ \omega_T,
\end{align*}$$

where $U_\varepsilon = \hat{U} + \varepsilon \delta U, \ H_\varepsilon = \hat{H} + \varepsilon \delta H, \ \varphi_\varepsilon = \hat{\varphi} + \varepsilon \delta \varphi; \ \delta \Psi$ is constructed from $\delta \varphi$ as in Lemma (2.1) and $\Psi_\varepsilon = \hat{\Psi} + \varepsilon \delta \Psi$. The source terms $f, F$ and $g$ appear to make the interior equations and the boundary conditions inhomogeneous.

We compute the exact form of the linearized equations (below we drop $\delta$):

$$\begin{align*}
P'(\hat{U}, \hat{\Psi})(U, \Psi) & = \left( L(\hat{W}, \hat{\Psi}) \right) U - \left\{ \begin{array}{c} \{ L(\hat{W}, \hat{\Psi}) \Psi \} \\
\partial_1 \hat{U} \end{array} \right\} \frac{\partial_1 \hat{U}}{\partial_1 \hat{\Phi}_1} = f, \\
V'(\hat{H}, \hat{\Psi})(H, \Psi) & = \nabla H \times \left( \begin{array}{c} \nabla \hat{H}_1 \\
\nabla \hat{H}_2 \end{array} \right) \cdot \nabla \Psi = F, \\
\end{align*}$$

(15)

(16)
In order to cancel out the first-order operators in \( \Psi \) from (15), (16) under the change of variables (18) and consider the effective linear operators 

\[
\begin{align*}
\mathbb{E}'(\hat{U}, \hat{\mathcal{H}}, \hat{\varphi})(U, \mathcal{H}, \varphi) &= \begin{pmatrix} \partial_t \varphi + \hat{v}_2 \partial_2 \varphi + \hat{v}_3 \partial_3 \varphi - v_N \hat{\varphi} \\ q - (\hat{\mathcal{H}}, \mathcal{H}) \\ \mathcal{H}_N - \hat{\mathcal{H}}_2 \partial_2 \varphi - \hat{\mathcal{H}}_3 \partial_3 \varphi \end{pmatrix} = g,
\end{align*}
\]

where \( L(\hat{W}, \hat{\Psi})U = L_1(\hat{W}, \hat{\Psi})W + \begin{pmatrix} \nabla \cdot q \\ \mathcal{H}_N - \hat{\mathcal{H}}_2 \partial_2 \varphi - \hat{\mathcal{H}}_3 \partial_3 \varphi \end{pmatrix} \)

\( \mathfrak{V}(\hat{W}, \hat{\Psi}) \) being the differential operator defined in (10) (with \( (W, \Psi) = (\hat{W}, \hat{\Psi}) \)), and

\[
C(\hat{W}, \hat{\Psi})W = \begin{pmatrix} C_1(\hat{W}, \hat{\Psi})W \\ C_2(\hat{W}, \hat{\Psi})W \end{pmatrix} = \frac{1}{\partial_1 \hat{\Phi}_1} \begin{pmatrix} (u, \nabla) \hat{\varphi} - (h, \nabla) \hat{\mathcal{H}} \\ (u, \nabla) \mathcal{H} - (h, \nabla) \hat{\varphi} \end{pmatrix}.
\]

In order to cancel out the first-order operators in \( \Psi \) from \( \mathbb{E}'(\hat{U}, \hat{\mathcal{H}}, \hat{\varphi})(U, \mathcal{H}, \varphi) \) and \( \mathfrak{V}'(\hat{\mathcal{H}}, \hat{\mathcal{H}}, \hat{\varphi}) \), as in [1], the linearized problem is rewritten in terms of the “good unknown”

\[
\hat{U} := U - \frac{\Psi}{\partial_1 \hat{\Phi}_1} \partial_1 \hat{U}, \quad \hat{\mathcal{H}} := \mathcal{H} - \frac{\Psi}{\partial_1 \hat{\Phi}_1} \partial_1 \hat{\mathcal{H}}.
\]

As in [1, 8], we drop the zeroth-order terms in \( \Psi \) appearing into (15), (16) under the change of variables (18) and consider the effective linear operators

\[
\mathbb{E}'_e(\hat{U}, \hat{\mathcal{H}}, \hat{\varphi})U := \begin{pmatrix} \mathbb{L}'_e(\hat{U}, \hat{\Psi})U \\ \nabla \cdot \hat{u} \end{pmatrix} = f,
\]

\[
\mathfrak{V}(\hat{\mathcal{H}}, \hat{\varphi}) = \begin{pmatrix} \nabla \times \hat{\mathcal{H}} \\ \nabla \cdot \hat{\mathcal{H}} \end{pmatrix} = \mathcal{F},
\]

where

\[
\mathbb{L}'_e(\hat{U}, \hat{\Psi})U = L(\hat{W}, \hat{\Psi})U = L_1(\hat{W}, \hat{\Psi})W + \begin{pmatrix} \nabla \cdot \hat{q} \\ 0 \end{pmatrix} + C(\hat{W}, \hat{\Psi})W
\]

and all the “dot”- values in (19), (20) are determined like the corresponding values for \( (U, \mathcal{H}) \), with \( \varphi, \Psi \) replaced by \( \hat{\varphi}, \hat{\Psi} \). After (18), the boundary conditions become

\[
\mathbb{E}'(\hat{U}, \hat{\mathcal{H}}, \hat{\varphi})(U, \mathcal{H}, \varphi) := \begin{pmatrix} \partial_t \varphi + \hat{v}_2 \partial_2 \varphi + \hat{v}_3 \partial_3 \varphi - v_N \hat{\varphi} - \hat{\varphi} \partial_1 \hat{v}_N \\ \hat{q} - (\hat{\mathcal{H}}, \mathcal{H}) + [\partial_1 \hat{\varphi}] \varphi \\ \mathcal{H}_N - \partial_2 (\hat{\mathcal{H}}_2 \varphi) - \partial_3 (\hat{\mathcal{H}}_3 \varphi) \end{pmatrix} = g, \text{ on } \omega_T.
\]

We consider the linearized problem (19), (20), (22). The source terms in (20) have to satisfy suitable compatibility conditions, see [5]. We assume that the source terms \( (f, \mathcal{F}) \) and the boundary data \( g \) vanish in the past and consider the case of zero initial data. We reduce problem (19), (20), (22) to one for which most of the data are homogeneous [5]:

\[
\mathbb{L}'_e(\hat{U}, \hat{\Psi})U = \begin{pmatrix} f_u \\ 0 \end{pmatrix},
\]

\[
\text{div } u = 0, \quad \text{in } Q^+_T,
\]

\[
\mathfrak{V}(\hat{\mathcal{H}}, \hat{\varphi}) = 0, \quad \text{in } Q^+_T,
\]

\[
\mathbb{E}'(\hat{U}, \hat{\mathcal{H}}, \hat{\varphi})(U, \mathcal{H}, \varphi) = 0, \quad \text{on } \omega_T.
\]
Solutions to problem (23) also satisfy
\[ \text{div } h = 0 \text{ in } Q_T^+, \quad H_\omega = \tilde{H}_2 \partial_2 \varphi + \tilde{H}_3 \partial_3 \varphi - \varphi \partial_1 \tilde{H}_N \text{ on } \omega_T. \] (24)

4. Function Spaces. For \( \gamma \geq 1 \) and \( s \in \mathbb{R} \), we set \( \lambda^{s,\gamma}(\xi) := (\gamma^2 + |\xi|^2)^{s/2} \).

For \( \gamma \geq 1 \), \( H_s^s(\mathbb{R}^n) \) will denote the Sobolev space of order \( s \), equipped with the \( \gamma \)-depending norm \( \| \cdot \|_{s,\gamma} \) defined by
\[ \| u \|^2_{s,\gamma} := (2\pi)^{-n} \int_{\mathbb{R}^n} \lambda^{s,\gamma}(\xi) |\hat{u}(\xi)|^2 d\xi, \] (25)
\( \hat{u} \) being the Fourier transform of \( u \). The norms defined by (25), with different values of \( \gamma \), are equivalent each other. For \( s \in \mathbb{N} \), (25) turns to be equivalent, uniformly with respect to \( \gamma \), to the norm
\[ \| u \|^2_{H^s_\omega(\mathbb{R}^n)} := \sum_{|\alpha| \leq s} \gamma^{2(s-|\alpha|)} \| \partial^\alpha u \|^2_{L^2_\omega(\mathbb{R}^n)}, \] (26)
where, for \( \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n \), we set \( \partial^\alpha := \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n} \) and \( |\alpha| := \alpha_1 + \cdots + \alpha_n \).

For functions defined over \( Q_T^+ \), we will consider the weighted Sobolev spaces \( H^m_\gamma(Q_T^+) \) equipped with the norm (where \( \partial^\alpha := \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n} \) and \( \partial_0 = \partial_t \))
\[ \| u \|^2_{H^m_\gamma(Q_T^+)} := \sum_{|\alpha| \leq m} \gamma^{2(m-|\alpha|)} \| \partial^\alpha u \|^2_{L^2_\omega(Q_T^+)}. \]

We also use the homogeneous space \( H^1(\omega_T) := \{ u \in L^1_{loc}(Q_T^+) \mid \nabla u \in L^2(Q_T^+) \} \).

5. The main result. Recall that \( U = (q, v, H) \), where we drop the dot from the variables for simplicity. The main result of the paper reads as follows.

**Theorem 5.1.** Let \( T > 0 \). Let the basic state (11) satisfy assumptions (12)-(14) and
\[ |\tilde{H} \times \tilde{H}| \geq \delta > 0, \text{ on } \omega_T, \]
where \( \delta \) is a fixed constant. Then there exists \( \gamma_0 \geq 1 \) such that for all \( \gamma \geq \gamma_0 \) and for all \( f, \gamma \in H^1_\gamma(Q_T^+) \) vanishing for \( t < 0 \), problem (23) has a unique solution \((U_\gamma, \gamma, \varphi_\gamma)\) such that \((q_\gamma, W_\gamma, H_\gamma, \varphi_\gamma) \in H^1_\gamma(Q_T^+) \times H^1_\gamma(Q_T^+) \times H^1_\gamma(\omega_T) \times H^1_\gamma(\omega_T)\) with the trace \((q_\gamma, u_1, \gamma, h_1, \gamma, H_\gamma) \big|_{\omega_T} \in H^{1/2}_\omega(\omega_T)\) and obeys the a priori estimate
\[ \gamma \left( \| W_\gamma \|^2_{H^1_\gamma(Q_T^+)} + \| \nabla q_\gamma \|^2_{L^2_\omega(Q_T^+)} + \| H_\gamma \|^2_{H^1_\gamma(Q_T^+)} \right) + \| (q_\gamma, u_1, \gamma, h_1, \gamma, H_\gamma) \big|_{\omega_T} \|_{H^{1/2}_\omega(\omega_T)}^2 + \gamma^2 \| \varphi_\gamma \|^2_{H^1_\gamma(\omega_T)} \leq C \gamma \| f, \gamma \|^2_{H^1_\gamma(Q_T^+)} \] (27)
where we have set \( U_\gamma := e^{-\gamma t} U, \quad H_\gamma := e^{-\gamma t} H, \quad \varphi_\gamma := e^{-\gamma t} \varphi \) and so on, and where \( C = C(K, T, \delta, \gamma > 0) \) is a constant independent of the data \( f, \gamma \) and the parameter \( \gamma \).

**Idea of the proof.** Problem (23) is a nonstandard initial-boundary value problem. For its resolution we introduce a fully hyperbolic approximation. Concerning the plasma part, we replace the incompressible MHD equations with their “compressible” counterpart by introducing an evolution equation for the total pressure involving a small parameter \( \varepsilon \) which corresponds to the reciprocal of the sound speed in the fluid. As for the vacuum part, we consider a “hyperbolic” regularization of the elliptic system (23) by introducing a new auxiliary unknown \( E \) which plays the role of the vacuum electric field, and the same small parameter \( \varepsilon \) as above is now associated with the physical parameter \( 1/c \), being \( c \) the speed of light. We also regularize the second boundary condition in (23) and introduce two boundary conditions for the unknown \( E \).
5.1. **Hyperbolic regularization of the reduced problem.** Let us denote $U^\varepsilon = (q^\varepsilon, v^\varepsilon, H^\varepsilon)$ (we also set $W^\varepsilon = (v^\varepsilon, H^\varepsilon)$). The regularized system for the plasma part reads

$$
\varepsilon^2 \left\{ \partial_t q^\varepsilon - (\partial_t \hat{H}, H^\varepsilon) - (\hat{H}, \partial_t H^\varepsilon) + \frac{1}{\partial_t \Phi_1} (\hat{w}, \nabla q^\varepsilon) - \frac{1}{\partial_t \Phi_1} (\hat{w}, (\nabla \hat{H}, H^\varepsilon)) \right\} + \frac{1}{\partial_t \Phi_1} \text{div } u^\varepsilon = 0, \tag{28a}
$$

$$
\partial_t v^\varepsilon + \frac{1}{\partial_t \Phi_1} \left\{ (\hat{w}, \nabla) v^\varepsilon - (\hat{h}, \nabla) H^\varepsilon \right\} + \nabla \hat{q}^\varepsilon + C_1 (\hat{W}, \hat{\Psi}) W^\varepsilon = f_v, \tag{28b}
$$

$$
\partial_t H^\varepsilon + \frac{1}{\partial_t \Phi_1} \left\{ (\hat{w}, \nabla) H^\varepsilon - (\hat{h}, \nabla) v^\varepsilon \right\} + C_2 (\hat{W}, \hat{\Psi}) W^\varepsilon + \frac{\hat{H}}{\partial_t \Phi_1} \text{div } u^\varepsilon = 0 \quad \text{in } Q_T^+, \tag{28c}
$$

where the matrices $C_1$ and $C_2$ were defined in (17), and $u^\varepsilon$ is defined through $v^\varepsilon$ like $u$ is defined through $v$. The latter system with $\varepsilon = 1$ looks like the linearized system of compressible isentropic MHD equations reduced to a dimensionless form.

Let us denote $V^\varepsilon = (H^\varepsilon, E^\varepsilon)$. We consider the following regularized system for the vacuum part:

$$
\varepsilon \partial_t h^\varepsilon + \nabla \times E^\varepsilon = 0, \quad \varepsilon \partial_t e^\varepsilon - \nabla \times \psi^\varepsilon = 0 \quad \text{in } Q_T^- \tag{29},
$$

where $E^\varepsilon = (E^\varepsilon_1, E^\varepsilon_2, E^\varepsilon_3)$, $E^\varepsilon_1 = (E^\varepsilon_1 \hat{\Phi}_1, E^\varepsilon_2, E^\varepsilon_3)$, $E^\varepsilon_2 = (E^\varepsilon_2 \hat{\Phi}_1, E^\varepsilon_2 \hat{\Phi}_1, E^\varepsilon_3)$, $E^\varepsilon_3 = (E^\varepsilon_3 \hat{\Phi}_1, E^\varepsilon_3 \hat{\Phi}_1, E^\varepsilon_3)$, $k = 2, 3$. All the other notations for $\mathcal{H}^\varepsilon$ (i.e. $h^\varepsilon$ and $\psi^\varepsilon$) are analogous of those for $\mathcal{H}$. If $\varepsilon = 1$, the above system coincides with the vacuum Maxwell equations.

We couple equations (28) and (29) with the following regularized boundary conditions

$$
\partial_t \phi^\varepsilon = v^\varepsilon_N - \hat{v}_2 \partial_2 \phi^\varepsilon - \hat{v}_3 \partial_3 \phi^\varepsilon + \phi^\varepsilon \partial_1 \hat{v}_N,
$$

$$
q^\varepsilon = (\hat{H}, \mathcal{H}^\varepsilon) - [\partial_t \hat{q}] \phi^\varepsilon - \varepsilon (\hat{E}, E^\varepsilon) \tag{30}
$$

$$
E^\varepsilon_2 = \varepsilon \partial_t (\hat{H}_2 \psi^\varepsilon) - \varepsilon \partial_2 (\hat{E}_1 \psi^\varepsilon),
$$

$$
E^\varepsilon_3 = -\varepsilon \partial_3 (\hat{H}_2 \psi^\varepsilon) - \varepsilon \partial_3 (\hat{E}_1 \psi^\varepsilon) \quad \text{on } \omega_T, \tag{30}
$$

where $E = (\hat{E}_1, \hat{E}_2, \hat{E}_3)$ and the coefficients $\hat{E}_j$ are given functions that must be chosen in a suitable way. Their choice will be crucial in order to make the boundary conditions dissipative.

We consider the regularized problem given by (28) - (30). For fixed $\varepsilon$, this problem looks as the linearized plasma-vacuum problem for compressible MHD studied by Secchi, Trakhinin [6]. The same analysis performed there leads to produce an energy estimate for sufficiently smooth solutions to problem (28) - (30) which is uniform with respect to the parameter $\varepsilon$. The main difficulty is to find a suitable change of unknown functions reducing the systems (28), (29) into an equivalent form, where the the matrix coefficients depend on $\varepsilon$ in such a way to guarantee the uniformity in $\varepsilon$ of the corresponding estimates. To solve the same difficulty, we perform a new symmetrization of the vacuum system (29), given by

$$
\widehat{K}^{-1} (\partial_t h^\varepsilon + \frac{1}{\varepsilon} \nabla \times E^\varepsilon) + \widehat{K}^{-1} (\partial_t e^\varepsilon - \frac{1}{\varepsilon} \nabla \times \psi^\varepsilon) \times e \nu + \frac{\nu}{\partial_t \Phi_1} \text{div } h^\varepsilon = 0,
$$

$$
\widehat{K}^{-1} (\partial_t e^\varepsilon - \frac{1}{\varepsilon} \nabla \times \psi^\varepsilon) - \widehat{K}^{-1} (\partial_t h^\varepsilon + \frac{1}{\varepsilon} \nabla \times E^\varepsilon) \times e \nu + \frac{\nu}{\partial_t \Phi_1} \text{div } e^\varepsilon = 0, \tag{31}
$$

where $\nu$ is a suitable viscosity coefficient.
where $\mathring{K} = \begin{pmatrix} 1 & -\partial_2 \Psi & -\partial_3 \Psi \\ 0 & \partial_1 \mathring{\Phi}_1 & 0 \\ 0 & 0 & \partial_1 \mathring{\Phi}_1 \end{pmatrix}$, while $\nu = (\nu_1, \nu_2, \nu_3)$ and $\nu_i = \nu_i(t, x)$ ($i = 1, 2, 3$) are arbitrary functions that will be chosen later on. For the proof of the equivalence of (29) and (31) with $\nu \neq 0$, we refer to [6, Lemma 16]. The above symmetrization allows to treat the boundary terms in the energy estimate for the regularized problem, by making the following choice of $\nu$ and $\mathring{E}$:

$$\nu_1 = \mathfrak{v}_1 = \hat{v}_2 \partial_2 \hat{\varphi} + \hat{v}_3 \partial_3 \hat{\varphi}, \quad \nu_k = \hat{v}_k, \quad k = 2, 3 \text{ on } \omega_T,$$

$$\mathring{E} := -\mathfrak{v} \times \hat{\mathcal{H}}, \quad \mathfrak{v} := (\mathfrak{v}_1, \hat{v}_2, \hat{v}_3).$$

Once the a priori estimate uniform in $\varepsilon$ is derived for the $\varepsilon-$regularized problem, one proves the existence of a solution of the regularized problem for each $\varepsilon$, see [6]. The uniformity in $\varepsilon$ of the estimate gives the boundedness, in the suitable functional spaces, of the sequence of solutions as $\varepsilon$ goes to zero. Up to a subsequence, one can prove the sequence is weakly convergent as $\varepsilon \to 0$ and the weak limit is a solution of the incompressible problem (23). The a priori estimate (27) is obtained by passing to the limit as $\varepsilon \to 0$ in the a priori estimate for the regularized problem.

REFERENCES


E-mail address: alessandro.morando@ing.unibs.it
E-mail address: paola.trebeschi@ing.unibs.it
E-mail address: trakhin@math.nsc.ru
RELAXATION-BASED APPROXIMATE RIEMANN SOLVERS
FOR ELASTIC RODS

Knut Waagan

Department of Applied Mathematics, University of Washington
Guggenheim Hall 414, Box 352420
Seattle, WA 98195-2420, USA

Abstract. We present Lax-Friedrichs and Suliciu type schemes for the equations of elastic rod theory. A simple rotational test problem reveals that the extension of these gas dynamics schemes to elasticity is not so straightforward.

1. System formulation. We consider the planar motion of a planar elastic rod, see [2]. The rod’s configuration at a time \( t \) is described by a parameterized curve \( r(s, t) \) in the plane, and the angle \( \theta(s, t) \) of the cross-sections with respect to a fixed reference. Define the internal coordinates by the orthonormal basis

\[
a(\theta) := \cos \theta \mathbf{i} + \sin \theta \mathbf{j}, \quad b(\theta) := -\sin \theta \mathbf{i} + \cos \theta \mathbf{j},
\]

where \( \mathbf{i}, \mathbf{j} \) is a fixed orthonormal basis. The equations can then be written

\[
\begin{align*}
\rho A_r r_{tt} &= (Na + Hb)_s, \quad (1.2a) \\
\rho J \theta_{tt} &= M_s + k \cdot [r_s \times (Na + Hb)]. \quad (1.2b)
\end{align*}
\]

We define the internal strain variables by \( r_s = \nu a + \eta b \) and \( \mu = \theta_s \). The internal contact forces only depend on the strain variables \( \nu, \eta \) and \( \mu \), which ensures that the system is invariant under rigid motion (translation and rotation). We assume that the rod is of strain rate type, i.e. that there exists a potential energy \( W(\nu, \eta, \mu) \) such that

\[
N = W_\nu, \quad H = W_\eta, \quad M = W_\mu. \quad (1.2c)
\]

Hyperbolicity of the system is ensured by assuming the Jacobian matrix

\[
W'' = \begin{bmatrix}
W_{\nu\nu} & W_{\nu\eta} & W_{\nu\mu} \\
W_{\eta\nu} & W_{\eta\eta} & W_{\eta\mu} \\
W_{\mu\nu} & W_{\mu\eta} & W_{\mu\mu}
\end{bmatrix} = \begin{bmatrix}
N_\nu & N_\eta & N_\mu \\
H_\nu & H_\eta & H_\mu \\
M_\nu & M_\eta & M_\mu
\end{bmatrix}
\]

is positive definite. We can rewrite (1.2) as a first order system by defining the velocity variables.

\[
v := r_t, \quad \omega := \theta_t.
\]
The most convenient velocity components are in internal coordinates, hence we define $u = v \cdot a$ and $v = v \cdot b$. This yields

\begin{align*}
\nu_t &= u_s - \mu v + \omega \eta, \\
\eta_t &= v_s + \mu u - \omega \nu, \\
\mu_t &= \omega_s, \\
\rho A u_t &= N^E_s - \mu H^E + \rho A \omega v, \\
\rho A v_t &= H^E_s + \mu N^E - \rho A \omega u, \\
\rho J \omega_t &= M^E_s + \nu H^E - \eta N^E.
\end{align*}

(1.4)

The first three equations, the compatibility equations, express equality of partial derivatives.

1.1. **Weak form.** The physically appropriate conservative form of the equations reads $r_{tt} = n_s$, and

\begin{align*}
(r \times r_t + \rho J \theta)_t &= (r \times n + M)_s, \\
\rho A (r \times r_t) + \rho J \theta_t &= (r \times n + M)_s,
\end{align*}

the latter equation expressing the conservation of angular momentum. One can show that the more convenient weak form based on the intrinsic coordinates is equivalent. Hence we will consider the first order form (1.4).

For solutions with shocks additional admissibility criteria are needed. A physically admissible shock should arise as a vanishing viscosity limit, where the viscosity terms are physically meaningful. There may be several such limits for elasticity problems. In gas dynamics the admissible shocks can often be sought by thermodynamical considerations. For isothermal solids, this procedure is easily copied. For smooth solutions, we have energy conservation

\begin{align*}
E_t &= (u N + v H + \omega M)_s, \\
E &= \frac{1}{2} (Au^2 + Av^2 + J \omega^2) + W(\nu, \eta, \mu).
\end{align*}

(1.6)

Around shock waves this becomes an inequality, stating that energy gets dissipated.

1.2. **Characteristic velocities.** The characteristic velocities are an important component of shock-capturing schemes. We will restrict ourselves to schemes that only require a good upper estimate for the fastest traveling wave. The squared characteristic velocities for the rod system equal the eigenvalues of the Hesse matrix (1.2d) of \( W \). To simplify expressions, we use that the eigenvalues of a symmetric, positive definite 3x3 matrix \( a_{ij} \) are bounded above by the sum of the diagonal entries \( a_{ii} \). Hence the characteristic velocities are bounded by

\[ \Lambda = \sqrt{N \nu / A + H \eta / A + M \mu / J}. \]

(1.7)

in 2D rod theory. A sharp value can be found using a power method or analytically finding the root of a third order equation.

2. **Numerical schemes.** The equations (1.4) may be written as the one-dimensio- nal hyperbolic balance law

\[ U_t + F(U)_x = G(U). \]

(2.1)

For shock admissibility, they may be equipped with the ‘entropy inequality’

\[ E(U)_t + Q(U)_x \leq 0, \]

(2.2)
with $E$ the energy (1.6), and $Q$ the corresponding flux, i.e. $Q = -(uN + vH + \omega M)$. Note that $E'(U) \cdot G(U) = 0$, hence the right hand side $G$ is energy preserving. We now review some standard numerical schemes for (2.1).

2.1. **Homogenous case.** First, we treat the classical conservation law case that $G = 0$. We will consider conservative schemes i.e. schemes of the form

$$U_{i}^{n+1} = U_{i}^{n} - \frac{\Delta t}{\Delta x} \left( F_{i+\frac{1}{2}}^{n} - F_{i-\frac{1}{2}}^{n} \right),$$

(2.3)

Hence, we are left with having to choose a consistent numerical flux function $F_{i+\frac{1}{2}}^{n}$. To restrict the choice we can impose a discrete version of the admissibility condition (2.2). We say a scheme is entropy stable if (2.3) implies the discrete entropy inequality

$$E(U_{i}^{n+1}) \leq E(U_{i}^{n}) - \frac{\Delta t}{\Delta x} \left( Q_{i+\frac{1}{2}}^{n} - Q_{i-\frac{1}{2}}^{n} \right),$$

(2.4)

with $Q_{i+\frac{1}{2}}$ consistent with the entropy flux $Q$. Some of the schemes we consider satisfy somewhat stronger entropy stability restrictions than this, but that is beyond our scope. We just note that simply having equality in (2.4) leads to strong oscillations around shocks.

The most basic conservative entropy stable scheme is the Godunov scheme which consists of the following iterated steps: (i) Approximate the state $U(x, \cdot)$ with piecewise constants $U_{i}$ by taking averages over a prescribed interval partition (these intervals are often called cells). (ii) Evolve the system with $U_{i}$ as initial data up to a time increment $\Delta t$. Working out the exact evolution in the second step is complicated, hence several approximate algorithms have been developed. We present two of these.

2.2. **Lax-Friedrichs scheme.** Consider first the homogenous hyperbolic system

$$U_{t} + F(U)_{x} = 0.$$  

(2.5)

Then formally, $F_{t} + F'(U)F_{x} = 0$. Hence we get a linearised system

$$U_{t} + \tilde{F}_{x} = 0, \quad \tilde{F}_{t} + C^{2}U_{x} = 0,$$

(2.6)

by choosing a constant matrix $C \approx F'(U)$. A consistent, conservative numerical scheme for (2.5) is given by a Godunov scheme for (2.6) with the addition that we enforce $\tilde{F} = F(U)$ after each time step. If we take $C = cI$ for a scalar $c$, we get the Lax-Friedrichs scheme given by the numerical fluxes

$$F_{i+\frac{1}{2}} = \frac{1}{2}(F(U_{i}) + F(U_{i+1})) + \frac{c}{2}(U_{i+1} - U_{i}).$$

(2.7)

It is easily seen that the Lax-Friedrichs scheme is equivalent to a second order accurate (in space) scheme for the viscous system

$$U_{t} + F(U)_{x} = \frac{c\Delta x}{2} U_{xx}.$$  

(2.8)

This right hand term is often referred to as numerical viscosity, and the equation (2.8) referred to as the model equation for the Lax-Friedrichs scheme. Numerical viscosity is an important component of shock-capturing schemes, however it is not always as easy to explicitly express as in (2.8).

We are left with a single parameter $c$ to choose, which can be interpreted as the either the signal velocity in (2.6) or the diffusion coefficient in (2.8). For entropy stability of the scheme, the choice of $c$ must be larger than the fastest characteristic speed. Several estimates have been suggested, and we will return to the issue for
our concrete system. For heterogeneous solutions, local values of \( c \) are preferred, leading to the Rusanov and HLL fluxes.

2.3. Suliciu scheme, boundary conditions. We next recount an approach for continuum dynamics that is more faithful to the characteristic wave structure than the Lax-Friedrichs scheme, for a detailed account of the gas dynamical case, see [3]. It was originally implemented to provide better resolution of material contact waves in gas dynamics. Consider the wave equation system

\[
\begin{align*}
\nu_t &= u_x, \quad u_t = \psi'(\nu)_x, \quad \psi'(\nu) > 0. 
\end{align*}
\]

If \( \psi(\nu) = N(\nu, 0, 0) \) this models pure longitudinal waves along a straight rod. The Lax-Friedrichs linearization takes the form

\[
\begin{align*}
\nu_t &= \hat{u}_x, \quad u_t = \hat{\psi}_x, \\
\hat{u}_t &= c^2 \nu_x, \quad \hat{\psi}_t = c^2 u_x. 
\end{align*}
\]

However, leaving the compatibility equation unchanged also produces a linear system, namely

\[
\begin{align*}
\nu_t &= u_x, \\
u_t &= \hat{\psi}_x, \\
\hat{\psi}_t &= c^2 u_x. 
\end{align*}
\]

This yields, analogously to Lax-Friedrichs, a numerical scheme approximating up to second order in space the viscous system

\[
\begin{align*}
\nu_t &= u_x + \frac{\Delta x}{2c} \psi_{xx}, \\
\nu_t &= \psi(\nu)_x + \frac{c \Delta x}{2} u_{xx}. 
\end{align*}
\]

Typical admissible boundary conditions for (2.9) consist of imposing either the ‘mass’ flux \( u \) or the momentum flux \( \psi \). These are also the natural boundary conditions for (2.11), with \( \hat{\psi} \) replaced by its counterpart \( \hat{\psi} \). On the other hand, the Lax-Friedrichs system (2.10) requires additional boundary data, due to having a larger number of characteristics with nonzero speed.

Stable choices for the signal velocity \( c \) can be derived as follows. The conditions that must be satisfied for entropy stability are that \( c^2 \geq \psi''(\nu) \) for all states in the Riemann solution of the linearisation system. Note that the values of \( \nu \) in the wave fan are given by

\[
\begin{align*}
\nu_l^* &= \nu_l + \frac{1}{2c} (u_r - u_l) + \frac{1}{2c^2} (\hat{\psi}_l - \hat{\psi}_r), \\
\nu_r^* &= \nu_r + \frac{1}{2c} (u_r - u_l) + \frac{1}{2c^2} (\hat{\psi}_r - \hat{\psi}_l). 
\end{align*}
\]

Hence, for \( \psi'' > 0 \), we can take

\[
c^2 \geq \psi' \left( \max(\nu_l, \nu_r) + \frac{1}{2c} (u_r - u_l)_+ + \frac{1}{2c^2} |\hat{\psi}_l - \hat{\psi}_r| \right). 
\]

This can be ensured by taking

\[
c^2 \geq \psi' \left( \max(\nu_l, \nu_r) + \frac{1}{2c} (u_r - u_l)_+ + \frac{1}{2c^2} |\hat{\psi}_l - \hat{\psi}_r| \right), \quad \hat{\psi} = \sqrt{\psi''(\max[\nu_l, (\nu_r)])}.
\]

Often, \( \psi' \) is readily given as a sum of increasing and decreasing functions, which allows similar estimates. For gas dynamics, the sound speed, i.e. \( \sqrt{\psi''} \) vanishes at
vacuum, i.e. at $\nu = \infty$, which leads to blow-up of $c$ unless we make more careful estimates (and use a slightly more general scheme). This is not an issue for elastic materials. However, we need to ensure that $\nu$ stays positive. This means that we must have

$$\nu_{l,r} + \frac{1}{2c} (u_r - u_l) - \frac{1}{2c^2} |\pi_l - \pi_r| > 0.$$  \hfill (2.16)

Solving for $c$, we get

$$c > \frac{2|\pi_l - \pi_r|}{\sqrt{(u_r - u_l)^2 + 8\nu_{l,r} |\pi_l - \pi_r| + (u_r - u_l)}}.$$  \hfill (2.17)

More generally we are interested in vectorial systems of form

$$p_t = q_s, \quad q_t = F(u)_s.$$  \hfill (2.18)

We linearize with

$$p_t = q_s, \quad q_t = \hat{F}_s, \quad \hat{F}_t = C^2 q_s.$$  \hfill (2.19)

We can analyze the scheme by interpreting it as a BGK relaxation scheme

$$p_t = q_s, \quad q_t = \hat{F}_s, \quad \hat{F}_t = C^2 q_s \frac{F - \hat{F}}{\epsilon},$$  \hfill (2.20)

where $C^2$ denotes a fixed positive definite matrix we need to choose. A Chapman-Enskog expansion yields that we have up to first order in $\epsilon$,

$$p_{st} = q_s, \quad q_t = p_s + \epsilon (C^2 q_s - F' \cdot q_s).$$  \hfill (2.21)

Consequently, dissipativity requires that the matrix $C^2 - F'$ is positive definite. Due to hyperbolicity $F'$ has a full orthogonal system of eigenvectors. Hence we may take a diagonal $C = cI$, with $c^2$ greater than the largest eigenvalue of $F'$. The numerical scheme can be analyzed in a similar way to the BGK scheme with the time step $\Delta t$ taking the role of $\epsilon$. Note that this construction and analysis applies to the homogenous part of (1.4). The positivity condition (2.17) is still valid, because the linearized equations for each component direction decouple.

### 2.4. Low order terms

Next, we consider the balance law

$$U_t + F(U)_x = G(U).$$  \hfill (2.22)

Often, the source term $G(U)$ is treated separately from the homogenous part rather than solving a Godunov scheme for (2.22). For example, let $U^{n}_{LF}$ denote a single time step update in the Lax-Friedrichs scheme (2.7). Then a first order accurate scheme for the balance law (2.22) is given by

$$U^{n+1} = U^{n}_{LF} + \Delta t G(U^n).$$  \hfill (2.23)

If the source term is stiff, often an implicit ODE solver is put in place of the last term. Recall that the system

$$U_t = G(U)$$  \hfill (2.24)

is energy conservative, i.e. it implies $E(U)_t = 0$, while the dissipation of energy is an important stability criterion. Generally forward Euler for the ODE system adds
an amount of energy proportional to $G(U)E''(U)G(U)\Delta t^2$ to leading order in $\Delta t$. As an alternative, the implicit Euler scheme

\[ U^{n+1} = U^n_F + \Delta tG(U^{n+1}). \quad (2.25) \]

tends to be dissipative. However, it is costly and complicated to compute. We propose the following ‘quasi-implicit’ scheme as an efficient alternative

\[ U^{n+1} = U^n + \Delta tG[U^n + \Delta tg(U^n)]. \quad (2.26) \]

The system (2.22) may have interesting steady states $U^*$ satisfying

\[ F(U^*)_x = G(U^*). \]

Sometimes it is desirable to ensure that the numerical scheme preserves the steady state. An example of such a scheme is given by

\[ U^{n+1} = [U^n_F + \Delta tG(U^n)] - [U^n_* + \Delta tG(U^*)]. \quad (2.27) \]

This construction obviously works quite generally, the disadvantage is that the steady state of interest must be explicitly implemented.

3. A simple numerical problem. We consider a problem from [1]. The same constitutive function is used with $h = A = 0.05$ and $J = h^3/12$. This problem has a simple steady state where all the intrinsic coordinate components are constants. It corresponds to a circular ring rotating at a constant angular velocity $\Omega_0$. There are no boundaries to take into account, instead we seek a periodic solution. The steady state longitudinal strain $\nu_0$ was calculated using the Matlab routine fsolve. When this steady state is used as initial data the solution is steady with our numerical schemes.

The initial velocities were perturbed so that $u(0,s) = \nu_0\omega_0 + 10^{-3}\cos(s)$ and $v(0,s) = 10^{-3}\sin(s)$ (the ring has circumference $2\pi$ at rest). In this case the schemes using the forward Euler scheme blow up. The total integrated energy

\[ E = \frac{1}{2}(Au^2 + Av^2 + J\omega^2) + W(\nu, \eta, \mu) \]

increases before visible onset of instability, which indicates that the schemes do not have enough dissipation. We plot the total energy as a function of time in figure 3.1 for the Lax-Friedrichs scheme. Replacing the forward Euler time integration with quasi-implicit time integration, seems to resolve this issue. The resulting time evolution of total energy for the Lax-Friedrichs scheme is in figure 3.2. The energy stabilizes around a value slightly smaller than the analytical value. The results with the Suliciu scheme were very similar.

This simple example illustrates that extending schemes that work well for gas dynamics to solid mechanics has complications. In the future we plan to perform further tests, including problems with shocks, and to study boundary conditions. A problem with the schemes presented here, as demonstrated in [1], is that the model equations (2.8) and (2.12) represent constitutive equations that are not invariant under rigid motion. Hence we want to compare our schemes to the invariant schemes of [1].
Figure 3.1. Time evolution of total energy with forward Euler.

Figure 3.2. Time evolution of total energy with quasi-implicit integration of low order terms.

REFERENCES


E-mail address: kwaagan@uw.edu
CONTROL OF A SCALAR CONSERVATION LAW WITH A NONLOCAL VELOCITY

JEAN-MICHEL CORON
Institut universitaire de France
Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie
4 place Jussieu, F-75005 Paris, France

ZHIQIANG WANG
Shanghai Key Laboratory for Contemporary Applied Mathematics
School of Mathematical Sciences, Fudan University
Shanghai 200433, China

Abstract. In this paper, we survey some results on control problems for a scalar conservation law with a nonlocal velocity, that models a highly re-entrant manufacturing system in semi-conductor production. For the open-loop system, we show the results on an optimal control problem, state controllability and out-flux controllability. While for the closed-loop system, we prove, by using a Lyapunov function approach, the exponential stabilization results in certain cases.

1. Introduction. This paper studies some control problems for control systems governed by the following scalar hyperbolic conservation law with a nonlocal velocity:

\[ \partial_t \rho(t, x) + \partial_x (\rho(t, x) \lambda(W(t))) = 0 \quad \text{where} \quad W(t) = \int_0^1 \rho(t, x) \, dx, \]

on a rectangular domain \((0, T) \times (0, 1)\) or the semi-infinite strip \((0, \infty) \times (0, 1)\). We assume that the velocity function \(\lambda\) is in \(C^1 (\mathbb{R}; (0, +\infty))\). Let us recall that the special case

\[ \lambda(s) = \frac{1}{1 + s}, \quad s \in [0, +\infty), \]

was, for example, used in [5, 17]. Compared to the results in [12, 13, 14], the non-negativeness of the data is eliminated in the whole paper, since \(\lambda\) is defined on \(\mathbb{R}\) (with suitable extension if necessary) instead of \([0, +\infty)\).

In the manufacture system, the initial data is given as the product density:

\[ \rho(0, x) = \rho_0(x), \quad x \in (0, 1). \]

For the open-loop system, the control is acted on the influx \(u(t) := \rho(t, 0) \lambda(W(t))\):

\[ u(t) = h(t). \]

While for the closed-loop control system, the following output feedback law is used:

\[ u(t) - \overline{\rho}(\overline{\rho}) = k(y(t) - \overline{\rho}(\overline{\rho})), \]
in which the output/measurement is the out-flux \( y(t) := \rho(t, 1)\lambda(W(t)), k \in \mathbb{R} \) is a tuning parameter, \( \rho \in \mathbb{R} \) is the equilibrium that we want to stabilize as \( t \to +\infty \).

The conservation law (1) arises in modeling of the semiconductor manufacturing systems, see e.g. [5, 17]. These systems are characterized by their highly re-entrant feature with very high volume (number of parts manufactured per unit time) and very large number of consecutive production steps as well. The main character of this model is described in terms of the velocity function \( \lambda \) which is a function of the total quantity of products in the plant(s).

The control problems for conservation laws and general hyperbolic systems have been widely studied. As for the controllability of nonlinear hyperbolic equations, we refer to [8, 15, 19] for solutions without shocks, and to [1, 2, 3, 4, 6, 16, 20] for solutions with shocks. As for asymptotic stability/stabilization of hyperbolic systems, we refer to [18, 21] for the strategy of careful analysis of solutions along the characteristic curves, and to [7, 9, 10, 23, 24] for the Lyapunov function approach.

Concerning the manufacturing model of (1) itself, an optimal control problem, motivated by [5, 17], related to the Demand Tracking Problem was studied in [12] (see also [22] for a generalized system where \( \lambda = \lambda(x, W(t)) \)). Later, the statement controllability and out-flux controllability for the open-loop system have been obtained in [13]. We show these results without proof in Section 3.

In Section 4, we prove the exponential stabilization for the closed-loop system (under the feedback law (4)) by Lyapunov function approach. The Lyapunov functions that we construct are inspired by [7, 11, 23] with necessary modifications according to the nonlocal feature of the nonlinear system, see [14]. The stabilization results depend on the equilibrium \( \overline{\rho} \neq 0 \) that we want to stabilize and the velocity function \( \lambda \) through the following quantity

\[
d := \frac{\overline{\rho}\lambda'(\overline{\rho})}{\lambda(\overline{\rho})}.
\]

### 2. Well-posedness of Cauchy problems

As for the open-loop system, we recall first the definition of the weak solution to the Cauchy problem (see, e.g., [9, Section 2.1]) and then its well-posedness; see [12, 22] for its proof.

**Definition 2.1.** Let \( T > 0, p \in [1, +\infty), \rho_0 \in L^p(0, 1) \) and \( h \in L^p(0, T) \) be given. A weak solution of the Cauchy problem

\[
\begin{cases}
\rho_t(t, x) + (\rho(t, x)\lambda(W(t)))_x = 0, & t \in (0, T), x \in (0, 1), \\
\rho(0, x) = \rho_0(x), & x \in (0, 1), \\
u(t) = h(t), & t \in (0, T)
\end{cases}
\]

is a function \( \rho \in C^0([0, T]; L^p(0, 1)) \) such that, for every \( \tau \in [0, T] \) and every \( \varphi \in C^1([0, \tau] \times [0, 1]) \) satisfying \( \varphi(\tau, x) = 0, \forall x \in [0, 1] \) and \( \varphi(1, t) = 0, \forall t \in [0, \tau] \), one has

\[
\int_0^\tau \int_0^1 \rho(t, x)(\varphi_t(t, x) + \lambda(W(t))\varphi_x(t, x))dxdt
+ \int_0^\tau h(t)\varphi(t, 0)dt + \int_0^1 \rho_0(x)\varphi(0, x)dx = 0.
\]

**Theorem 2.2 (Well-posedness of open-loop system).** Let \( T > 0 \) and \( p \in [1, +\infty) \) be given. For any \( \rho_0 \in L^p(0, 1) \) and \( h \in L^p(0, T) \), the Cauchy problem
admits a unique weak solution \( \rho \in C^0([0,T]; L^p(0,1)) \), and additionally \( \rho \in C^0([0,1]; L^p(0,T)) \). Moreover, the following maps

\[
(\rho_0, h) \in L^p(0,1) \times L^p(0,T) \mapsto \rho \in C^0([0,T]; L^p(0,1)),
\]

\[
(\rho_0, h) \in L^p(0,1) \times L^p(0,T) \mapsto y \in L^p(0,T)
\]

are continuous.

Remark 1. Theorem 2.2 has been generalized to the case where \( \lambda = \lambda(x, W(t)) \) in [22].

As for the closed-loop system, the definition and the well-posedness of the weak solution to the Cauchy problem are as follows.

Definition 2.3. Let \( \overline{\rho} \in \mathbb{R}, p \in [1, +\infty), k \in \mathbb{R} \) and \( \rho_0 \in L^p(0,1) \) be given. A weak solution of the Cauchy problem

\[
\begin{aligned}
\rho_t(t, x) + (\rho(t, x)\lambda(W(t)))_x &= 0, \quad t \in (0, +\infty), x \in (0, 1), \\
\rho(0, x) &= \rho_0(x), \quad x \in (0, 1), \\
u(t) - p\lambda(\overline{\rho}) &= k(y(t) - p\lambda(\overline{\rho})), \quad t \in (0, +\infty)
\end{aligned}
\]

is a function \( \rho \in C^0([0, +\infty); L^p(0,1)) \) such that, for every \( T > 0 \), every \( \tau \in [0,T] \) and every \( \varphi \in C^1([0,\tau] \times [0,1]) \) satisfying \( \varphi(\tau, x) = 0, \forall x \in [0,1] \) and \( \varphi(t,1) = 0, \forall t \in [0,\tau] \), one has

\[
- \int_0^1 \rho(t,x)(\varphi_t(t,x) + \lambda(W(t))\varphi_x(t,x))dx dt - \int_0^1 \rho_0(x)\varphi(0,x)dx
\]

\[
+ \int_0^T \left( y(t)\varphi(t,1) - [ky(t) + (1-k)p\lambda(\overline{\rho})]\varphi(t,0) \right)dt = 0.
\]

Theorem 2.4 (Well-posedness of closed-loop system). Let \( \overline{\rho} \in \mathbb{R}, p \in [1, +\infty) \) and \( k \in \mathbb{R} \) be given. For any given \( \rho_0 \in L^p(0,1) \), the Cauchy problem (10) has a unique weak solution \( \rho \in C^0([0, +\infty); L^p(0,1)) \). Moreover, for every \( T > 0 \), the following maps

\[
\rho_0 \in L^p(0,1) \mapsto \rho \in C^0([0,T]; L^p(0,1)),
\]

\[
\rho_0 \in L^p(0,1) \mapsto (u, y) \in L^p(0,T) \times L^p(0,T),
\]

are continuous.

Sketch of proof. We introduce the characteristic curve:

\[
\frac{d\xi}{ds} = \lambda(W(s)), \quad s \geq 0 \quad \text{with} \quad W(s) = \int_0^s \rho(s,x)dx.
\]

Since \( \rho \) is constant along the characteristics, we define a solution candidate in terms of \( \xi \):

\[
\rho(t,x) = \begin{cases} 
\rho_0(x - \xi(t)), & \text{if } 0 \leq \xi(t) \leq x \leq 1, \\
kp(\xi^{-1}(\xi(t) - x),1) + \frac{(1-k)p\lambda(\overline{\rho})}{\lambda(W(\xi^{-1}(\xi(t) - x)))}, & \text{if } 0 \leq x \leq \xi(t) - n + 1 \leq 1, \text{ or } 0 \leq \xi(t) - n \leq x \leq 1 \text{ for } n \in \mathbb{N}.
\end{cases}
\]

By fixed point argument as in [12], we prove that (15) is indeed the unique weak solution to the closed-loop system (6). Thanks to (15), we also get the continuity of the maps (12) and (13): see the proof of [22, Theorem 4.1].

\( \Box \)
3. Control problems of the open-loop system. In this section, we give some results on control problems for the open-loop system (6).

First we obtain the existence of the \( L^2 \)-optimal control for the demand tracking problem. For its proof, we refer to [12].

**Theorem 3.1 (Optimal control problem).** Let \( T > 0 \) be given. For any \( \rho_0 \in L^2(0,1) \) and any \( y_d \in L^2(0,T) \), let \( J \) be the functional on \( L^2(0,T) \):

\[
J(h) := \int_0^T |h(t)|^2 dt + \int_0^T |y(t) - y_d(t)|^2 dt,
\]

where \( y(t) = \rho(t,1)\lambda(W(t)) \) is the out-flux corresponding to the influx \( h \in L^2(0,T) \) and initial data \( \rho_0 \). Then, there exists \( h_\infty \in L^2(0,T) \) such that \( J(h_\infty) = \inf_{h \in L^2(0,T)} J(h) \).

Next, we have three theorems on controllability of the system (6); see [13] for proofs.

**Theorem 3.2 (Local state controllability).** Let \( \overline{\rho} \in \mathbb{R} \) be the given constant equilibrium and let

\[
T_0 := \frac{1}{\lambda(\overline{\rho})}.
\]

Then, for any \( T > T_0 \), any \( \varepsilon > 0 \) and any \( p \in [1, +\infty) \), there exists \( \nu > 0 \) such that, for any \( \rho_0 \in L^p(0,1) \) and any \( \rho_1 \in L^p(0,1) \) with \( \|\rho_0(\cdot) - \overline{\rho}\|_{L^p(0,1)} \leq \nu \) and \( \|\rho_1(\cdot) - \overline{\rho}\|_{L^p(0,1)} \leq \nu \), there exists \( h \in L^p(0,T) \) with

\[
\|h(\cdot) - \overline{\rho}\lambda(h(\cdot))\|_{L^p(0,T)} \leq \varepsilon,
\]

such that the weak solution \( \rho \in C^0([0,T]; L^p(0,1)) \) to the Cauchy problem (6) satisfies the final condition

\[
\rho(T, x) = \rho_1(x), \quad x \in (0,1)
\]

and the estimate

\[
\|\rho(t, \cdot) - \overline{\rho}\|_{L^p(0,1)} \leq \varepsilon, \quad \forall t \in [0,T].
\]

**Theorem 3.3 (Global state controllability).** For any \( p \in [1, +\infty) \), any \( \rho_0 \in L^p(0,1) \) and any \( \rho_1 \in L^p(0,1) \), there exists \( T_1 > 0 \) (depending on \( \rho_0 \) and \( \rho_1 \)) such that the following holds: For any \( T \geq T_1 \), there exists \( h \in L^p(0,T) \) such that the weak solution \( \rho \in C^0([0,T]; L^p(0,1)) \) to the Cauchy problem (6) satisfies the final condition (19).

**Theorem 3.4 (Out-flux controllability).** Let \( \overline{\rho} \in \mathbb{R} \) be the given constant equilibrium and let \( T_0 \) be given by (17). For any \( p \in [1, +\infty) \), any \( \varepsilon > 0 \) and any \( T_1, T \) with \( T_0 < T_1 < T \), there exists \( \nu > 0 \) such that the following holds: For any \( \rho_0 \in L^p(0,1) \) and any \( y_d \in L^p(T_1,T) \) with \( \|\rho_0(\cdot) - \overline{\rho}\|_{L^p(0,1)} \leq \nu \) and \( \|y_d(\cdot) - \overline{\rho}\lambda(\overline{\rho})\|_{L^p(T_1,T)} \leq \nu \), there exists \( h \in L^p(0,T) \) satisfying (18) such that the weak solution \( \rho \in C^0([0,T]; L^p(0,1)) \) to the Cauchy problem (6) satisfies the estimate (20) and the out-flux condition

\[
y(t) = y_d(t), \quad t \in (T_1, T).
\]
4. Exponential stabilization of the closed-loop system. In this section, we stabilize the nonlinear system to an equilibrium \( \overline{p} \in \mathbb{R} \) by using Lyapunov function approach. The stabilization results have been divided into two cases: \( \overline{p} = 0 \) and \( \overline{p} \neq 0 \), because the situation of \( \overline{p} \neq 0 \) is much more complicated than that of \( \overline{p} = 0 \). The main assumptions, \( d > -1 \) and \( k \in (-1, 1) \), of exponential stability are due to the spectral analysis of the linearized system of (10) near \( \overline{p} \), see [14] for the details. Based on these facts, we construct the candidate Lyapunov functions for the linearized system, with which we obtain exponential stabilization for the nonlinear system as a perturbation to the linearized one. The proofs are written for regular solutions, but they are true for weak solution as well because of the continuous dependence of the weak solution with respect to the initial data. More precisely, we have the following theorems.

**Theorem 4.1 (Stabilization to \( \overline{p} = 0 \)).** Let \( k \in (-1, 1) \). For every \( R > 0 \), there exist constants \( C = C(k, R) > 0 \) and \( \alpha = \alpha(k, R) > 0 \) such that for any \( \rho_0 \in L^2(0, 1) \) with

\[
\| \rho_0 \|_{L^1(0,1)} \leq R,
\]

the solution \( \rho \in C^0([0, +\infty); L^2(0, 1)) \) to the Cauchy problem

\[
\begin{aligned}
\rho_t(t, x) + (\rho(t, x) \lambda(W(t)))_x &= 0, & t &\in (0, +\infty), x &\in (0, 1), \\
\rho(0, x) &= \rho_0(x), & x &\in (0, 1), \\
\rho(t) &= k\rho(t), & t &\in (0, +\infty)
\end{aligned}
\]

satisfies

\[
\| \rho(t, \cdot) \|_{L^2(0,1)} \leq C e^{-\alpha t} \| \rho_0 \|_{L^2(0,1)}, \quad \forall t \in [0, +\infty).
\]

**Sketch of proof.** Motivated by the Lyapunov functions used in [7, 11, 23], we introduce the following Lyapunov function (equivalent to \( \| \rho(t, \cdot) \|_{L^2(0,1)}^2 \)):

\[
L_1(t) := \int_0^1 e^{-\beta x} \rho^2(t, x) dx, \quad \forall t \in [0, +\infty),
\]

where \( \beta > 0 \) is a positive constant to be determined. Then, along the trajectories of (23),

\[
\dot{L}_1(t) = -\beta \lambda(W(t))L_1(t) + \left( \lambda(W(t)) \right)^{-1}(2 - e^{-\beta})y^2(t) \leq -\beta \lambda(W(t))L_1(t)
\]

by letting \( \beta > 0 \) be such that \( k^2 < e^{-\beta} < 1 \).

In order to get exponential decay of the solution as \( t \to +\infty \), it suffices to prove the uniform boundedness of \( W(\cdot) \). In fact, \( t \mapsto \int_0^1 |\rho(t, x)| dx \) is a nonincreasing function which leads to that

\[
|W(t)| \leq \| \rho_0 \|_{L^1(0,1)} \leq R, \quad \forall t \in [0, +\infty).
\]

Then, from (22), (26) and (27), we conclude the proof of Theorem 4.1.

**Theorem 4.2 (Stabilization to \( \overline{p} \neq 0 \)).** Assume that \( d > -1 \). Let \( k \in (-1, 1) \). Then there exist constants \( \varepsilon = \varepsilon(\overline{p}, k) > 0 \), \( C = C(\overline{p}, k) > 0 \) and \( \alpha = \alpha(\overline{p}, k) > 0 \) such that the following holds: For every \( \rho_0 \in L^2(0, 1) \) with

\[
\| \rho_0(\cdot) - \overline{p} \|_{L^2(0,1)} \leq \varepsilon,
\]

the weak solution \( \rho \in C^0([0, +\infty); L^2(0, 1)) \) to the Cauchy problem (10) satisfies

\[
\| \rho(t, \cdot) - \overline{p} \|_{L^2(0,1)} \leq C e^{-\alpha t} \| \rho_0(\cdot) - \overline{p} \|_{L^2(0,1)}, \quad \forall t \in [0, +\infty).
\]
Sketch of proof. The proof is divided into two cases.

Case 1: $|d| < 1$. We construct a Lyapunov function (equivalent to $\|\rho(t, \cdot) - \overline{p}\|_{L^2(0,1)}^2$):

$$L_2(t) := \int_0^1 e^{-\beta x} (\rho(t,x) - \overline{p})^2 dx + a(W(t) - \overline{p})^2, \quad \forall t \in [0, +\infty),$$

with

$$a := \frac{e^{-\beta} - k}{1 - k}, \quad k^2 < e^{-\beta} < 1.$$  \hspace{1cm} (31)

Note that

$$\dot{W}(t) = \int_0^1 \rho_t(t,x) dx = u(t) - y(t).$$  \hspace{1cm} (32)

Note also that

$$\lambda(W(t)) = \lambda(\overline{p}) + \lambda'(\overline{p})(W(t) - \overline{p}) + o(1)(W(t) - \overline{p}), \quad \forall t \in [0, +\infty)$$  \hspace{1cm} (33)

since $\lambda$ is of class $C^1$. Here and in the following, $o(1)$ denotes various quantities which tend to 0 as $|W(t) - \overline{p}| \to 0$. Let us compute the time derivative of $L_2(t)$ for any classical solution of (10):

$$\dot{L}_2(t) = -\lambda(W(t)) \int_0^1 e^{-\beta x} ((\rho(t,x) - \overline{p})^2)_x dx + 2a(W(t) - \overline{p}) \dot{W}(t)$$

$$= -\beta \lambda(W(t)) \int_0^1 e^{-\beta x} (\rho(t,x) - \overline{p})^2 dx + BT,$$

where

$$BT = \frac{(u(t) - \overline{p} \lambda(W(t)))^2 - e^{-\beta}(y(t) - \overline{p} \lambda(W(t)))^2}{\lambda(W(t))} + 2a(W(t) - \overline{p})(u(t) - y(t))$$

$$\leq \lambda(\overline{p}) |d^2(1 - e^{-\beta}) + o(1)| (W(t) - \overline{p})^2.$$  \hspace{1cm} (35)

due to (5), (31), (32) and (33). Then, it follows from (30), (31), (33), (34), (35) and Hölder inequality that for some constant $c > 0$,

$$\dot{L}_2(t) \leq -c \beta \lambda(\overline{p}) |1 - d^2(e^\beta - 1)^2 e^{-\beta} \beta^{-2} + o(1)| L_2(t).$$

(36)

Note that $1 - d^2(e^\beta - 1)^2 e^{-\beta} \beta^{-2} \to 1 - d^2 > 0$ as $\beta \to 0^+$. Finally, we choose firstly $\beta$, then $\varepsilon$, sufficiently small, so that $1 - d^2(e^\beta - 1)^2 e^{-\beta} \beta^{-2} + o(1) > (1 - d^2)/2 > 0$ by the continuity of the mapping $\rho_0 \mapsto \rho$. This concludes the proof of Theorem 4.2 in the case $|d| < 1$.

Case 2: $d \geq 1$. Let

$$V_1(t) := \int_0^1 (\rho(t,x) - \overline{p})^2 dx + d(W(t) - \overline{p})^2.$$  \hspace{1cm} (37)

Then, by (32), (33) and the Cauchy-Schwarz inequality,

$$\dot{V}_1(t) = -\lambda(W(t)) \int_0^1 ((\rho(t,x) - \overline{p})^2)_x dx + 2d(W(t) - \overline{p}) \dot{W}(t)$$

$$= \lambda(W(t))(k^2 - 1) \xi^2(t,1) + o(1)W(t) \xi(t,1) + o(1)W^2(t)$$

$$\leq \lambda(\overline{p})(k^2 - 1 + o(1)) \xi^2(t,1) + o(1)W^2(t),$$

(38)
where $\xi(t, x) := (\rho(t, x) - \overline{\rho}) + d(W(t) - \overline{\rho})$. It is easy to check that $\xi$ satisfies the following Cauchy problem

$$
\begin{cases}
\xi_t(t, x) + \lambda(W(t))\xi_x(t, x) = d\dot{W}(t), & t \in (0, +\infty), x \in (0, 1),
\xi(0, x) = (\rho_0(x) - \overline{\rho}) + d(W(0) - \overline{\rho}), & x \in (0, 1),
\xi(t, 0) = k\xi(t, 1) + (k - 1) \left( \frac{p(\lambda(W(t)) - \lambda(\overline{\rho}))}{\lambda(W(t))} - d(W(t) - \overline{\rho}) \right), & t \in (0, +\infty).
\end{cases}
$$

(39)

Let

$$
V_2(t) := \int_0^1 e^{-x} \xi^2(t, x) dx.
$$

(40)

Then, by (32), (33), (39) and the Cauchy-Schwarz inequality,

$$
\begin{align*}
\dot{V}_2(t) &= -\lambda(W(t)) \int_0^1 e^{-x}(\xi^2(t, x))_x dx + 2d\dot{W}(t) \int_0^1 e^{-x}\xi(t, x) dx \\
&= -\lambda(W(t))V_2(t) + \lambda(W(t))(k^2 - e^{-1})\xi^2(t, 1) + o(1)W(t)\xi(t, 1) + o(1)W^2(t) \\
&\quad + 2d(k - 1)\lambda(W(t))(\xi(t, 1) + o(1)W(t)) \int_0^1 e^{-x}\xi(t, x) dx \\
&\leq (-\lambda(\overline{\rho}) + o(1))V_2(t) + A(1 + o(1))\xi^2(t, 1) + o(1)W^2(t),
\end{align*}
$$

(41)

for some constant $A > 0$.

Finally, we define the Lyapunov function (equivalent to $\|\rho(t, \cdot) - \overline{\rho}\|_{L^2(0, 1)}^2$) as follows:

$$
V(t) := \frac{2A}{\lambda(\overline{\rho})(1 - k^2)}V_1(t) + V_2(t), \quad \forall t \in [0, +\infty).
$$

(42)

Again, letting $\varepsilon$ small and noting the continuity of the mapping $\rho_0 \mapsto \rho$, we conclude, from (38), (41) and (42), the proof of Theorem 4.2 in the case $d \geq 1$. 

\textbf{Acknowledgments.} This work was partially supported by the ERC advanced grant 266907 (CPDENL) of the 7th Research Framework Programme (FP7). Zhiquiang Wang was also supported by Shanghai Pujiang Program (No. 11PJ1401200), by the Natural Science Foundation of Shanghai (No. 11ZR1402500) and by the National Science Foundation of China (No. 11271082).

\textbf{REFERENCES}


E-mail address: coron@ann.jussieu.fr
E-mail address: wzq@fudan.edu.cn
BURGERS’ EQUATION WITH A FILTERED VELOCITY

NILS HENRIK RISEBRO AND FRANZISKA WEBER

CMA, University of Oslo
P.O. Box 1053 Blindern
0316 Oslo, Norway

Abstract. The properties of the solution to the convectively filtered Burgers’ equation, a regularization of Burgers’ equation with the convective velocity replaced by a nonlocal averaged velocity, are examined. It is found that the limit of solutions, as the regularizing parameter $\alpha$ goes to zero, does not satisfy an entropy inequality owing to the reversibility of the equation and the absence of an $L^1$-contraction estimate for the limit of solutions.

1. Introduction. We consider the convectively filtered Burgers’ (CFB) equation,

\begin{align}
\frac{d}{dt}u^\alpha + \pi^\alpha_x &= 0, \quad (x, t) \in \mathbb{R} \times (0, T), \\
\pi^\alpha(x) &= g_\alpha * u^\alpha(x), \quad x \in \mathbb{R}, \\
u^\alpha(x, 0) &= u_0(x), \quad x \in \mathbb{R},
\end{align}

(1a, 1b, 1c)

where $\alpha > 0$, $u_0 \in BV(\mathbb{R}) \cap L^\infty(\mathbb{R})$ is the initial data, $g_\alpha(x) = \frac{1}{\alpha}g(\frac{x}{\alpha})$, and $g$ is a nonnegative, symmetric and non-increasing (with respect to the absolute value of the argument) function with $\int g = 1$. (We will call a function $g$ with the latter four properties a filter.) Setting $\alpha = 0$ in the above equation, one obtains the inviscid Burgers’ equation in nonconservative form,

\begin{align}
\frac{d}{dt}u + uu_x &= 0, \quad (x, t) \in \mathbb{R} \times (0, T), \\
u(x, 0) &= u_0(x), \quad x \in \mathbb{R}.
\end{align}

(2a, 2b)

It is well-known that the unique entropy solution to inviscid Burgers’ equation can be obtained as the strong limit in $C([0, T]; L^1_{loc}(\mathbb{R}))$ of the viscous Burgers’ equation

\begin{align}
\frac{d}{dt}u^\epsilon + \frac{(u^\epsilon_x)^2}{2} x &= \epsilon u^\epsilon_{xx}, \quad (x, t) \in \mathbb{R} \times (0, T), \\
u^\epsilon(x, 0) &= u_0(x), \quad x \in \mathbb{R},
\end{align}

(3a, 3b)

when $\epsilon \to 0$. The aim of this note, is to investigate, whether the sequence of solutions $(u^\alpha)_{\alpha > 0}$ also tends to a limit function and if such limits are entropy of the inviscid Burgers’ equation (2). This question is motivated by the fact that many methods in computational fluid dynamics use averaged or filtered quantities, as an example we mention the Leray-alpha model for the Navier-Stokes equation introduced in 1934 by Leray [5].

2000 Mathematics Subject Classification. Primary: 35L60, 35L65.

Key words and phrases. Conservation laws, averaged variables, Leray-alpha model, nonlocal partial differential equation, Burgers equation, convergence.
1.1. Well-Posedness and Convergence. Global existence and uniqueness of a solution to (1) has been shown in [1], under the assumptions that the initial data $u_0 \in L^\infty(\mathbb{R})$ can be written as the sum of a bounded Lipschitz continuous function and an $L^1$-function, and that the filter $g$ satisfies $g \in W^{2,1}(\mathbb{R})$ and $g'' \in L^\infty(\mathbb{R})$. If $u_0 \in BV(\mathbb{R})$, the same proof can be slightly modified to show existence and uniqueness for $g \in W^{1,1}(\mathbb{R})$, which includes in particular the Helmholtz filter

$$g(x) = \frac{1}{2}e^{-|x|}. \hspace{1cm} (4)$$

The proof in [1] uses the fact that the characteristics of the equation,

$$\frac{d}{dt}\eta(X,t) = \pi^\alpha(\eta(X,t), t),$$

$$\eta(X,0) = X, \hspace{1cm} (5)$$

do not cross. This implies that the map $\eta$ is a diffeomorphism for all times $t$ and that if $u_0$ is smooth, $u^\alpha(\cdot, t)$ is smooth for all times. Moreover, by (5) and (1a), the solution can be expressed as a reparametrization of the initial data at any time $t$:

$$u(x,t) = u_0(\eta^{-1}(x,t)). \hspace{1cm} (6)$$

In particular, this implies that the $L^\infty$-norm and the total variation of the solution $u^\alpha$ and the filtered velocity $\pi^\alpha$ are preserved in time and that $u^\alpha$ and $\pi^\alpha$ are Lipschitz continuous in time uniformly for all $\alpha > 0$. So convergence of a subsequence $(u^\alpha_n)_n$ of solutions of (1) and of filtered velocities $(\pi^\alpha_n)_n$ to a limit function $u$ in $C([0,T]; L^1_{loc}(\mathbb{R}))$ is guaranteed by Kolmogorov’s compactness theorem (see e.g. [7, Theorem A.8]). It is then natural to ask, whether the limit $u$ is a weak or an entropy solution of the inviscid Burgers’ equation (2).

Indeed, for the Helmholtz filter (4) it has been shown in addition that the limit function $u$ is a weak solution of Burgers’ equation [2]. For other filters, which cannot be represented as Green’s functions of differential operators, the question of convergence is, to our knowledge, not yet resolved. Similarly, proving convergence to the unique entropy solution seems to be complicated, see [3]. This note aims to identify some reasons for this.

2. Reversibility. One problem related to the regularization (1), is that this regularization does not imply an entropy inequality, in contrast to the viscous regularization (3):

**Proposition 1.** The convectively filtered Burgers’ equation (1) does not imply an entropy inequality

$$\eta(u)_t + q(u)_x \leq 0, \hspace{1cm} (7)$$

where $\eta$ is a convex function and $q$ with $q'(u) = \eta'(u)$ its corresponding entropy flux, for the limit $u = \lim_{\alpha \to 0} u^\alpha$ in $L^1$ of solutions $u^\alpha$ to (1).

**Proof.** We will prove the proposition by a contradiction argument.

Let $u^\alpha$ denote the solution of the CFB equation (1) and $v^\alpha(x,t) := u^\alpha(-x,-t)$, $(x,t) \in \mathbb{R} \times (0,T)$ its reflection in time and space. Observing that

$$\pi^\alpha(x,t) = \int_\mathbb{R} u^\alpha(-x-y,-t)g_\alpha(y)dy = \pi^\alpha(-x,-t)$$

and

$$\frac{d}{dt}v^\alpha(x,t) = -u^\alpha_t(-x,-t), \hspace{1cm} \frac{d}{dx}v^\alpha(x,t) = -u^\alpha_x(-x,-t),$$

we see that

$$\frac{d}{dt}v^\alpha(x,t) + q(v^\alpha)_x \leq 0$$

where $v^\alpha(\cdot,t)$ is the reflected solution $u^\alpha(\cdot,-t)$. Hence, if $u^\alpha$ is an entropy solution, then $v^\alpha$ is an entropy solution and

$$\eta(v^\alpha)_t + q(v^\alpha)_x \leq 0. \hspace{1cm} (8)$$

This is a contradiction, as $v^\alpha$ is the reflection of $u^\alpha$. Therefore, $u^\alpha$ is not an entropy solution, and thus the limit $u$ is not an entropy solution of Burgers’ equation.
thus \( v^\alpha \) satisfies the equation
\[
-v^\alpha_t(x, t) - \varphi^\alpha(x, t)v^\alpha_x(x, t) = 0 \quad (x, t) \in \mathbb{R} \times (0, T)
\]
which is the same differential equation as the one \( u^\alpha \) satisfies (except for the initial data which is reflected). If we assume that the limit function \( u = \lim_{\alpha \to 0} u^\alpha \), is a weak solution to inviscid Burgers’ equation (2), then the limit \( v = \lim_{\alpha \to 0} v^\alpha \) of the sequence of functions \( v^\alpha(x, t) \), as \( \alpha \to 0 \), will also be a weak solution of Burgers’ equation with reflected initial data. If we furthermore assume that (1a) implies that the limit \( u \) satisfies the entropy inequality (7), then, since \( v \) is the limit of the sequence \( v^\alpha \), this would imply that \( v \) satisfies the entropy inequality as well. So
\[
0 \geq \eta(v)_t + q(v)_x
= -\eta(u) - q(v)_x,
\]
and \( u \) can satisfy (7) only with equality. This is a contradiction: It is well-known that the entropy solution to Burgers’ equation (2) develops shocks so that (7) is a strict inequality in some cases. \( \Box \)

An explicit example of an initial data for (1) for which the solution \( u^\alpha \) does not converge to the entropy solution of Burgers’ equation was provided in the paper of Bhat and Fetecau [1]. There, the solution to (1) for a Riemann Problem with \( u_L < u_R \) is calculated and it is shown that in the limit \( \alpha \to 0 \), it converges to a weak solution with a non-entropic shock wave.

Nevertheless it has been conjectured, and numerical experiments provide some evidence it might be true [2], that for continuous initial data, \( u^\alpha \) converges to the entropy solution of Burgers’ equation. However, as we shall show in the next section, there is no uniform \( L^1 \)-stability in \( \alpha \) with respect to the initial data, which makes a possible proof of this convergence hard.

Another approach would be to investigate a filter which depends in addition on the time variable \( t \), thus breaking the symmetry \( t \to -t \). This led to satisfactory results in the context of two phase fluid flow [4], but our investigations have shown that the same filter applied in (1) doesn’t imply the desired entropy inequality (7) either [8].

3. \( L^1 \)-continuity of the CFB equation with respect to initial data. We have the following \( L^1 \)-stability estimate with respect to the initial data:

**Proposition 2.** Let \( u_0 \in W^{1,\infty}(\mathbb{R}) \cap BV(\mathbb{R}) \) and \( v_0 \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R}) \). Denote \( M_1 = \max\{\|u_0\|_\infty, \|v_0\|_\infty\} \), \( M_2 = \|u_0\|_\infty \) and \( M_3 = TV(u_0) \). Let \( u^\alpha, v^\alpha \) be the solutions to (1a), (1b) with initial condition \( u^0, v^0 \) respectively and \( g \in W^{1,1}(\mathbb{R}) \).

Then we have for \( t \in (0, T) \) and a fixed \( \alpha > 0 \),
\[
\|u^\alpha(\cdot, t) - v^\alpha(\cdot, t)\|_{L^1(\mathbb{R})} \leq e^{t(k_1 M_2 + M_3)} \|u_0 - v_0\|_{L^1(\mathbb{R})} \quad (8)
\]
where \( k_1 \) and \( k_2 \) are constants depending on the filter \( g \).

**Proof.** For simplicity we write \( u := u^\alpha \) and \( v := v^\alpha \). Subtracting (1a) for \( v \) from the same equation for \( u \) and adding and subtracting the term \( \nu u_x \), we obtain
\[
(u - v)_t + (\nu - \varphi) u_x + \varphi(u_x - v_x) = 0.
\]
We multiply the equation by \( \text{sgn}(u - v) \), bring the second and third term to the right-hand side, and integrate over the spatial domain,

\[
\int |u - v|_t \, dx = - \int \text{sgn}(u - v)(\overline{v} - v)u_x \, dx - \int v|u - v|_x \, dx.
\]

Integrating by parts, we get

\[
\frac{d}{dt} \int |u - v| \, dx = - \int \text{sgn}(u - v)(u_x - v_x)(u - v) \, dx + \int v_x|u - v|_x \, dx.
\]

Integrating by parts, we get

\[
\int |u - v| \, dx = - \int \text{sgn}(u - v)(u_x - v_x)(u - v) \, dx + \int v_x|u - v|_x \, dx.
\]

We estimate \( \|\overline{v}\|_{L^1} \) and \( \|u_x\|_\infty \):

\[
|\overline{v}(x,t)| = \left| \int g'_\alpha(x - y)v(y,t) \, dy \right| \\
\leq \frac{1}{\alpha} \|g'\|_{L^1(\mathbb{R})} \|v(\cdot,t)\|_{\infty} \\
\leq \frac{1}{\alpha} \|g'\|_{L^1(\mathbb{R})} \|v_0\|_{\infty} \\
\leq \frac{k_1 M_1}{\alpha}
\]

where we used that (1) preserves the maximum in the second inequality. We continue to estimate \( \|u_x\|_{\infty} \). Using the reparametrization of \( u \) in terms of the initial data (6) we find,

\[
\frac{d}{dx} u(x,t) = \frac{d}{dx} (u(\eta^{-1}(x,t))) = (u^0)'(\eta^{-1}(x,t)) \frac{d}{dx}\eta^{-1}(x,t) \\
= (u^0)'(\eta^{-1}(x,t)) \frac{1}{\eta_X(\eta^{-1}(x,t),t)}
\]

and therefore

\[
\|u_x\|_{\infty} \leq \|(u_0)'\|_{\infty} \sup_{X \in \mathbb{R}} \left| \frac{1}{\eta_X(X,t)} \right|
\]

An estimate on \( |\eta_X(X,t)| \) is derived in [1] for the purpose of proving global existence of the solution of (1). Specifically, in that paper, the ordinary differential equation (5) describing the evolution of the characteristics is rewritten as

\[
\frac{d}{dt} \eta(X,t) = \int g_\alpha(\eta(X,t) - \eta(Y,t))u_0(Y)\gamma(Y,t) \, dY.
\]

Then taking derivative with respect to \( X \),

\[
\frac{d}{dt} \eta_X(X,t) = \eta_X(X,t) \int g'_\alpha(\eta(X,t) - \eta(Y,t))u_0(Y)\gamma(Y,t) \, dY,
\]

dividing the equation by \( \eta_X(X,t) \) and integrating in \( t \) one obtains,

\[
|\eta_X(X,t)| = \exp\left( \int_0^t \int g_\alpha(\eta(X,s) - \eta(Y,s))(u(0)')(Y) \, dY \, ds \right)
\]
This can be bounded from above as follows:
\[
\left| \int_0^t \int_\mathbb{R} g_\alpha(\eta(X, s) - \eta(Y, s))(u^0)'(Y) \, dY \, ds \right| \leq t \| g_\alpha \|_\infty \| (u_0)' \|_{L^1(\mathbb{R})} \leq \frac{t k_2}{\alpha} \| (u_0)' \|_{L^1(\mathbb{R})},
\]
and therefore,
\[
|\eta_X(X, t)| \geq \exp\left(-\frac{t k_2}{\alpha} M_3\right).
\]
Combining this estimate with (11), we obtain
\[
\| u_x \|_\infty \leq \| (u_0)' \|_\infty \exp\left(\frac{t k_2}{\alpha} M_3\right).
\] (13)

Having estimated \( \| u_x \|_\infty \) and \( \| \pi_x \|_\infty \) by (10) and (13), we can apply Gronwall’s inequality in (9) to obtain the result.

Obviously, the estimate of Theorem 2 is of no use in the limit \( \alpha \to 0 \). One rather wishes to have an estimate of the form
\[
\| u^\alpha(\cdot, t) - v^\alpha(\cdot, t) \|_{L^1(\mathbb{R})} \leq C \| u_0 - v_0 \|_{L^1(\mathbb{R})},
\] (14)
where \( C \) is a constant not depending on \( \alpha \). This would then imply the same stability for the limits \( u \) and \( v \) of the sequences \( (u^\alpha)_{\alpha > 0} \) and \( (v^\alpha)_{\alpha > 0} \) (this would follow automatically if the limit functions satisfied the entropy inequality (7)). Unfortunately, this is not possible, as the following counterexample shows. For \( K \geq 3 \), we let
\[
v_0(x) = \begin{cases} 
0, & x < 0, \\
1, & x \in [0, K], \\
0, & x > K,
\end{cases}
\] (15)
and
\[
u_{0, \delta}(x) = \begin{cases} 
0, & x < 0, \\
0.5, & x \in [0, \delta), \\
1, & x \in [\delta, K], \\
0, & x > K,
\end{cases}
\] (16)
for some \( \delta > 0 \). We have \( \| v_0 - u_{0, \delta} \|_{L^1(\mathbb{R})} = \delta/2 \) and thus \( u_{0, \delta} \to v_0 \) as \( \delta \to 0 \) in \( L^1(\mathbb{R}) \). We calculate the solutions to (1a), (1b) augmented with (15), (16) respectively. Since we know from the global existence result that the characteristics of the equation do not cross, we only need to find the speeds of the discontinuities, then we know that the solutions are constant inbetween. We will denote the discontinuities by \( s_1 \) and \( s_2 \) (from the left to the right) for the first problem and by \( s_1 \), \( s_2 \) and \( s_3 \) for the second problem. In the first problem, we have
\[
\frac{d}{dt} s_1(t) = \int_0^{(s_2 - s_1)(t)} g_\alpha(y) \, dy, \quad s_1(0) = 0,
\]
\[
\frac{d}{dt} s_2(t) = \int_{(s_1 - s_2)(t)}^0 g_\alpha(y) \, dy, \quad s_2(0) = K,
\]
and therefore
\[
\frac{d}{dt} (s_2 - s_1)(t) = 0, \quad (s_2 - s_1)(0) = K.
\]
Thus \((s_2 - s_1)(t) = K\) for all \(t\) and the discontinuities satisfy
\[
\frac{d}{dt}s_i(t) = \int_0^{K/\alpha} g(y) \, dy =: G(\alpha), \quad i = 1, 2. \tag{17}
\]
Hence, the solution of \((1a), (1b)\) and \((15)\) is
\[
v^\alpha(x, t) = \begin{cases} 
0, & x < G(\alpha)t, \\
1, & G(\alpha)t < x < G(\alpha)t + K, \\
0, & x > G(\alpha)t + K.
\end{cases} \tag{18}
\]
We have that \(\lim_{\alpha \to 0} G(\alpha) = 1/2\). Therefore,
\[
v(x, t) = \lim_{\alpha \to 0} v^\alpha(x, t) = \begin{cases} 
0, & x < t/2, \\
1, & t/2 < x < t/2 + K, \\
0, & x > t/2 + K.
\end{cases}
\]
For the problem with initial data \((16)\), the solution cannot be calculated explicitly for arbitrary filters. Therefore, we rather provide some estimates which suffice our purposes. We assume from now on \(\delta < 1\) and \(\alpha < 1\). For the discontinuities of \(u^\alpha\), we have,
\[
\frac{d}{dt}s_1(t) = \frac{1}{2} \int_0^{(s_2 - s_1)(t)} g_\alpha(y) \, dy + \int_{(s_2 - s_1)(t)}^{(s_3 - s_1)(t)} g_\alpha(y) \, dy, \quad s_1(0) = 0,
\]
\[
\frac{d}{dt}s_2(t) = \frac{1}{2} \int_0^{(s_1 - s_2)(t)} g_\alpha(y) \, dy + \int_{(s_1 - s_2)(t)}^{(s_3 - s_2)(t)} g_\alpha(y) \, dy, \quad s_2(0) = \delta,
\]
\[
\frac{d}{dt}s_3(t) = \frac{1}{2} \int_{(s_1 - s_3)(t)}^{(s_2 - s_3)(t)} g_\alpha(y) \, dy + \int_{(s_2 - s_3)(t)}^{(s_3 - s_1)(t)} g_\alpha(y) \, dy, \quad s_3(0) = K,
\]
Set \(d_{l,m,r}\) to be the distances between the discontinuities, that is, \(d_l := s_2 - s_1\), \(d_r := s_3 - s_2\) and \(d_m = d_l + d_r\). They satisfy
\[
\frac{d}{dt}d_l(t) = \int_0^{(s_3 - s_2)(t)} g_\alpha(y) \, dy - \int_{(s_2 - s_1)(t)}^{(s_3 - s_1)(t)} g_\alpha(y) \, dy \geq 0, \tag{19a}
\]
\[
\frac{d}{dt}d_r(t) = \frac{1}{2} \int_{(s_1 - s_2)(t)}^{(s_2 - s_3)(t)} g_\alpha(y) \, dy - \frac{1}{2} \int_{(s_1 - s_2)(t)}^{(s_3 - s_2)(t)} g_\alpha(y) \, dy \geq -\frac{1}{4}, \tag{19b}
\]
\[
\frac{d}{dt}d_m(t) = \frac{1}{2} \int_{(s_1 - s_3)(t)}^{(s_2 - s_3)(t)} g_\alpha(y) \, dy + \int_{(s_2 - s_3)(t)}^{(s_3 - s_1)(t)} g_\alpha(y) \, dy
\]
\[
- \frac{1}{2} \int_0^{(s_2 - s_1)(t)} g_\alpha(y) \, dy + \int_{(s_2 - s_1)(t)}^{(s_3 - s_1)(t)} g_\alpha(y) \, dy \geq 0, \tag{19c}
\]
(we have used that \(g\) is a non-increasing function for the inequalities). So for \(t < 4(K - \delta)\) the discontinuities will not meet, independently of \(\alpha\). For \(d_l\) we have in addition \(\frac{d}{dt}d_l \leq 1/2\). To find a lower bound on \(d_l\), we compute (omitting the
variable $t$):

$$\frac{d}{dt} d_1 = \int_0^{s_2-s_1} g_\alpha(y) \, dy - \int_{s_3-s_2}^{s_3-s_1} g_\alpha(y) \, dy$$

$$= \int_0^{(s_2-s_1)/\alpha} g(y) \, dy - \int_{(s_3-s_1)/\alpha}^{(s_3-s_2)/\alpha} g(y) \, dy$$

$$= \int_0^{d_1/\alpha} g(y) \, dy - \int_{d_1/\alpha}^{(d_1+\delta)/\alpha} g(y) \, dy$$

$$\geq \int_0^{d_1/\alpha} g(y) \, dy - \int_{(K-\delta-t/4+\delta)/\alpha}^{(K-\delta-t/4)/\alpha} g(y) \, dy,$$

(20)

where we have used the bound (19b) in the last inequality. We assume now that $t \leq 4/3(K-2)$. This implies that $d_1 \leq K - \delta - t/4$ and that we can estimate (20) further by

$$\frac{d}{dt} d_1 \geq \int_0^{d_1/\alpha} g(y) \, dy - \int_{d_1/\alpha}^{(K-t/4-\delta+d_1)/\alpha} g(y) \, dy,$$

$$\geq \int_0^{d_1/\alpha} g(y) \, dy - \int_{d_1/\alpha}^{\infty} g(y) \, dy,$$

$$\geq \int_0^{\delta/\alpha} g(y) \, dy - \int_{\delta/\alpha}^{\infty} g(y) \, dy,$$

(21)

Now we fix $\delta > 0$ and choose $\alpha(\delta)$ so small that the last expression in (21) becomes $\geq 1/8$ for all $\alpha \leq \alpha(\delta)$. This is possible, since $g$ is non-increasing and integrable. Then $d_1$ satisfies

$$d_1(t) \geq \delta + t/8 \geq t/8, \quad \forall \alpha \leq \alpha(\delta).$$

(22)

We have chosen $t$ small enough such that the discontinuities do not interact. Thus, the solution $u^\alpha$ of the convectively filtered Burgers’ equation with initial data (16) has for all $0 < \alpha < 1$ and $t \leq 4/3(K-2)$ the form

$$u^\alpha,\delta(x,t) = \begin{cases} 
0, & x < s_1^\alpha(t), \\
0.5, & s_1^\alpha(t) < x < s_2^\alpha(t), \\
1, & s_2^\alpha(t) < x < s_3^\alpha(t), \\
0, & x > s_3^\alpha(t), 
\end{cases}$$

where $s_1^\alpha(t) + 2\delta \leq s_2^\alpha(t) + \delta \leq s_3^\alpha(t)$ uniformly in $\alpha$. So the limit function $u^\delta = \lim_{\alpha \to 0} u^\alpha,\delta$ has the form

$$u^\delta(x,t) = \lim_{\alpha \to 0} u^\alpha,\delta(x,t) = \begin{cases} 
0, & x < s_1^0(t), \\
0.5, & s_1^0(t) < x < s_2^0(t), \\
1, & s_2^0(t) < x < s_3^0(t), \\
0, & x > s_3^0(t), 
\end{cases}$$

for some $s_1^0(t) < s_2^0(t) < s_3^0(t)$. In addition, we know from (22) that $s_2^0(t) - s_1^0(t) > t/8$. In this region, $u^\delta$ takes the value 0.5 whereas $v^\cdot$ is either 0 or 1.

Consequently, their difference in the $L^1$-norm satisfies

$$\|u^\delta(\cdot,t) - v(\cdot,t)\|_{L^1(\mathbb{R})} \geq \frac{t}{16},$$

(23)
while
\[ \|u_{0,\delta} - v_0\|_{L^1(\mathbb{R})} \leq \frac{\delta}{2}. \]  
(24)

The estimates (23), (24) are valid for arbitrary small \( \delta > 0 \). This shows that an estimate of the form (14) cannot be achieved for the solutions to the convectively filtered Burgers' equation, since it would imply the same bound for the limit functions obtained when letting \( \alpha \to 0 \). This cannot hold owing to (23) and (24).

The missing \( L^1 \)-stability makes it difficult to find a numerical scheme which converges independently of the parameter \( \alpha \), since the natural norm with respect to which we would expect convergence is \( L^1 \).

**Remark 1.** It would be interesting to investigate whether it is also possible to find a counterexample to (14) with continuous initial data or whether on the opposite a uniform estimate of the form
\[ \|u^\alpha(\cdot, t) - v^\alpha(\cdot, t)\|_{L^1(\mathbb{R})} \leq C\|u_0 - v_0\|_{\infty} \]
is possible. This could be the subject of further investigations.

**Acknowledgments.** F. Weber thanks the Seminar for Applied Mathematics, ETH, for the financial support making it possible to attend the HYP-conference.

**REFERENCES**


E-mail address: nilshr@math.uio.no
E-mail address: franziska.weber@cma.uio.no
elman Approximation of the Gross-Pitaevskii Equation for General Initial Data

Chi-Kun Lin
Department of Applied Mathematics and
Center of Mathematical Modeling and Scientific Computing
National Chiao Tung University
Hsinchu 30010, Taiwan

Kung-Chien Wu
Department Mathematics
National Kaohsiung Normal University
Kaohsiung 82444, Taiwan

Abstract. We perform a rigorous analysis of the anelastic approximation for the Gross-Pitaevskii equation with \( x \)-dependent chemical potential. For general initial data and periodic boundary condition, we show that as \( \varepsilon \to 0 \), equivalently the Planck constant tends to zero, the density \(|\psi^\varepsilon|^2\) converges toward the chemical potential \( \rho_0(x) \) and the velocity field converges to the anelastic system.

1. Introduction. We consider a superfluid governed by the Gross-Pitaevskii equation [13]

\[
i\hbar \partial_t \psi + \frac{\hbar^2}{2m} \Delta \psi - (V_0|\psi|^2 - E)\psi = 0,
\]

appropriate to a weakly interacting Bose gas. This is a nonlinear Schrödinger equation for a single-particle wave function \( \psi(x, t) \) governing an assembly of bosons of mass \( m \), with \( V_0 \) the strength of the \( \delta \)-function interaction potential between the bosons, \( E \) the chemical potential, and \( \hbar \) the Planck constant. In this paper, we will assume the unit strength \( V_0 = 1 \) and the chemical potential \( E \) is a function of position, i.e., \( E = \rho_0(x) \);

\[
i\hbar \partial_t \psi + \frac{\hbar^2}{2m} \Delta \psi - (|\psi|^2 - \rho_0(x))\psi = 0.
\]

In order to investigate the singular limit, we introduce the scaled variables \( \tilde{t} = \varepsilon t \) and \( \tilde{x} = x \) with \( 0 < \varepsilon \ll 1 \), and the Planck constant is also rescaled as \( \hbar = \varepsilon^{1+\alpha} \), where \( 0 < \alpha < \infty \). Then after dropping the tilde, (1.2) becomes

\[
i\varepsilon^\alpha \partial_t \psi^\varepsilon + \frac{\varepsilon^{2\alpha}}{2} \Delta \psi^\varepsilon - \frac{1}{\varepsilon^{2\alpha}} (|\psi^\varepsilon|^2 - \rho_0(x)) \psi^\varepsilon = 0.
\]

The initial condition is complemented by

\[
\psi^\varepsilon(x, 0) = \psi_0^\varepsilon(x), \quad x \in \Omega.
\]
According to the conservative quantities we have the hydrodynamical variables: density (or charge) \( \rho^\varepsilon \), momentum (or current) \( J^\varepsilon \), and energy \( e^\varepsilon \) given respectively by

\[
\rho^\varepsilon = |\psi^\varepsilon|^2, \quad J^\varepsilon = \frac{i}{2} \varepsilon e^\alpha (\psi^\varepsilon \nabla \overline{\psi^\varepsilon} - \overline{\psi^\varepsilon} \nabla \psi^\varepsilon),
\]

\[
e^\varepsilon = \frac{1}{2} \varepsilon^2 |\nabla \psi^\varepsilon|^2 + \frac{1}{2 \varepsilon^2} (|\psi^\varepsilon|^2 - \rho_0)^2.
\]

Under the above definition, we can define the initial conditions as follows: initial charge \( \rho^\varepsilon_0(x) = \rho^\varepsilon(x,0) = |\psi^\varepsilon_0(x)|^2 \) and initial momentum (current) \( J^\varepsilon_0(x) = J^\varepsilon_0(x,0) \). The associated local conservations of mass(charge), momentum(current) and energy in terms of hydrodynamical variables are given by

\[
\partial_t \rho^\varepsilon + \nabla \cdot J^\varepsilon = 0, \tag{1.6}
\]

\[
\partial_t J^\varepsilon + \nabla \cdot \left( \frac{J^\varepsilon \otimes J^\varepsilon}{\rho^\varepsilon} \right) + \frac{1}{\varepsilon^2} \rho^\varepsilon \nabla (\rho^\varepsilon - \rho_0) = \frac{\varepsilon^2}{4} \nabla \cdot \left( \rho^\varepsilon \nabla^2 \log \rho^\varepsilon \right), \tag{1.7}
\]

\[
\partial_t e^\varepsilon + \nabla \cdot \left( e^\varepsilon \frac{J^\varepsilon}{\rho^\varepsilon} + \frac{1}{\varepsilon^2} J^\varepsilon (\rho^\varepsilon - \rho_0) \right) = \frac{\varepsilon^2}{4} \nabla \cdot \left( \rho^\varepsilon \Delta \rho^\varepsilon - \nabla \cdot J^\varepsilon \frac{\nabla \rho^\varepsilon}{\rho^\varepsilon} \right). \tag{1.8}
\]

Equations (1.6)–(1.7) comprise a closed system governing \( \rho^\varepsilon \) and \( J^\varepsilon \), which have the form of a perturbation of the compressible Euler equations with an extra potential \( \rho_0(x) \). Formally, letting \( \varepsilon \to 0 \), the uniform boundedness of energy \( e^\varepsilon \) and strict convexity of the potential will imply \( \rho^\varepsilon \to \rho_0 \). We also assume \( J^\varepsilon \to \rho_0 v \) for some proper function \( v \). The limit equations will then be the anelastic system

\[
\partial_t (\rho_0 v) + \nabla \cdot (\rho_0 v \otimes v) + \rho_0 \nabla \pi = 0, \tag{1.9}
\]

where the pressure \( \pi \) is the formal limit of \( \frac{\varepsilon^2}{4} \rho^\varepsilon \nabla^2 \log \rho^\varepsilon \). When \( \rho_0 \) is a constant, the anelastic system (1.9) will reduce to the incompressible Euler equation,

\[
\partial_t v + v \cdot \nabla v + \nabla \pi = 0, \tag{1.10}
\]

The reader is referred to [9] for the detail study of the incompressible Euler equations. For \( x \in \Omega \subset \mathbb{R}^2 \), (1.9) can be served as the lake equations which may be seen as the low Froude number limit of the usual inviscid shallow water equations when the initial height converges to a nonconstant function depending on the space variable. Here we have to interpret \( \pi(x,t) \) as the surface height variation (see [3, 5, 8] and references therein). For the viscous shallow water equations and the convergence to the quasi-geostrophic model we will refer to [1]. Note that the chemical potential \( \rho_0(x) \) plays the role of depth of the basin at location \( x \) in the lake equations. However, the above formal discussion of the singular limit \( \varepsilon \to 0 \) can not be made directly in (1.6)–(1.7) since \( \log \rho^\varepsilon \) may be undefined. Indeed, we can write the dispersive term on the right hand side of (1.7) in different ways:

\[
\frac{1}{4} \nabla \cdot (\rho^\varepsilon \nabla^2 \log \rho^\varepsilon) = \frac{1}{2} \rho^\varepsilon \nabla \left( \frac{\Delta \sqrt{\rho^\varepsilon}}{\sqrt{\rho^\varepsilon}} \right) = \frac{1}{4} \Delta \nabla \rho^\varepsilon - \nabla \cdot (\nabla \sqrt{\rho^\varepsilon} \otimes \nabla \sqrt{\rho^\varepsilon}). \tag{1.11}
\]
Instead of the momentum and energy equations (1.7)–(1.8), we prefer to represent them as

\[
\partial_t J^\varepsilon + \frac{1}{2} \varepsilon^{2\alpha} \nabla \cdot \left( \left( \nabla \psi^\varepsilon \otimes \nabla \psi^\varepsilon + \nabla \bar{\psi}^\varepsilon \otimes \nabla \bar{\psi}^\varepsilon \right) - \nabla^2 (|\psi^\varepsilon|^2) \right) + \frac{1}{2\varepsilon^2} \nabla (\rho^\varepsilon - \rho_0)^2 + \frac{1}{\varepsilon^2} \rho_0 \nabla (\rho^\varepsilon - \rho_0) = 0,
\]

and

\[
\partial_t e^\varepsilon - \frac{1}{2} \varepsilon^{2\alpha} \nabla \cdot \left( \nabla \psi^\varepsilon \partial_t \psi^\varepsilon + \nabla \bar{\psi}^\varepsilon \partial_t \bar{\psi}^\varepsilon \right) = 0,
\]

which are more suitable for the estimates of the modulated energy as we shall see in the next section.

The question of anelastic-type limits in fluid dynamics has received considerable attention recently. As is well-known, an anelastic approximation is a filtering approximation for the equations of motion that eliminates sound waves by assuming that the flow has velocities and phase speeds much smaller than the speed of sound. This approximation has been used to model astrophysical and geophysical fluids [10, 12]. The rigorous derivation of the anelastic limit starting from the diffusive systems have been recently studied from a mathematical point of view in [3, 6, 11, 14]. We will refer to [4] for the compressible Euler-type systems and [7] for the study of the asymptotic behavior of a fluid submitted to a strong external magnetic field.

The rest of the paper is organized as follows. In section 2, in addition to the main theorem, we also present the weighted Helmholtz decomposition which will play the essential role for the anelastic limit. Section 3 is devoted to the sketch proof of the main theorem.

2. Main Theorem and Function Space. Since \( \rho_0 \) is strictly positive, we can define the weighted space as follows. Let \( \Omega \subset \mathbb{R}^n \) be open and \( \sigma(x) = \frac{1}{\rho_0(x)} \) be the weighted function, then the weighted square integrable space \( L^2_u(\frac{dx}{\rho_0}) = L^2_u(\Omega) \) consists of all measurable functions \( f \) that satisfy

\[
\int_\Omega |f(x)|^2 \sigma(x) dx < \infty.
\]

The resulting \( L^2_u(\Omega) \)-norm of \( f \) is defined by

\[
\|f\|_{L^2_u(\Omega)} = \left( \int_\Omega |f(x)|^2 \sigma(x) dx \right)^{1/2}.
\]

(2.1)

The space \( L^2_u(\Omega) \) is naturally equipped with the following inner product:

\[
\langle f, g \rangle_\sigma = \int_\Omega f \cdot g \sigma(x) dx.
\]

(2.2)

The weighted Sobolev space \( H^1_u(\Omega) \) consists of all functions \( f \) with weak derivatives \( Df \) satisfying

\[
\|f\|_{H^1_u(\Omega)} = \left( \int_\Omega (|f(x)|^2 + |Df(x)|^2) \sigma(x) dx \right)^{1/2} < \infty.
\]

(2.3)

To derive the anelastic system we need to introduce the space \( \mathbb{H}_{\rho_0}[L^2_u(\Omega)] \) defined by

\[
\mathbb{H}_{\rho_0}[L^2_u(\Omega)] = \{ \rho_0 v \in L^2_u(\Omega) \mid \text{div} (\rho_0 v) = 0 \}.
\]

(2.4)
Since \( \mathbb{H}_{\rho_0}^\perp[L^2_\sigma] \) is a closed subspace of \( L^2_\sigma \), then by projection theorem the Hilbert space \( L^2_\sigma(\Omega) \) admits an orthogonal projection

\[
L^2_\sigma(\Omega) = \mathbb{H}_{\rho_0}[L^2_\sigma(\Omega)] \oplus \mathbb{H}_{\rho_0}^\perp[L^2_\sigma(\Omega)],
\]

where the orthogonal complement \( \mathbb{H}_{\rho_0}^\perp[L^2_\sigma(\Omega)] \) is given by

\[
\mathbb{H}_{\rho_0}^\perp[L^2_\sigma(\Omega)] = \left\{ \rho_0 \nabla \Psi \mid \Psi \in H^1_\sigma(\Omega), \int_\Omega \Psi dx = 0 \right\}.
\]

In the sequel, we will consider the periodic domain \( \Omega = \mathbb{T}^n \). Associated with the orthogonal projection \( (2.5) \) is the weighted Helmholtz decomposition in the form [6]

\[
\mathbb{H}_{\rho_0} f = f - \rho_0 \nabla \Psi, \quad \mathbb{H}_{\rho_0}^\perp f = \rho_0 \nabla \Psi,
\]

where \( \Psi \in H^1_\sigma(\mathbb{T}^n) \) is the unique solution of the problem

\[
\int_{\mathbb{T}^n} \rho_0 \nabla \Psi \cdot \nabla \eta dx = \int_{\mathbb{T}^n} f \cdot \nabla \eta dx, \quad \forall \eta \in H^1_\sigma(\mathbb{T}^n).
\]

That is, \( \Psi \) is a weak solution of the uniformly elliptic partial differential equation satisfying the zero mean condition

\[
\nabla \cdot (\rho_0 \nabla \Psi) = \nabla \cdot f, \quad \int_{\mathbb{T}^n} \Psi dx = 0.
\]

Note that \( \mathbb{H}_{\rho_0} \) is an orthogonal projection on \( L^2_\sigma \), and the two projectors \( \mathbb{H}_{\rho_0} f \) and \( \mathbb{H}_{\rho_0}^\perp f \) are orthogonal with respect to the inner product \( \langle \cdot, \cdot \rangle_\sigma \). The reader is referred to [2, 6, 9] for mathematical properties of such projectors.

In order to make sure that (1.3) and the limit systems (2.13) and (3.20) are well-defined, we shall from now on impose the following conditions on \( \psi_0^* \), \( \rho_0 \), \( J_0 \), \( v_0 \) and \( w_0 \):

(A1) \( \psi_0^* \in H^{\frac{n}{2}+3}(\mathbb{T}^n; \mathbb{C}) \), this will guarantee the local existence and uniqueness of classical solution of the Gross-Pitaevskii equation (1.3).

(A2) \( \rho_0 \in C^s(\mathbb{T}^n) \), \( s > \frac{n}{2} + 1 \), \( \rho_0 \geq c > 0 \), the initial kinetic, potential and quantum energies satisfy

\[
\frac{1}{\sqrt{\rho_0}} J_0 \to \frac{1}{\sqrt{\rho_0}} J_0 \quad \text{in} \quad L^2(\mathbb{T}^n),
\]

\[
\frac{\rho_0^c - \rho_0}{\varepsilon} \to \varphi_0 \quad \text{in} \quad L^2(\mathbb{T}^n),
\]

\[
\varepsilon^\alpha \nabla \sqrt{\rho_0} \to 0 \quad \text{in} \quad L^2(\mathbb{T}^n).
\]

(A3) \( J_0 = \mathbb{H}_{\rho_0} J_0 \oplus \mathbb{H}_{\rho_0}^\perp J_0 = \rho_0 v_0 + \rho_0 \nabla w_0 \), where \( \sqrt{\rho_0} v_0 \in H^s(\mathbb{T}^n) \) and \( (\varphi_0, \sqrt{\rho_0} \nabla w_0) \in H^s(\mathbb{T}^n) \) for \( s > \frac{n}{2} + 3 \). This condition will guarantee the local existence and uniqueness of smooth solution of the anelastic system (2.13) and the oscillating part (3.20).

**Theorem 2.1.** Let \( \alpha > 0 \) and \( \psi^* \) be the solution of the Gross-Pitaevskii equation (1.3) with \( \psi_0^* \) satisfying the assumptions (A1)–(A3). There then exists \( T > 0 \) such that

\[
\rho^* \to \rho_0 \quad \text{strongly in} \quad L^\infty ([0, T]; L^2(\mathbb{T}^n)),
\]

\[
J^* \to \rho_0 v \quad \text{weakly * in} \quad L^\infty ([0, T]; L^{4/3}(\mathbb{T}^n)).
\]
where \( v \) solves the anelastic system
\[
\begin{aligned}
\partial_t (\rho_0 v) + \nabla \cdot (\rho_0 v \otimes v) + \rho_0 \nabla \pi &= 0, \\
v(x,0) = v_0(x), \quad \nabla \cdot (\rho_0 v) &= 0.
\end{aligned}
\tag{2.13}
\]

3. Sketch Proof of the Main Theorem. We divide the proof into several steps:

**Step 1. Spectral analysis of the wave group.** Analogous to the low Mach number limit in fluid dynamics we consider the perturbation of \( \rho^\varepsilon \) near the equilibrium \( \rho_0 \)
\[
\rho^\varepsilon = \rho_0 + \varepsilon \phi^\varepsilon,
\tag{3.1}
\]
i.e., \( \phi^\varepsilon \) is the density fluctuation. By weighted Helmholtz decomposition the current \( J^\varepsilon \) can be rewritten as
\[
J^\varepsilon = \mathbb{H}_{\rho_0} J^\varepsilon + \mathbb{H}_{\rho_0}^\perp J^\varepsilon = \rho_0 u^\varepsilon + \rho_0 \nabla w^\varepsilon.
\]
Applying the operator \( \mathbb{H}_{\rho_0}^\perp \) to the momentum equation (1.12) and using the weighted incompressibility \( \nabla \cdot (\rho_0 u^\varepsilon) = 0 \), the equations (1.6) and (1.12) can be rewritten in terms of \( \phi^\varepsilon \) and \( w^\varepsilon \) as
\[
\begin{aligned}
&\frac{\partial}{\partial t} \phi^\varepsilon + \frac{1}{\varepsilon} \nabla \cdot (\rho_0 \nabla w^\varepsilon) = 0, \\
&\frac{\partial}{\partial t} (\sqrt{\rho_0} \nabla w^\varepsilon) + \frac{1}{\varepsilon} \sqrt{\rho_0} \nabla \phi^\varepsilon = \frac{1}{\sqrt{\rho_0}} F^\varepsilon,
\end{aligned}
\tag{3.2}
\]
where
\[
F^\varepsilon = -\frac{\varepsilon^2}{2} \mathbb{H}_{\rho_0} \nabla \cdot (\nabla \psi^\varepsilon \otimes \nabla \psi^\varepsilon + \nabla \overline{\psi^\varepsilon} \otimes \nabla \overline{\psi^\varepsilon}) - \frac{1}{2} \mathbb{H}_{\rho_0} \nabla (\psi^\varepsilon)^2 + \frac{\varepsilon^2}{4} \mathbb{H}_{\rho_0} \nabla \Delta \rho^\varepsilon.
\]
It is obvious from (3.2) that \( \partial_t \phi^\varepsilon \) and \( \partial_t (\sqrt{\rho_0} \nabla w^\varepsilon) \) are of order \( O(1/\varepsilon) \) and are highly oscillatory as \( \varepsilon \to 0 \). Therefore we have to introduce the wave group in order to filter out the fast oscillating wave. Let \( L(\tau) = e^{\tau L} \), \( \tau \in \mathbb{R} \) be the evolution group associated with the operator \( L \) which is defined, according to (3.2), on \( D = L^2(\mathbb{T}^n) \times \{ h = \sqrt{\rho_0} \nabla w : h \in L^2(\mathbb{T}^n) \} \) by
\[
L \left( \frac{\phi}{\sqrt{\rho_0} \nabla w} \right) = -\left( \nabla \cdot (\rho_0 \nabla w) \right) \sqrt{\rho_0} \nabla \phi.
\tag{3.3}
\]
The eigenvalues and the associated eigenfunctions of \( L \) are given by
\[
\left\{ i \sqrt{k_j}, \left( \frac{i}{\sqrt{k_j}} \frac{\chi_j}{\sqrt{\rho_0} \nabla \chi_j} \right) \right\} \quad \text{and} \quad \left\{ -i \sqrt{k_j}, \left( \frac{-i}{\sqrt{k_j}} \frac{\chi_j}{\sqrt{\rho_0} \nabla \chi_j} \right) \right\}.
\tag{3.4}
\]
where \( (k_j, \chi_j) \) is the spectrum of the uniformly elliptic operator
\[
- \nabla \cdot (\rho_0 \nabla).
\tag{3.5}
\]
Since \( V \in \mathbb{D} \) is real, we can represent it as the eigenfunction expansion
\[
V = \sum_{j=1}^{\infty} a_j^+ \left( \frac{i}{\sqrt{k_j}} \frac{\chi_j}{\sqrt{\rho_0} \nabla \chi_j} \right) + a_j^- \left( \frac{-i}{\sqrt{k_j}} \frac{\chi_j}{\sqrt{\rho_0} \nabla \chi_j} \right).
\tag{3.6}
\]
respectively a linear and a bilinear forms of $R^v$ where $v$ is the strong solution of the anelastic system (2.13). Here $Q_1$ and $Q_2$ are respectively a linear and a bilinear forms of $V$ defined by

$$Q_1(u, V) = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau \mathcal{L}(-s) \left( \frac{1}{\sqrt{\rho_0}} \frac{\partial}{\partial \rho} B_1(u, V) \right) \, ds$$

and

$$Q_2(V, V) = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau \mathcal{L}(-s) \left( \frac{1}{\sqrt{\rho_0}} \frac{\partial}{\partial \rho} B_2(V, V) \right) \, ds$$

For convenience, we denote $\mathcal{L}_1(\tau)$ the first component and $\mathcal{L}_2(\tau)$ the rest $n$ components of $\mathcal{L}(\tau)$. It is easy to see that the operator $\mathcal{L}(\tau)$ is isometry on $L^2(\mathbb{T}^n)$ and is bounded on $H^s(\mathbb{T}^n)$ for all $s \in \mathbb{R}$ and for all $\tau \in \mathbb{R}$.

In the sequel, we shall denote

$$U^\varepsilon = \left( \frac{\varphi^\varepsilon}{\sqrt{\rho_0}} \right), \quad V^\varepsilon = \mathcal{L}\left( \frac{-t}{\varepsilon} \right) \left( \frac{\varphi^\varepsilon}{\sqrt{\rho_0}} \right).$$

With this notation, we can rewrite (3.2) as

$$\partial_t U^\varepsilon = \frac{1}{\varepsilon} LU^\varepsilon + \frac{1}{\sqrt{\rho_0}} \tilde{F}^\varepsilon,$$

or equivalently

$$\partial_t V^\varepsilon = \mathcal{L}\left( \frac{-t}{\varepsilon} \right) \frac{1}{\sqrt{\rho_0}} \tilde{F}^\varepsilon,$$

where $\tilde{v}$ denotes the column vector $(0, v)^T$.

**Step 2. Analysis of the oscillating equation.** Let $V, V_1, V_2 \in D \cap L^2(\mathbb{T}^n)$, and $u$ be any vector in $\mathbb{R}^n$ such that $\nabla \cdot (\rho_0 u) = 0$, we define the linear form $B_1$ by

$$B_1(u, V) = \nabla \cdot \left( \sqrt{\rho_0} u \otimes \mathcal{L}_2(\frac{t}{\varepsilon}) V + \sqrt{\rho_0} \mathcal{L}_2(\frac{t}{\varepsilon}) V \otimes u \right)$$

and the bilinear form $B_2$ by

$$B_2(V_1, V_2) = \frac{1}{2} \nabla \cdot \left( \mathcal{L}_2(\frac{t}{\varepsilon}) V_1 \otimes \mathcal{L}_2(\frac{t}{\varepsilon}) V_2 + \mathcal{L}_2(\frac{t}{\varepsilon}) V_2 \otimes \mathcal{L}_2(\frac{t}{\varepsilon}) V_1 \right)$$

$$+ \frac{1}{2} \nabla \cdot (\mathcal{L}_1(\frac{t}{\varepsilon}) V_1 \cdot \mathcal{L}_1(\frac{t}{\varepsilon}) V_2).$$

We can rewrite (3.8) as

$$\partial_t V^\varepsilon = \mathcal{L}\left( \frac{-t}{\varepsilon} \right) \left( \frac{1}{\sqrt{\rho_0}} \frac{\partial}{\partial \rho} B_1(u^\varepsilon, V^\varepsilon) + \frac{1}{\sqrt{\rho_0}} \frac{\partial}{\partial \rho} B_2(V^\varepsilon, V^\varepsilon) \right)$$

$$+ \frac{\varepsilon^{2\alpha}}{2} \mathcal{L}\left( \frac{-t}{\varepsilon} \right) \left( \frac{1}{\sqrt{\rho_0}} \frac{\partial}{\partial \rho} \nabla \left( \frac{\Delta \tilde{v}}{\sqrt{\rho_0}} \right) \right).$$

If we had sufficient compactness in space, then we could pass to the limit as $\varepsilon \to 0$ in (3.11) and obtain the following system for the oscillating parts:

$$\partial_t V + Q_1(v, V) + Q_2(V, V) = 0,$$

where $v$ is the strong solution of the anelastic system (2.13).
for any divergence free vector fields $u \in L^2(T^n)$ and any $V = (\phi, \sqrt{\rho_0} \nabla q)^t \in L^2(T^n)$. We have the following basic propositions:

**Proposition 1.** For any $u$ satisfying $\nabla \cdot (\rho_0 u) = 0$, and for all vectors $V$ and $V_j$, $j = 1, 2$, we have

$$
\int_{T^n} Q_1(u, V) \cdot V dx = 0 ,
$$

(3.15)

$$
\int_{T^n} Q_1(u, V_1) \cdot V_2 + Q_1(u, V_2) \cdot V_1 dx = 0 .
$$

(3.16)

**Proposition 2.** For any $u$ satisfying $\nabla \cdot (\rho_0 u) = 0$, and for all vectors $V$ and $V_j$, $j = 1, 2$, we have

$$
\lim_{\varepsilon \to 0} \int_0^t \int_{T^n} B_2(V_1, V_2) \cdot u dx ds = 0 ,
$$

(3.17)

$$
\int_{T^n} Q_2(V, V) \cdot V dx = 0 ,
$$

(3.18)

$$
\int_{T^n} Q_2(V_1, V_1) \cdot V_2 + 2 Q_2(V_1, V_2) \cdot V_1 dx = 0 .
$$

(3.19)

There are some difficulties in the crossing eigenvalues phenomena, thus, we need studying the resonant effect of the oscillating part $Q_2(V, V)$ and it can be overcome because of oscillation-cancelation.

**Step 3.** The modulated energy functional and uniform estimates. Let $V^0$ be the solution of the oscillating part:

$$
\partial_t V^0 + Q_1(v, V^0) + Q_2(V^0, V^0) = 0 \tag{3.20}
$$

with initial condition

$$
V^0(x, 0) = V_{in}^0(x) = (\phi_0, \sqrt{\rho_0} w_0)^t , \tag{3.21}
$$

where $v$ is the strong solution of the anelastic system (2.13).

We can define the modulated energy $H^\varepsilon(t)$ of (1.3) as

$$
H^\varepsilon(t) = \frac{1}{2} \int_{T^n} \left\| \varepsilon^{\alpha} \nabla - i \left[ v + \frac{1}{\sqrt{\rho_0}} L_2(\frac{t}{\varepsilon}) V^0 \right] \right\|^2 \psi^\varepsilon dx 
$$

$$
+ \frac{1}{2} \int_{T^n} |\psi^\varepsilon - L_1(\frac{t}{\varepsilon}) V^0|^2 dx .
$$

(3.22)

It is nature to check the evolution of the modulated energy $H^\varepsilon(t)$, using the conservation laws (1.6), (1.12)–(1.13) of the Gross-Pitaevskii equation and the limiting systems (anelastic system (2.13) and the oscillating part (3.20)), we have

$$
H^\varepsilon(t) \leq C \int_0^t H^\varepsilon(s) ds + o_\varepsilon(1) ,
$$

(3.23)

this means

$$
H^\varepsilon(t) \to 0 \quad \text{as} \quad \varepsilon \to 0 .
$$

It is easy to rewrite the modulated energy as

$$
H^\varepsilon(t) = \frac{\varepsilon^{2\alpha}}{2} \int_{T^n} |\nabla \sqrt{\rho_0} |^2 dx + \frac{1}{2} \int_{T^n} \left| \psi^\varepsilon - L_1(\frac{t}{\varepsilon}) V^0 \right|^2 dx 
$$

$$
+ \frac{1}{2} \int_{T^n} \frac{1}{\sqrt{\rho_0}} \left( J^\varepsilon - \rho_0 v - \frac{\rho_0}{\sqrt{\rho_0}} L_2(\frac{t}{\varepsilon}) V^0 \right)^2 dx .
$$

(3.24)
and our theorem can be deduced from (3.24).

**Acknowledgments.** C.K. Lin is supported by National Science Council of Taiwan under the grant NSC98-2115-M-009-004-MY3. K.C. Wu is supported by the Tsz-Tza Foundation (Taiwan), National Science Council under the grant 102-2115-M-017-004-MY2 (Taiwan) and ERC grant MATKIT (European Union). K.C. Wu would like to thank Professor Dr. Clément Mouhot for his kind invitation to visit Cambridge during the 2011-2013 academic years.

**REFERENCES**


[8] C. D. Levermore, M. Oliver and E.S. Titi, Global well-posedness for the lake equations, Physica D., 98(1996), 492–509.


*E-mail address: ckltn@math.nctu.edu.tw*

*E-mail address: kungchienwu@gmail.com*
COMPARATIVE STUDY OF HIGH-ORDER
POSITIVITY-PRESERVING WENO SCHEMES

Dmitry Kotov
Center for Turbulence Research
488 Escondido Mall Building 500
Stanford University
Stanford, CA 94305-3035, USA

H.C. Yee
NASA Ames Research Center
MS 258-5
Moffett Field, CA, 94035, USA

Bjorn Sjögreen
Lawrence Livermore National Laboratory
Box 808, L-422
Livermore, CA 94551-0808, USA

Abstract. The objective of this study is to compare the results obtained by
different non-positivity-preserving methods with the two recently developed
high-order positivity-preserving schemes. Several test cases are considered,
including classical Noh and Sedov 3D cases and also shock tube calculations
related to NASA experiments. Studies indicate that the positivity preserving
schemes produce more stable results than regular methods for the considered
test cases. However some of the regular methods may obtain slightly more
accurate results than the new methods for the certain problems.

1. Introduction. Let us consider compressible gas flow. Physically, the density
\( \rho \) and the pressure \( p \) should both be positive. In a standard conservative numerical
scheme, however, the computed internal energy is obtained by subtracting the
kinetic energy from the total energy, thus resulting in a computed \( p \) that may be
negative, for example, for problems in which the dominant energy is kinetic. Negative \( \rho \) may often emerge in computing blast waves. In such situations the computed
eigenvalues of the Jacobian will become imaginary. Consequently, the initial value
problem for the linearized system will be ill-posed. This explains why failure of
preserving-positivity of density or pressure may cause blow-ups of the numerical
algorithm. The adhoc methods in numerical strategy by modifying the computed
negative density and/or computed negative pressure to be positive are neither a
conservative cure nor a stable solution. Conservative positivity-preserving schemes
are more appropriate for such flow problems.

The ideas of [17] and [2] precisely address the aforementioned issue. Zhang
& Shu constructed a new conservative positivity-preserving procedure to preserve

2000 Mathematics Subject Classification. Primary: 65M06, 65M08; Secondary: 76L05.
Key words and phrases. Positivity-preserving algorithms, WENO, Noh problem.
The support of the DOE/SciDAC SAP grant DE-AI02-06ER25796 is acknowledged.
positive density and pressure for high order WENO schemes by the Lax-Friedrichs flux (WENO/LLF). In general, WENO/LLF is too dissipative for flows such as turbulence with strong shocks computed in direct numerical simulations (DNS) and large eddy simulations (LES). The new conservative positivity-preserving procedure proposed in [2] can be used with any high-order shock-capturing scheme including high order WENO schemes using the Roe’s flux (WENO/Roe).

The goal of this study is to compare the results obtained by non-positivity-preserving methods with the recently developed positivity preserving schemes for representative test cases. In particular the more difficult 3D Noh and Sedov problems are considered. These test cases are chosen due to the negative pressure/density most often exhibited by standard high order shock-capturing schemes. The simulation of a hypersonic nonequilibrium viscous shock tube that is related to the NASA Electric Arc Shock Tube (EAST) is also included. EAST is a high temperature and high Mach number viscous nonequilibrium flow consisting of 13 species. In addition, as most common shock-capturing schemes have been developed for problems without source terms, when applied to problems with nonlinear and/or stiff source terms these methods can result in spurious solutions, even when solving a conservative system of equations with a conservative scheme. This kind of behavior can be observed even for a scalar case [5] as well as for the case considering two species and one reaction [9]. For further information concerning this issue see [5, 1, 4, 13]. This EAST example indicated that standard high order shock-capturing methods exhibit instability of density/pressure in addition to grid dependent discontinuity locations with insufficient grid points. The evaluation of these test cases is based on the stability of the numerical schemes together with the accuracy of the obtained solutions.

2. Positivity-preserving algorithms. Here we briefly describe the Hu et al. positivity-preserving method. For more details, see [2]. Readers are referred to [17] for their positivity-preserving WENO schemes that are only valid for the Lax-Friedrichs flux formulation. Consider the Euler equations:

\[ w_t + f(w)_x = 0, \]

\[ w = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix}, \quad f(w) = \begin{pmatrix} \rho u^2 + p \\ (E + p)u \end{pmatrix} \]

with

\[ m = \rho u, \quad E = \frac{1}{2} \rho u^2 + \rho e, \quad p = (\gamma - 1)\rho e, \]

where \( \rho \) is the density, \( u \) is the velocity, \( m \) is the momentum, \( E \) is the total energy, \( p \) is the pressure, \( e \) is the internal energy, and \( \gamma > 1 \) is a constant. The speed of sound is given by \( c = \sqrt{\gamma p/\rho} \) and the three eigenvalues of the Jacobian \( f'(w) \) are \( u - c, u \) and \( u + c \).

A general explicit \( k^{th} \)-order conservative scheme with Euler-forward time integration for Eq. (1) can be written as

\[ w_i^{n+1} = w_i^n - \lambda \left( \tilde{f}_{i+\frac{1}{2}} - \tilde{f}_{i-\frac{1}{2}} \right), \]

where the superscripts \( n \) and \( n+1 \) represent the old and new time steps, respectively, and \( \lambda = \Delta t/\Delta x \), where \( \Delta t \) is the time-step size and \( \Delta x \) is the grid-step size.
The Eq. (3) can be rewritten as follows:

\[ 2w_i^{n+1} = \left( w_i^n + 2\lambda f_{i+\frac{1}{2}}^{-} \right) + \left( w_i^n - 2\lambda f_{i+\frac{1}{2}}^{+} \right) = w_i^{+} + w_i^{-}. \quad (4) \]

The Hu et al. positivity-preserving procedure involves the first-order Lax-Friedrichs scheme with numerical flux:

\[ \hat{f}_{i+\frac{1}{2}}^{LF} = \left\{ \frac{1}{2} \left[ f_i + f_{i+1} + (|u| + c)_{\text{max}} (w_i^n - w_{i+1}^n) \right] \right\}. \quad (5) \]

Let \( w_i^{LF,\pm} = w_i^n \pm 2\lambda \hat{f}_{i+\frac{1}{2}}^{LF} \). The positive density is first enforced by the following cut-off flux limiter for positive density:

1. For all \( i \) initialize \( \theta_{i+1/2}^+ = 1, \theta_{i+1/2}^- = 1 \).
2. If \( \rho \left( w_i^+ \right) < \epsilon_p \), solve \( \theta_{i+1/2}^+ \) from \( (1 - \theta_{i+1/2}^+)\rho(w_i^{LF,+}) + \theta_{i+1/2}^+\rho(w_i^+) = \epsilon_p \).
3. If \( \rho \left( w_i^- \right) < \epsilon_p \), solve \( \theta_{i+1/2}^- \) from \( (1 - \theta_{i+1/2}^-)\rho(w_i^{LF,-}) + \theta_{i+1/2}^-\rho(w_i^-) = \epsilon_p \).
4. Set \( \theta_{p,i+1/2} = \min(\theta_{p,i+1/2}^+, \theta_{p,i+1/2}^-) \).

Here, \( \epsilon_p = \min\{10^{-13}, \rho_{\text{min}}^0 \} \), where \( \rho_{\text{min}} \) is the minimum density in the initial condition, \( \hat{f}_{i+\frac{1}{2}}^\pm \) is the limited flux, and \( 0 \leq \theta_{i+1/2}^\pm \leq 1 \).\( \hat{f}_{i+\frac{1}{2}}^\pm \) are the limiting factors corresponding to the two neighboring cells, which share the same flux \( \hat{f}_{i+\frac{1}{2}}^\pm \). After applying this flux limiter, Eq. (4) becomes

\[ 2w_i^{n+1} = \left( w_i^n + 2\lambda \hat{f}_{i+\frac{1}{2}}^+ \right) + \left( w_i^n - 2\lambda \hat{f}_{i+\frac{1}{2}}^- \right) = w_i^{+} + w_i^{-}. \quad (6) \]

The positive pressure is further enforced by the following cut-off flux limiter for positive pressure:

1. For all \( i \) initialize \( \theta_{i+1/2}^+ = 1, \theta_{i+1/2}^- = 1 \).
2. If \( \rho \left( w_i^+ \right) < \epsilon_p \), solve \( \theta_{i+1/2}^+ \) from \( (1 - \theta_{i+1/2}^+)\rho(w_i^{LF,+}) + \theta_{i+1/2}^+\rho(w_i^+) = \epsilon_p \).
3. If \( \rho \left( w_i^- \right) < \epsilon_p \), solve \( \theta_{i+1/2}^- \) from \( (1 - \theta_{i+1/2}^-)\rho(w_i^{LF,-}) + \theta_{i+1/2}^-\rho(w_i^-) = \epsilon_p \).
4. Set \( \theta_{p,i+1/2} = \min(\theta_{p,i+1/2}^+, \theta_{p,i+1/2}^-) \).

Again, \( \epsilon_p = \min\{10^{-13}, \rho_{\text{min}}^0 \} \), where \( \rho_{\text{min}} \) is the minimum pressure in the initial condition, and \( \hat{f}_{i+\frac{1}{2}}^\pm \) is the further limited flux. With these limited fluxes, the original scheme (3) is modified as

\[ w_i^{n+1} = w_i^n - \lambda \left( \hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}} \right). \quad (7) \]

The above new procedure can be applied at each sub-stage of a TVD Runge-Kutta [8] method, which is a convex combination of Euler-forward time steps. Note that any high order numerical flux can be used in this formulation.

3. Numerical Results. The numerical experiments include 1D and 3D Noh problems, 3D Sedov blast, a Mach 2000 Jet test case, and the 1D EAST hypersonic viscous nonequilibrium shock tube. The comparative study includes the following numerical schemes:

- **UPWIND** - first-order upwind scheme using Roe average state
- **TVD** - second-order TVD [11, 12]
- **Standard fifth and seventh WENO (WENO5 and WENO7)** with local Lax-Friedrichs flux (WENO5-LLF and WENO7-LLF) and Roe flux (WENO5-Roe), see [3]
- **WENO5fi+split** and **WENO7fi+split** - high-order nonlinear filter counterparts of WENO5 and WENO7, see [10, 15]
Figure 1. Density plot for the Noh-1D problem obtained by UPWIND (line 1), WENO5-LLF (line 2), WENO5-Roe (line 3), WENO5GP (line 4), WENO5P (line 5), WENO7-LLF (line 6), WENO7P (line 7), WENO5afi (line 8) and WENO7afi (line 9) on grid $N = 267$. Reference solution: line 10. Right figure is zoomed in the vicinity of $x = 0$.

Figure 2. Density plot for the Noh-1D problem obtained by WENO5-Roe on grid $N = 267$ with different values of the entropy fix parameter: $\epsilon = 0.20$ (line 1), $\epsilon = 0.25$ (line 2), $\epsilon = 0.30$ (line 3), $\epsilon = 0.35$ (line 4). Reference solution: line 5.

- Zhang & Shu positivity-preserving WENO5 and WENO7 local Lax-Friedrichs flux (WENO5P, WENO7P) and of fifth order with global Lax-Friedrichs flux (WENO5GP).
- Hu et al. positivity-preserving WENO5 with local Lax-Friedrichs flux (WENO5PH) and Roe’s flux (WENO5PH-Roe)

3.1. Noh problem. The first test case is the well-known 1D and 3D spherical Noh implosion problem [6]. The initial conditions are $\rho = 1$, $p = 0$ and $u =$ unit vector directed toward the origin, with $\gamma = 5/3$. In this problem an infinite-strength shock
expands outward from the origin at a constant velocity of 1/3. The goal is to test the ability of the scheme to preserve spherical symmetry and produce the correct entropy jump for adiabatic shock compression.

The results obtained using different schemes for the 1D case are shown in Fig. 1. The grid cell size is $h = 0.002$. The second order TVD scheme does not obtain a solution for this case. Increasing the WENO order from fifth to seventh gives slightly better results. In addition, the regular WENO-LLF schemes are slightly more accurate than their positive counterparts. The WENO5-Roe scheme here performs with the same accuracy as WENO5-LLF using a large entropy fix. Its accuracy can be improved by using the appropriate entropy fix parameter $\epsilon$. Figure 2 shows the results by the WENO5-Roe scheme with different values of $\epsilon$. The value $\epsilon = 0.3$ produces an error close to zero in the vicinity of $x = 0$ instead of an error of 0.5% obtained by the scheme with the regular value of $\epsilon = 0.25$.

The 2D slices at $z = 0$ obtained for the 3D case by UPWIND, WENO5P, WENO7P, WENO5PH, WENO7PH and WENO5PH-Roe are shown in Fig. 3. The grid size is $134 \times 134 \times 134$. Note that regular WENO5-LLF and WENO5-Roe do not obtain the solution for the Noh 3D problem because of pressure reaching negative values. The results by WENO5P and WENO7P also contain some points with small negative pressure values. This can be fixed by applying adaptive timestep using the positivity preservation condition. The results obtained using WENO5PH,
The second test case is the 3D spherical Sedov blast wave [7]. The initial conditions are $\rho = 1$, $u = 0$ and $e = 0.1528415451 \exp(-R^2/R_0^2)/R_0^3$, where $R = \sqrt{x^2 + y^2 + z^2}$ and $R_0 = 2/h$. The grid cell size is $h = 0.02$. Here we use $\gamma = 1.4$. The density and temperature contours are shown in Figs. 5 and 6. The deviation from the spherical symmetry for positive schemes is bigger than for regular WENO, which is well observed on the temperature contours. However, the solution obtained using standard WENO schemes contains some points with small negative pressure values, whereas WENOP and WENOPfi obtain all positive pressure and density values. Switching to Roe’s flux causes even further deviation from the symmetry (see results obtained using WENO5PH-Roe). Note that the density contours obtained by all of these methods look very similar.

3.3. Mach 2000 Jet. The same Mach 2000 jet problem as in [16] is considered here with $\gamma = 5/3$. The computational domain is $[0, 1] \times [-0.25, 0.25]$. The initial flow condition is the ambient gas with $(\rho, u, v, p) = (0.5, 0, 0, 0.4127)$. The boundary conditions for the right, top and bottom are outflow. For the left boundary,
Figure 5. Density contours for the Sedov-3D problem on grid $128 \times 128 \times 128$ (slice $z = 0$). Top row: WENO5-LLF, WENO5Pfi+split and WENO5PH. Bottom row: WENO5P, WENO7P and WENO5PH-Roe.

$(\rho, u, v, p) = (5, 800, 0, 0.4127)$ if $y \in [-0.05, 0.05]$ and $(\rho, u, v, p) = (0.5, 0, 0, 0.4127)$ otherwise. The terminal time is 0.001. The speed of the jet is 800, which is around Mach 2100 with respect to the sound speed in the jet gas.

For this problem very small initial CFL value is required for high-order computation (about 0.01). For this reason, a variable time step control is used in the computation. After each RK stage the solution is tested using the positivity condition. If the condition is not satisfied, the timestep is divided by a factor of 2 and the current RK step is repeated again. In this way the computation can be carried out with an average CFL value 4–8 times larger than the fixed CFL.

The results by different schemes on the uniform grid $800 \times 400$ are shown in the Fig. 7. Note that regular WENO5-LLF and WENO5-Roe do not obtain the solution for this problem because of pressure reaching negative values. Here the quality of the results is improved when increasing the order of the numerical scheme. The exception is WENO7P which obtains essential oscillations. The results obtained using WENO5PH-Roe are a bit less dissipative than in the case using WENO5PH-LLF or WENO5P. The entropy fix parameter used in WENO5PH-Roe for this case was $\epsilon = 0.3$.

3.4. 13 species 1D EAST simulation. The computational domain has a total length of 8.5m. The left part of the domain with length 0.1m is a high pressure region. The right part of the domain with length 8.4m is a low pressure region. The gas mixture consists of 13 species:
Figure 6. Temperature contours for the Sedov-3D problem on grid $128 \times 128 \times 128$ (slice $z = 0$). Top row: WENO5-LLF, WENO5Pfi+split and WENO5PH. Bottom row: WENO5P, WENO7P and WENO5PH-Roe.

Table 1. High (left) and low (right) pressure region initial data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>1.10546 kg/m$^3$</td>
</tr>
<tr>
<td>$T$</td>
<td>6000 K</td>
</tr>
<tr>
<td>$p$</td>
<td>12.7116 MPa</td>
</tr>
<tr>
<td>$Y_{He}$</td>
<td>0.9856</td>
</tr>
<tr>
<td>$Y_{N_2}$</td>
<td>0.0144</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$3.0964 \times 10^{-4}$ kg/m$^3$</td>
</tr>
<tr>
<td>$T$</td>
<td>300 K</td>
</tr>
<tr>
<td>$p$</td>
<td>26.711 Pa</td>
</tr>
<tr>
<td>$Y_{O_2}$</td>
<td>0.21</td>
</tr>
<tr>
<td>$Y_{N_2}$</td>
<td>0.79</td>
</tr>
</tbody>
</table>

The initial conditions of the high and low pressure regions are listed in the Table 1. For the left-side boundary the Euler (slip) wall condition is applied, and for the right-side, the zero gradient condition is applied for all variables.

Figure 8 shows the results from the computation using the Harten-Yee second-order TVD scheme [11, 12] for four grids with $\Delta x = 10^{-3}$ m, $5 \times 10^{-4}$ m, $5 \times 10^{-5}$ m and $2.5 \times 10^{-5}$ m at time $t_{end} = 0.325 \times 10^{-4}$ sec. One can observe a significant shift in the shear (left discontinuity) and the shock (right discontinuity) locations as the grid is refined. The distance between the shear and the shock shrinks as the grid is refined. The difference between shock locations obtained on the grids with $\Delta x = 5 \times 10^{-5}$ m and $2.5 \times 10^{-5}$ m is less than 0.3%. Thus the solution using $\Delta x = 5 \times 10^{-5}$ m can be considered as the reference solution.
Figure 7. From top to bottom: density contours for the Mach 2000 Jet problem obtained by TVD, WENO5GP, WENO5P, WENO5PH-Roe and WENO7P on grid 800 × 400. Scales are logarithmic.
The left subfigure of Fig. 9 shows a comparison among five methods obtained on a coarse grid ($\Delta x = 10^{-3} \text{ m}$) with the reference solution. The scheme labels are defined as follows:

- WENO5-llf: Fifth-order WENO (WENO5) using the local Lax-Friedrichs flux.
- WENO5P-llf: Positive WENO5 of [17] using the local Lax-Friedrichs flux.

The right subfigure of Fig. 9 shows a comparison of ACMTVDfi using a different weight $\kappa$ parameter of the ACM flow sensor. The smaller the $\kappa$, the smaller the amount of TVD dissipation that is used. Among the considered schemes, Fig. 9 indicates that the least dissipative scheme predicts the shear and shock locations best when compared with the reference solution. The results indicate that ACMTVDfi is slightly more accurate than WENO5-llf. This is due to the fact that ACMTVDfi reduces the amount of numerical dissipation away from high gradient regions. Using the subcell resolution method of Wei et al for one reaction case [9] by applying it to only one of the reactions in this multi-reaction flow does not improve the performance over standard schemes. Further research on generalization of subcell resolution to multi-reactions needs to be explored.

4. Summary. The positivity preserving schemes produce more stable behavior than regular WENO for the considered problems. The scheme by [2] achieves slightly better positivity preservation than the scheme by [17] using the same CFL
Figure 9. 1D, 13 species EAST problem: Comparison among methods using 601 point grids with $CFL = 0.6$ and $t_{\text{end}} = 3.25 \times 10^{-5}$ sec. Left subfigure: Reference solution (TVD on a 10,001 point grid) (line 1), TVD (line 2), ACMTVDfi (TVDfi) using $\kappa = 0.5$ (line 3), WENO5-Iff (line 4), WENO5P-Iff (line 5), WENOPH-Iff (line 6). Right: ACMTVDfi, $\kappa = 0.15$ (line 2), $\kappa = 0.2$ (line 3), $\kappa = 0.3$ (line 4), $\kappa = 0.5$ (line 5), $\kappa = 1$ (line 6). See text for method notation.

number. These positivity-preserving schemes also are more diffusive than their standard WENO counterparts. The accuracy issue can be improved when using Roe’s flux with the Hu et al. scheme instead of using Lax-Friedrichs, which is required by the Zhang-Shu scheme.

Acknowledgments. The support of the DOE/SciDAC SAP grant DE-AI02-06ER25796 is acknowledged. The work was performed while the first author is a postdoc fellow at the Center for Turbulence Research, Stanford University. The authors are grateful to X. Zhang and X. Y. Hu for providing their positive scheme routines. The financial support from the NASA Fundamental Aeronautics (Hypersonic) program for the second author is gratefully acknowledged. The work by the third author was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

REFERENCES

E-mail address: dkotov@stanford.edu
E-mail address: Helen.M.Yee@nasa.gov
E-mail address: sjogreen2@llnl.gov
MATHER MEASURES IN SEMICLASSICAL ANALYSIS

LORENZO ZANELLI
Centre de Mathématiques Laurent Schwartz
Ecole Polytechnique
Route de Saclay 91120 Palaiseau, France

Abstract. We discuss the Mather’s problem of Action minimizing compactly supported measures which are invariant under the Lagrangian flow, within the framework of semiclassical Analysis on the torus. More precisely, we show that the semiclassical limit of the Wigner transform for some energy quasimodes of the Weyl quantization of Tonelli Hamiltonians coincides with the Legendre transform of some Mather measures.

1. Introduction. In this paper we deal with the class of smooth wave functions defined on the flat torus \( \mathbb{T}^n := (\mathbb{R}/2\pi \mathbb{Z})^n \) taking the form
\[
\psi_{\hbar, P}(x) := a(\hbar, P, x) e^{i(P \cdot x + v(\hbar, P, x))}
\]
where \( 0 < \hbar \leq 1, \ h^{-1} \in \mathbb{N}, P \in \mathbb{Z}^n \). Here the phase functions \( v(\hbar, P, \cdot) \in C^\infty(\mathbb{T}^n; \mathbb{R}) \) are \( \hbar \)-close to \( C^1 \), \( \hbar \)-critical subsolutions of the stationary Hamilton-Jacobi equation (see [2])
\[
H(x, P + \nabla_x v(P, x)) \leq \bar{H}(P),
\]
where \( H \in C^\infty(\mathbb{T}^n \times \mathbb{R}^n) \) is such that \( x \mapsto H(x, \eta) \) is strictly convex and there exists \( C > 0 \) such that
\[
\frac{\eta^2}{2} \leq \nabla^2_x H(x, \eta) \omega \cdot \omega \leq C |\omega|^2 \quad \forall (x, \eta) \in \mathbb{T}^n \times \mathbb{R}^n \quad \forall \omega \in \mathbb{R}^n,
\]
where \( \bar{H} \) is the so-called effective Hamiltonian (see Section 2.1). We require
\[
\| \nabla_x v(\hbar, P, \cdot) - \nabla_x v(P, \cdot) \|_{C^0} \leq C_P \hbar^\alpha, \quad \| v(\hbar, P, \cdot) \|_{C^2} \leq C_P \hbar^{-\alpha},
\]
for some \( c_P, C_P > 0 \) and \( 0 < \alpha < 1 \). The amplitudes \( a(\hbar, P, \cdot) \in C^\infty(\mathbb{T}^n; \mathbb{R}_+) \) are selected in such a way \( a^2(\hbar, P, \cdot) \) are suitable \( \hbar \)-regularizations of some Borel probability measures \( d\sigma_P \) on \( \mathbb{T}^n \). More precisely, we require the energy constraint
\[
\bar{H}(P) = \int_{\mathbb{T}^n} H(x, P + \nabla_x v(P, x))d\sigma_P(x)
\]
and the following continuity equation written in the measure sense
\[
\int_{\mathbb{T}^n} \nabla_x g(x) \cdot \nabla_x H(x, P + \nabla_x v(P, x))d\sigma_P(x) = 0, \quad \forall g \in C^\infty(\mathbb{T}^n; \mathbb{R}),
\]
see Section 3. The first objective of this paper is to prove that the semiclassical limit of the Wigner transform \( W_{\hbar, P} \psi_{h, P} \) equals the Legendre transform of Mather measures. The second target is to prove that the wave functions \( \psi_{\hbar, P} \) are energy quasimodes for the Weyl operator \( \text{Op}_w^\hbar(H) \) where the order of the \( L^2 \)-quasimode estimate that we provide depends on the parameter \( \alpha \) in (3).

In the following, \( H \) is Tonelli as prescribed above and moreover \( H \in S^2(\mathbb{T}^n \times \mathbb{R}^n) \), namely it belongs to a toroidal Hörmander’s symbol class (see Section 2.3).

2000 Mathematics Subject Classification. Primary: 81S30, 37J50; Secondary: 35F21.
Key words and phrases. Semiclassical Analysis, Aubry-Mather theory, weak KAM theory.
We are now ready to provide the first result of the paper, where we use a suitable family of smooth test functions $\Phi$ (see Def. 2.1).

**Theorem 1.1.** Let $\psi_{h,P}$ be as in (1) with phase functions as in (3) with $0 < \alpha < 1$. Then, there exists $R_{P,\delta} > 0$ such that

$$\left| \sum_{\eta \in \mathbb{Z}^n} \int_{T^n} \phi(x,\eta)W_{h}\psi_{h,P}(x,\eta)dx - \int_{T^n \times \mathbb{R}^n} \phi(x,\eta)dw_{P}(x,\eta) \right| \leq R_{P,\delta} h^\delta$$

for all $\phi \in \Phi$, $\delta := \alpha/2$, and where

$$\int_{T^n \times \mathbb{R}^n} \phi(x,\eta)dw_{P}(x,\eta) = \int_{T^n} \phi(x, P + \nabla_x v(P,x))d\sigma_P(x)$$

equals the Legendre transform of a Mather $P$-minimal measure.

The well posedness of the continuity equation (5) with the energy constraint (4) is showed within Theorem 1.1. Indeed, we also prove that the Legendre transform of any Mather measure takes the form (7) and that the push forward by the canonical projection $d\sigma_P = \pi_\star (dw_P)$ fulfill (4) and (5). The measures of type (7) have been already studied within Th. 3.4 of [7] involving Lipschitz continuous solutions of the Hamilton-Jacobi equation instead of critical subsolutions (see Section 2.1). In fact, the characterization (7) of the Legendre transform of Mather $P$-minimal measures can also be obtained as a consequence of the Lemma 3.1 in [12].

The second result of the paper reads:

**Theorem 1.2.** Let $\psi_{h,P}$ be as in (1). Then

$$\|\text{Op}_h^{w}(H)\psi_{h,P} - \bar{H}(P)\psi_{h,P}\|_{L^2} = O(h^\beta),$$

for all $0 < \beta < \delta/2$, $\delta := \alpha/2$ and $0 < \alpha < 1$.

We underline that here the best $L^2$-quasimode estimate is closed to $O(h^{1/4})$. Within general settings, there exist better quasimode estimates, for example by the use of coherent states (see [4]) there exist estimates of order $O(h^{1/2})$; the semiclassical limit of the Wigner transform localizes measures supported on points of the phase space. However, in our paper we deal with invariant measures as in (7) with support contained in the so-called Aubry set $A_P$ in the phase space (see Section 2.2) which generalizes the notion of smooth invariant Lagrangian torus, namely the set $\Delta_P := \text{Graph}(P + \nabla v(P,\cdot))$ when $v(P,\cdot)$ is a $C^2$ solution of Hamilton-Jacobi equation

$$H(x, P + \nabla_x v(P,x)) = \bar{H}(P).$$

In [8] the author suggests a quantum analog of weak KAM theory, which is in particular realized by going through the study of a class of critical points of the quantum Action functional on the torus, and by proving a quasimode estimate of order $O(h)$; same result of quasimode estimate in [10] where the phases of the wave functions are suitable $h$-regularizations for viscosity solutions of the Hamilton-Jacobi equation (9), whereas for the amplitudes it is used a regularized setting similar to our one. We remark that in that papers the study of the Wigner transform is not exhibited. In [21] the problem investigated is about the convergence of the Wigner transform of wave functions of type (1) towards measures taking the form (7) with $\sigma_P$ solving continuity equation (5) but associated to Lipschitz continuous viscosity solutions of Hamilton-Jacobi equation (9). In our paper, the use of $C^{1,1}$-critical subsolutions allows the energy quasimode estimate (8). In the papers [15] and [17]
the authors study the Wigner transform for the quantum states like as in [8], and
prove the convergence to some Mather measures in the semiclassical limit.

2. Preliminaries.

2.1. The weak KAM theory. Let us consider the Hamilton-Jacobi equation:

\[ H(x, P + \nabla_x v(P, x)) = \bar{H}(P), \quad (10) \]

for \( C^2 \)-Tonelli Hamiltonians. The function \( \bar{H} : \mathbb{R}^n \to \mathbb{R} \) is called the effective Hamiltonian and (see [3], [9]) can be expressed by

\[ \bar{H}(P) = \inf_{v \in C^\infty(T^n; \mathbb{R})} \sup_{x \in T^n} H(x, P + \nabla_x v(x)). \quad (11) \]

We recall that for a solution of the Hamilton-Jacobi equation (10), different notions exists: critical subsolutions, weak KAM solutions and viscosity solutions. Here we give only the basics. For an exhaustive treatment we refer to A. Fathi [11].

A Lipschitz continuous function \( v(P, \cdot) : T^n \to \mathbb{R} \) is a critical subsolution of the Hamilton-Jacobi equation (10) if \( H(x, P + \nabla_x v(P, x)) \leq \bar{H}(P) \) almost everywhere. For energy levels greater than \( \bar{H}(P) \), there exist \( C^\infty \) subsolutions, see [5]. Fathi and Siconolfi proved in [13] that there exist critical subsolutions with \( C^1 \) regularity. Moreover, Bernard showed in [2] the existence of \( C^{1,1} \) critical subsolutions, and exhibit a mechanical type example for which \( C^2 \) critical subsolutions do not exist.

The Lax-Oleinik semigroup is given by

\[ T_t^x u(x) := \inf_{\gamma} \{ u(\gamma(0)) + \int_0^t \mathcal{L}(\gamma(s), \dot{\gamma}(s)) ds \} \]

where the infimum is taken over all absolutely continuous curves \( \gamma : [0, t] \to T^n \) such that \( \gamma(t) = x \). Here we need to consider \( L(x, \xi) - P \cdot \xi \) where \( L \) is the Fenchel-Legendre Transform of \( H \). A function \( u : T^n \to \mathbb{R} \) is a negative weak KAM solution of (10) if \( T_t^x v(x) = v(x) - \bar{H}(P)t \) for all \( t > 0 \), whereas it is a positive weak KAM solution of (10) if \( T_t^x v(x) = v(x) + \bar{H}(P)t \) for all \( t > 0 \). Fathi proved that Lipschitz continuous weak KAM solutions exist and that the negative ones coincide with the so-called viscosity solutions (see [6], [11]).

2.2. The Aubry-Mather theory. Aubry-Mather theory proves the existence of invariant and Action-minimizing measures and sets in the phase space. Here we review only a few results that we are going to use in the subsequent sections. For a detailed treatment we refer to ([18], [20], [23]) and references quoted therein.

A Borel probability measure \( d\mu \) defined on \( T^n \times \mathbb{R}^n \) is called invariant with respect to the Euler-Lagrange flow \( \varphi_t^L : T^n \times \mathbb{R}^n \to T^n \times \mathbb{R}^n \) related to a Tonelli Lagrangian \( L \) if it is a fixed point of the push forward \((\varphi_t^L)_\ast (d\mu) = d\mu \) \( \forall t \in \mathbb{R} \). A probability measure \( d\mu \) is said to be closed if \( \forall g \in C^\infty(T^n; \mathbb{R}) \) it holds \( \int_{T^n \times \mathbb{R}^n} \xi \cdot \nabla_x g(x) \ d\mu(x, \xi) = 0 \).

A compactly supported Borel probability measure \( d\mu_P \) is \( P \)-minimal if, for any given \( P \in \mathbb{R}^n \), it minimizes the Action

\[ \inf_{d\mu} \int_{T^n \times \mathbb{R}^n} L(x, \xi) - P \cdot \xi \ d\mu(x, \xi) = -\bar{H}(P) \quad (12) \]

on the set of invariant probability measures on \( T^n \times \mathbb{R}^n \). It has been also proved that Mather measures of a Tonelli Lagrangian are those which minimize the Action in
the class of all compactly supported closed measures (see for example [1] Theorem 7). The Mather set \( M_P \) involves the supports of all Mather measures, namely
\[
M_P := \bigcup_{d\mu_P} \text{supp } d\mu_P.
\]
(13)

We recall that Mather proved in [20] that \( M_P \) is compact and that it is a Lipschitz graph above a compact part of \( T^n \). The Aubry set in the cotangent bundle \( T^n \times \mathbb{R}^n \),
\[
A^\star_P := \bigcap_v \left\{ (x, P + \nabla_x v(P, x)) \mid v(P, x) \text{ is differentiable in } x \right\},
\]
(14)
is the intersection can be taken over all \( v(P, \cdot) \) Lipschitz continuous critical subsolutions of the stationary Hamilton-Jacobi equation. This set is also invariant under the Hamiltonian dynamics, and it has been proved in [3] that
\[
\mathcal{L}(M_P) \subseteq A^\star_P
\]
(15)
where \( \mathcal{L} \) is the Legendre transform. Hence the Legendre transform of Mather \( P \)-minimal measures are supported on the Aubry set \( A^\star_P \).

2.3. The Weyl quantization on the torus. Consider the class of Hörmander’s symbols \( b \in S^m(T^n \times \mathbb{R}^n) \), \( m \in \mathbb{R} \), consisting of those functions in \( C^\infty(T^n \times \mathbb{R}^n) \) which are \( 2\pi \)-periodic in \( x \) (that is, in each variable \( x_j \), \( 1 \leq j \leq n \)) and for which for all \( \alpha, \beta \in \mathbb{Z}_+^n \) there exists \( C_{\alpha\beta} > 0 \) such that \( \forall (x, \eta) \in T^n \times \mathbb{R}^n \)
\[
|\partial_x^\alpha \partial_\eta^\beta b(x, \eta)| \leq C_{\alpha\beta} \langle \eta \rangle^{m-|\beta|}
\]
(where \( \langle \eta \rangle := (1 + |\eta|^2)^{1/2} \)). We assume here \( C_{\alpha\beta} = C \). We next recall the Weyl quantization on the torus as used in [21]. Given a symbol \( b \in S^m(T^n \times \mathbb{R}^n) \), we put for \( \psi \in C^\infty(T^n) \)
\[
\text{Op}_h^w(b)\psi(x) := (2\pi)^{-n} \sum_{\eta \in \mathbb{Z}^n} \int_{T^n} e^{i(x-y, \eta)} b(y, h\eta/2) \psi(2y - x) dy,
\]
(16)
The Wigner distribution is therefore given by
\[
\langle \psi, \text{Op}_h^w(b)\psi \rangle_{L^2(T^n)} = \sum_{\eta \in \mathbb{Z}^n} \int_{T^n} b(x, \eta) W_h \psi(x, \eta) dx,
\]
(17)
where the Wigner transform \( W_h \psi \) is
\[
W_h \psi(x, \eta) := (2\pi)^{-n} \int_{T^n} e^{2\pi i (x-z, \eta)} \psi(x+z) \overline{\psi(x+z)} dz.
\]
(18)

We now introduce a family of test functions which will be useful for our semi-classical estimates.

**Definition 2.1.** Let \( \mathcal{F} : \mathbb{R}^n \to \mathbb{R}^n \) be the Fourier transform, let us define the family \( \Phi \) of test functions \( \phi \in C^\infty(T^n \times \mathbb{R}^n; \mathbb{R}) \) such that \( \exists \mathcal{F}_\eta^{-1}\phi(x, \cdot) \forall x \in T^n \), the function \( \mathcal{F}_\eta^{-1}\phi : T^n \times \mathbb{R}^n \to \mathbb{R} \) is continuous and \( \text{supp}(\mathcal{F}_\eta^{-1}\phi) \) is compact.
3. The class of wave functions. Let \( v(P, \cdot) \in C^{1,1}(\mathbb{T}^n; \mathbb{R}) \) be a critical subsolution of the Hamilton-Jacobi equation (2). We take functions \( v(h, P, \cdot) \in C^\infty(\mathbb{T}^n; \mathbb{R}) \) such that for a fixed \( 0 < \alpha < 1 \) and constants \( c_P, C_P > 0 \) uniformly bounded for \( P \) belonging to a bounded set of \( \mathbb{R}^n \), one has for \( h \in (0, 1] \)

\[
\| \nabla_x v(h, P, \cdot) - \nabla_x v(P, \cdot) \|_{C^0} \leq c_P h^\alpha, \quad \| v(h, P, \cdot) \|_{C^2} \leq C_P h^{-\alpha} \tag{19}
\]

Now select Borel probability measures \( d\sigma_P \) on \( \mathbb{T}^n \) such that

\[
\int_{\mathbb{T}^n} \nabla_x g(x) \cdot \nabla_H H(x, P + \nabla_x v(P, x)) d\sigma_P(x) = 0, \quad \forall g \in C^\infty(\mathbb{T}^n; \mathbb{R}). \tag{21}
\]

The related amplitude functions are defined as

\[
a(h, P, x) := \left\{ \int_{\mathbb{T}^n} \frac{1}{c_P} [h^\epsilon + \sum_{k \in \mathbb{Z}^n} h^{-\gamma} \rho \left( \frac{x - y - 2\pi k}{h^\gamma} \right)] d\sigma_P(y) \right\}^{1/2}, \tag{22}
\]

where \( \epsilon, \gamma > 0 \) and \( 0 < \epsilon + \gamma (n + 1) < 2 \), the function \( \rho \in C^\infty_{0}(\mathbb{R}^n) \) being such that \( 0 \leq \rho \leq 1 \), supp(\( \rho \)) \( \subset Q_n \) and \( \int \rho = 1 \), and where \( c_P = c_P(h) = \| h^\epsilon + \rho \|_{L^1(Q_n)} = 1 + O(h) \), \( Q_n := [0, 2\pi]^n \simeq \mathbb{T}^n \).

**Definition 3.1.** Let \( 0 < h \leq 1 \), \( h^{-1} \in \mathbb{N} \), \( P \in \mathbb{Z}^n \). We define

\[
\psi_{h, P}(x) := a(h, P, x) e^{\frac{1}{h}(P - x + v(h, P, x))}
\]

with \( a(h, P, \cdot) \) as in (22) and \( v(h, P, \cdot) \) as in (19).

4. Proof of the main results.

**Proof of Theorem 1.1.** By following the same computations as in the proof of Theorem 1.1 in [21] we get the semiclassical estimate (6) with \( R_{P, \phi} = \tilde{R}_P \| F^{-1}_\phi \|_\infty \) for some \( \tilde{R}_P > 0 \).

In order to prove that (7) are the Legendre transform of Mather measures, define for \( \tilde{b} \in C^0(\mathbb{T}^n \times \mathbb{R}^n; \mathbb{R}) \)

\[
\int_{\mathbb{T}^n \times \mathbb{R}^n} \tilde{b}(x, \eta) d\mu_P(x, \eta) := \int_{\mathbb{T}^n} b(x, P + \nabla_x v(P, x)) d\sigma_P(x), \tag{23}
\]

and \( d\mu_P := (\mathcal{L}^{-1})_*(d\mu_P) \) by the push forward of the inverse Legendre transform \( \mathcal{L}^{-1} \). Hence, for \( \tilde{b} \in C^0(\mathbb{T}^n \times \mathbb{R}^n; \mathbb{R}) \) this measure reads

\[
\int_{\mathbb{T}^n \times \mathbb{R}^n} \tilde{b}(x, \xi) d\mu_P(x, \xi) = \int_{\mathbb{T}^n \times \mathbb{R}^n} \tilde{b}(x, \xi) (\mathcal{L}^{-1})_*(d\mu_P)(x, \xi) = \int_{\mathbb{T}^n \times \mathbb{R}^n} \tilde{b}(\mathcal{L}^{-1}(x, \eta)) d\mu_P(x, \eta) \tag{24}
\]

namely,

\[
\int_{\mathbb{T}^n} \tilde{b}(\mathcal{L}^{-1}(x, P + \nabla_x v(P, x))) d\sigma_P(x) = \int_{\mathbb{T}^n} \tilde{b}(x, \nabla_H H(x, P + \nabla_x v(P, x))) d\sigma_P(x).
\]

We claim that \( d\mu_P \) is a closed measure. In fact,

\[
\int_{\mathbb{T}^n \times \mathbb{R}^n} \nabla_x g(x) \cdot \xi d\mu_P(x, \xi) = \int_{\mathbb{T}^n} \nabla_x g(x) \cdot \nabla_H H(x, P + \nabla_x v(P, x)) d\sigma_P(x) = 0
\]
thanks to the assumption on $d\sigma_p$. Moreover,

$$\text{supp}(d\mu_P) \subseteq T^n \times B_R(0)$$

where $R := \sup_{x \in T^n} |\nabla \eta H(x, P + \nabla_x v(P, x))| < +\infty$ since $H$ is Tonelli, $d\sigma_p$ is a probability measure and $v(P, \cdot) \in C^{1,1}(T^n \times \mathbb{R})$. Hence, $d\mu_P$ is compactly supported. We need now to show the minimizing property. Consider

$$\int_{T^n \times \mathbb{R}^n} L(x, \xi) - P \cdot \xi \, d\mu_P(x, \xi)$$

$$= \int_{T^n} L(x, \nabla \eta H(x, P + \nabla_x v(P, x)) - P \cdot \nabla \eta H(x, P + \nabla_x v(P, x)) \, d\sigma_P(x),$$

and apply the Fenchel-Legendre inequality $L(x, \xi) - \eta \cdot \xi \geq -H(x, \eta)$, which becomes an equality if and only if $\xi = \nabla \eta H(x, \eta)$ (resp. $\eta = \nabla_x L(x, \xi)$). Thus,

$$= \int_{T^n} -H(x, P + \nabla_x v(P, x)) \, d\sigma_P(x) = -\hat{H}(P),$$

thanks to the energy constraint (20). Now apply Theorem 7 in [1] and we conclude that $d\mu_P$ is a Mather measure.

Conversely, take a Mather measure $d\mu_P$ and $\Lambda_P := \text{Graph}(P + \nabla v(P, \cdot))$. Then, thanks to (15) it holds $\text{supp}(\mathcal{L}_*(d\mu_P)) \subseteq \Lambda_P$ and hence we can write

$$\int_{T^n \times \mathbb{R}^n} f(x, \eta) \mathcal{L}_*(d\mu_P)(x, \eta) = \int_{\Lambda_P} f(x, \eta) \mathcal{L}_*(d\mu_P)(x, \eta)$$

for all $f \in C^0(T^n \times \mathbb{R}^n; \mathbb{R})$, where on the new domain $\Lambda_P$ we have defined the restriction of the Borel sigma-algebra of $T^n \times \mathbb{R}^n$ (since $\Lambda_P$ is the graph of a continuous function). Moreover, the canonical projection $\pi : \Lambda_P \subseteq T^n \times \mathbb{R}^n \to T^n$ is a measurable mapping between this sigma-algebra and the Borel sigma-algebra defined on $T^n$. Thus, we can apply the change of coordinates by the map $\pi : \Lambda_P \subseteq T^n \times \mathbb{R}^n \to T^n$, to have

$$\int_{T^n} f(x, P + \nabla v(P, x)) \pi_*(\mathcal{L}_*(d\mu_P))(x)$$

and define $d\sigma_P := \pi_*(\mathcal{L}_*(d\mu_P))$. The measure $d\mu_P$ is closed, which means that $0 = \int_{T^n \times \mathbb{R}^n} \nabla_x \phi(x) \cdot \xi \, d\mu_P(x, \xi)$ and by the Legendre transform and change of variables, we get

$$0 = \int_{T^n \times \mathbb{R}^n} \nabla_x \phi(x) \cdot \nabla \eta H(x, \eta) \, \mathcal{L}_*(d\mu_P)(x, \eta)$$

$$= \int_{T^n} \nabla_x \phi(x) \cdot \nabla \eta H(x, P + \nabla v(P, x)) \pi_*(\mathcal{L}_*(d\mu_P))(x)$$

$$= \int_{T^n} \nabla_x \phi(x) \cdot \nabla \eta H(x, P + \nabla v(P, x)) \, d\sigma_P(x).$$

Proof of Theorem 1.2. The assumption $H \in S^2(T^n \times \mathbb{R}^n)$ provides a well defined setting for the quasimode estimate of the Weyl Operator $\text{Op}_h^w(H)$.

Since $\text{Op}_h^w(H)$ is selfadjoint with respect to $L^2$-scalar product,

$$\|\text{Op}_h^w(H)\psi - E\psi\|_{L^2}^2 = \langle \psi, \text{Op}_h^w(H)^2\psi\rangle_{L^2} - 2E\langle \psi, \text{Op}_h^w(H)\psi\rangle_{L^2} + E^2.$$

Now recall the well known semiclassical expansion for the composition of symbols $a \in S^0(T^n \times \mathbb{R}^n)$ and $b \in S^1(T^n \times \mathbb{R}^n)$

$$\text{Op}_h^w(a) \circ \text{Op}_h^w(b) = \text{Op}_h^w(ab)$$
where $a\sharp b = a \cdot b - \frac{ih}{2}\{a, b\} + r_h$, where $r_h = O(h^2)$ in $S^{p+q-1}(\mathbb{T}^n \times \mathbb{R}^n)$. Hence

$$\text{Op}_h^w(H)^2 = \text{Op}_h^w(H^2) + \text{Op}_h^w(r_h)$$

with $r_h = O(h^2)$ in $S^3(\mathbb{T}^n \times \mathbb{R}^n)$; apply for example Th. 4.3 in [22]. This gives

$$\|\text{Op}_h^w(H)\psi - E\psi\|_{L^2}^2 = \langle \psi, \text{Op}_h^w(H^2)\psi \rangle_{L^2} - 2E\langle \psi, \text{Op}_h^w(H)\psi \rangle_{L^2} + E^2$$

$$+ \langle \psi, \text{Op}_h^w(r_h)\psi \rangle_{L^2}.$$  \hspace{1cm} (31)

Now we apply (see Section 2.3)

$$\langle \psi, \text{Op}_h^w(f)\psi \rangle_{L^2}(\mathbb{T}^n) = \sum_{\eta \in \frac{1}{2}\mathbb{Z}^n} \int_{\mathbb{T}^n} f(x, \eta) W_h \psi(x, \eta) dx$$  \hspace{1cm} (32)

for symbols $f = H, H^2, r_h$. Recall that (6) holds for $R_{p, \phi} = \tilde{R}_P \|F^{-1}_w\phi\|_{L^\infty}$. Now, take a partition $\text{Op}_h^w(f) = \text{Op}_h^w(\sum_{k \in \mathbb{Z}^n} f \cdot P_k) \psi = \sum_{k \in \mathbb{Z}^n} \text{Op}_h^w(f \cdot P_k) \psi + O(h^\infty)$ (by using similar arguments as in Lemma 2.8.2 in [19]) and test functions $\phi, \psi \in \Phi$ with $j \in \mathbb{N}$ such that $\|\partial_x^j \psi \|_{L^\infty(\text{supp}(x_k))} \leq c_j \|\psi\|_{L^\infty(\text{supp}(x_k))}$ for $0 < \epsilon < \delta$. After some computations and the application of the toroidal version of Calderon-Vaillancourt Theorem (see Prop 3.5 of [16]), we get $R_{1, P, \epsilon} > 0$ such that

\[
\begin{align*}
|\langle \psi_{h, p}, \text{Op}_h^w(H)\psi_{h, p} \rangle_{L^2} - \int_{\mathbb{T}^n \times \mathbb{R}^n} H(x, \eta) \, dw_{P}(x, \eta)| &\leq R_{1, P, \epsilon} h^\gamma \\
|\langle \psi_{h, p}, \text{Op}_h^w(H^2)\psi_{h, p} \rangle_{L^2} - \int_{\mathbb{T}^n \times \mathbb{R}^n} H^2(x, \eta) \, dw_{P}(x, \eta)| &\leq R_{1, P, \epsilon} h^\gamma \\
|\langle \psi_{h, p}, \text{Op}_h^w(r_h)\psi_{h, p} \rangle_{L^2} - \int_{\mathbb{T}^n \times \mathbb{R}^n} r_h(x, \eta) \, dw_{P}(x, \eta)| &\leq R_{3, P, \epsilon} h^\gamma
\end{align*}
\]

for $0 < \gamma := \delta - c_n \epsilon < \delta < 1/2$ and some $c_n > 0$. The constraint $\text{supp}(dw_P) \subset H^{-1}(E)$ with $E := \tilde{H}(P)$ gives

$$|\langle \psi_{h, p}, \text{Op}_h^w(H)\psi_{h, p} \rangle_{L^2} - E| \leq R_{1, P, \epsilon} h^\gamma$$

$$|\langle \psi_{h, p}, \text{Op}_h^w(H^2)\psi_{h, p} \rangle_{L^2} - E^2| \leq R_{2, P, \epsilon} h^\gamma.$$  \hspace{1cm} (33)

Now recall that $K_P := \text{supp}(dw_P)$ is compact, hence

$$\sup_{(x, \eta) \in K_P} |r_h(x, \eta)| = O(h^2).$$  \hspace{1cm} (34)

Since $dw_P$ is a probability measure, we have

$$\left| \int_{\mathbb{T}^n \times \mathbb{R}^n} r_h(x, \eta) \, dw_{P}(x, \eta) \right| \leq \sup_{(x, \eta) \in K_P} |r_h(x, \eta)| = O(h^2).$$  \hspace{1cm} (35)

We directly get

$$\|\text{Op}_h^w(H)\psi_{h, p} - E\psi_{h, p}\|_{L^2} \leq R_{2, P, \epsilon} h^\gamma + 2ER_{1, P, \epsilon} h^\gamma + R_{3, P, \epsilon} h^\gamma + O(h^2)$$  \hspace{1cm} (36)

and conclude

$$\|\text{Op}_h^w(H)\psi_{h, p} - E\psi_{h, p}\|_{L^2} = O(h^\beta)$$  \hspace{1cm} (37)

for $0 < \beta < \delta/2$, since $\min(1; \delta/2) = \delta/2$ when $0 < \delta < 1/2$. \hfill ∎
REFERENCES


E-mail address: lorenzo.zanelli@ens.fr
# Author Index

<table>
<thead>
<tr>
<th>Author</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abreu, Eduardo</td>
<td>233</td>
</tr>
<tr>
<td>Achleitner, Franz</td>
<td>241</td>
</tr>
<tr>
<td>Alberti, Giovanni</td>
<td>399</td>
</tr>
<tr>
<td>Alessandro Morando</td>
<td>1007</td>
</tr>
<tr>
<td>Almeida, Luís</td>
<td>293</td>
</tr>
<tr>
<td>Amadori, Debora</td>
<td>641</td>
</tr>
<tr>
<td>Ancona, Fabio</td>
<td>249, 447</td>
</tr>
<tr>
<td>Andreianov, Boris</td>
<td>259</td>
</tr>
<tr>
<td>Angot, Philippe</td>
<td>277</td>
</tr>
<tr>
<td>Aregba-Driollet, Denise</td>
<td>269, 365</td>
</tr>
<tr>
<td>Arun, Kootuttugal Revi</td>
<td>439</td>
</tr>
<tr>
<td>Auphan, Thomas</td>
<td>277</td>
</tr>
<tr>
<td>Bae, Hantaek</td>
<td>285</td>
</tr>
<tr>
<td>Bae, Soohyun</td>
<td>125</td>
</tr>
<tr>
<td>Bagnerini, Patrizia</td>
<td>293</td>
</tr>
<tr>
<td>Ballew, Joshua</td>
<td>301</td>
</tr>
<tr>
<td>Bärwolff, Günter</td>
<td>317</td>
</tr>
<tr>
<td>Bauzet, Caroline</td>
<td>309</td>
</tr>
<tr>
<td>Benfield, Michael</td>
<td>749</td>
</tr>
<tr>
<td>Berger, Marsha J.</td>
<td>675</td>
</tr>
<tr>
<td>Berres, Stefan</td>
<td>317</td>
</tr>
<tr>
<td>Berselli, Luigi C.</td>
<td>325, 967</td>
</tr>
<tr>
<td>Berthon, Christophe</td>
<td>541</td>
</tr>
<tr>
<td>Bianchini, Stefano</td>
<td>399</td>
</tr>
<tr>
<td>Blokhin, Aleksander</td>
<td>991</td>
</tr>
<tr>
<td>Borsche, Raul</td>
<td>333, 575</td>
</tr>
<tr>
<td>Boscarino, Sebastiano</td>
<td>61</td>
</tr>
<tr>
<td>Boudin, Laurent</td>
<td>341</td>
</tr>
<tr>
<td>Bourdarias, Christian</td>
<td>615</td>
</tr>
<tr>
<td>Boutin, Benjamin</td>
<td>349</td>
</tr>
<tr>
<td>Bürger, Rainmund</td>
<td>381</td>
</tr>
<tr>
<td>Bressan, Alberto</td>
<td>357</td>
</tr>
<tr>
<td>Briani, Maya</td>
<td>365</td>
</tr>
<tr>
<td>Buet, Christophe</td>
<td>509</td>
</tr>
<tr>
<td>Čanić, Sunčica</td>
<td>389</td>
</tr>
<tr>
<td>Caravenna, Laura</td>
<td>399</td>
</tr>
<tr>
<td>Carvalho, Filipe</td>
<td>407</td>
</tr>
<tr>
<td>Castañeda, Pablo</td>
<td>415, 791</td>
</tr>
<tr>
<td>Castelli, Pierre</td>
<td>709</td>
</tr>
<tr>
<td>Cavalli, Fausto</td>
<td>423</td>
</tr>
<tr>
<td>Chalons, Christophe</td>
<td>607</td>
</tr>
<tr>
<td>Chemetov, Nikolai</td>
<td>431</td>
</tr>
<tr>
<td>Chen, Guoxian</td>
<td>439</td>
</tr>
<tr>
<td>Chiodaroli, Elisabetta</td>
<td>1</td>
</tr>
<tr>
<td>Choi, Sun-Ho</td>
<td>125</td>
</tr>
<tr>
<td>Cipriano, Fernanda</td>
<td>431</td>
</tr>
<tr>
<td>Coclite, Giuseppe Maria</td>
<td>447</td>
</tr>
<tr>
<td>Coco, Armando</td>
<td>455</td>
</tr>
<tr>
<td>Colombo, Rinaldo M.</td>
<td>147, 467, 575, 641, 889</td>
</tr>
<tr>
<td>Coquel, Frédéric</td>
<td>349, 631</td>
</tr>
<tr>
<td>Corli, Andrea</td>
<td>475</td>
</tr>
<tr>
<td>Coron, Jean-Michel</td>
<td>1023</td>
</tr>
<tr>
<td>Crippa, Gianluca</td>
<td>109, 957</td>
</tr>
<tr>
<td>de Almeida, Valmor F.</td>
<td>623</td>
</tr>
<tr>
<td>De Lellis, Camillo</td>
<td>1</td>
</tr>
<tr>
<td>Debrabant, Kristian</td>
<td>483</td>
</tr>
<tr>
<td>Degond, Pierre</td>
<td>179</td>
</tr>
<tr>
<td>Delis, A.I.</td>
<td>491</td>
</tr>
<tr>
<td>Després, Bruno</td>
<td>509</td>
</tr>
<tr>
<td>Dimarco, Giacomo</td>
<td>865</td>
</tr>
<tr>
<td>Dinu, Liviu Florin</td>
<td>515</td>
</tr>
<tr>
<td>Donadello, Carlotta</td>
<td>957</td>
</tr>
<tr>
<td>Donatelli, Donatella</td>
<td>525</td>
</tr>
<tr>
<td>Duan, Renjun</td>
<td>533</td>
</tr>
<tr>
<td>Duran, Arnaud</td>
<td>541</td>
</tr>
<tr>
<td>Dziuk, Gerhard</td>
<td>775</td>
</tr>
<tr>
<td>Elling, Volker</td>
<td>873</td>
</tr>
<tr>
<td>Ernest, Jan</td>
<td>551</td>
</tr>
<tr>
<td>Falcone, Maurizio</td>
<td>559</td>
</tr>
<tr>
<td>Fan, Haitao</td>
<td>475</td>
</tr>
<tr>
<td>Feireisl, Eduard</td>
<td>11</td>
</tr>
<tr>
<td>Festa, Adriano</td>
<td>559</td>
</tr>
<tr>
<td>Franck, Emmanuel</td>
<td>509</td>
</tr>
<tr>
<td>Frid, Hermano</td>
<td>567</td>
</tr>
<tr>
<td>Frouvelle, Amic</td>
<td>179</td>
</tr>
<tr>
<td>Fu, Xuedong</td>
<td>685</td>
</tr>
<tr>
<td>Fuchs, Franz</td>
<td>41</td>
</tr>
<tr>
<td>Author</td>
<td>Page Numbers</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Furtado, Frederico</td>
<td>415</td>
</tr>
<tr>
<td>Gérard-Varet, David</td>
<td>29</td>
</tr>
<tr>
<td>Garavello, Mauro</td>
<td>467, 575</td>
</tr>
<tr>
<td>Gasser, Ingenuin</td>
<td>907</td>
</tr>
<tr>
<td>Gersbacher, Christoph</td>
<td>591</td>
</tr>
<tr>
<td>Giesselmann, Jan</td>
<td>599</td>
</tr>
<tr>
<td>Girardin, Mathieu</td>
<td>607</td>
</tr>
<tr>
<td>Gisclon, Marguerite</td>
<td>615</td>
</tr>
<tr>
<td>Glass, Olivier</td>
<td>249</td>
</tr>
<tr>
<td>Glimm, James</td>
<td>623</td>
</tr>
<tr>
<td>Goatin, Paola</td>
<td>501, 889</td>
</tr>
<tr>
<td>Godlewski, Edwige</td>
<td>631</td>
</tr>
<tr>
<td>Gowda, G. D. Veerappa</td>
<td>733</td>
</tr>
<tr>
<td>Grunert, Katrin</td>
<td>193</td>
</tr>
<tr>
<td>Guès, Olivier</td>
<td>277</td>
</tr>
<tr>
<td>Guermond, Jean-Luc</td>
<td>119</td>
</tr>
<tr>
<td>Guerra, Graziano</td>
<td>641</td>
</tr>
<tr>
<td>Ha, Seung-Yeal</td>
<td>125</td>
</tr>
<tr>
<td>Hantke, Maren</td>
<td>651</td>
</tr>
<tr>
<td>Hashimoto, Itsuko</td>
<td>659</td>
</tr>
<tr>
<td>Haspot, Boris</td>
<td>667</td>
</tr>
<tr>
<td>Helzel, Christiane</td>
<td>675</td>
</tr>
<tr>
<td>Herty, Michael</td>
<td>147</td>
</tr>
<tr>
<td>Hittmeir, Sabine</td>
<td>241</td>
</tr>
<tr>
<td>Holden, Helge</td>
<td>193</td>
</tr>
<tr>
<td>Hong, Zhenying</td>
<td>685</td>
</tr>
<tr>
<td>Huth, Frank</td>
<td>317</td>
</tr>
<tr>
<td>Jakobsen, Espen Robstad</td>
<td>483</td>
</tr>
<tr>
<td>James, François</td>
<td>693</td>
</tr>
<tr>
<td>Jenssen, Helge Kristian</td>
<td>749</td>
</tr>
<tr>
<td>Jiu, Quansen</td>
<td>701</td>
</tr>
<tr>
<td>Junca, Stéphane</td>
<td>615, 709</td>
</tr>
<tr>
<td>Kabil, Buğra</td>
<td>373</td>
</tr>
<tr>
<td>Kalise, Dante</td>
<td>717</td>
</tr>
<tr>
<td>Kall, Jochen</td>
<td>333</td>
</tr>
<tr>
<td>Karimou, Mohamed Gazibo</td>
<td>583</td>
</tr>
<tr>
<td>Kemm, Friedemann</td>
<td>725</td>
</tr>
<tr>
<td>Klingenberg, Christian</td>
<td>741</td>
</tr>
<tr>
<td>Kogan, Irina A.</td>
<td>749</td>
</tr>
<tr>
<td>Kokh, Samuel</td>
<td>607</td>
</tr>
<tr>
<td>Koley, Ujjwal</td>
<td>741, 759</td>
</tr>
<tr>
<td>Kotov, Dmitry</td>
<td>1047</td>
</tr>
<tr>
<td>Kränkle, Mirko</td>
<td>767</td>
</tr>
<tr>
<td>Kröner, Dietmar</td>
<td>775</td>
</tr>
<tr>
<td>Kumar, K. Sudarshan</td>
<td>733</td>
</tr>
<tr>
<td>Lattanzio, Corrado</td>
<td>163</td>
</tr>
<tr>
<td>Lécureux-Mercier, Magali</td>
<td>467</td>
</tr>
<tr>
<td>LeFloch, Philippe G.</td>
<td>349</td>
</tr>
<tr>
<td>Lim, Hyunkyung</td>
<td>623</td>
</tr>
<tr>
<td>Lin, Chi-Kun</td>
<td>1039</td>
</tr>
<tr>
<td>Liu, Jian-Guo</td>
<td>179</td>
</tr>
<tr>
<td>Lukáčová-Medvid’ová, Maria</td>
<td>931</td>
</tr>
<tr>
<td>Müller, Thomas</td>
<td>775, 801</td>
</tr>
<tr>
<td>Mailybaev, Alexei A.</td>
<td>783</td>
</tr>
<tr>
<td>Marcato, Pierangelo</td>
<td>525</td>
</tr>
<tr>
<td>Marche, Fabien</td>
<td>541</td>
</tr>
<tr>
<td>Marchesín, Dan</td>
<td>415, 791, 881, 949</td>
</tr>
<tr>
<td>Masmoudi, Nader</td>
<td>29</td>
</tr>
<tr>
<td>Mathiaud, Julien</td>
<td>341</td>
</tr>
<tr>
<td>Matos, Vítor</td>
<td>791</td>
</tr>
<tr>
<td>May, Georg</td>
<td>915</td>
</tr>
<tr>
<td>McMurry , Andrew</td>
<td>41</td>
</tr>
<tr>
<td>Meister, Andreas</td>
<td>823</td>
</tr>
<tr>
<td>Mishra, Siddhartha</td>
<td>41</td>
</tr>
<tr>
<td>Monache, Maria Laura Delle</td>
<td>501</td>
</tr>
<tr>
<td>Morosi, Carlo</td>
<td>847</td>
</tr>
<tr>
<td>Muha, Boris</td>
<td>389</td>
</tr>
<tr>
<td>Mulet, Pep</td>
<td>381</td>
</tr>
<tr>
<td>Natalini, Roberto</td>
<td>365</td>
</tr>
<tr>
<td>Nguyen, Khai T.</td>
<td>249</td>
</tr>
<tr>
<td>Nikolos, I.K.</td>
<td>491</td>
</tr>
<tr>
<td>Nishibata, Shinya</td>
<td>817</td>
</tr>
<tr>
<td>Noelle, Sebastian</td>
<td>439</td>
</tr>
<tr>
<td>Nolte, Martin</td>
<td>809</td>
</tr>
<tr>
<td>Ohnawa, Masashi</td>
<td>817</td>
</tr>
<tr>
<td>Ortleb, Sigrun</td>
<td>823</td>
</tr>
<tr>
<td>Panov, Evgeniy Yu.</td>
<td>831</td>
</tr>
<tr>
<td>Paola Trebeschi</td>
<td>1007</td>
</tr>
<tr>
<td>Pareschi, Lorenzo</td>
<td>865</td>
</tr>
<tr>
<td>Author Name</td>
<td>Page Numbers</td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Pelanti, Marica</td>
<td>839</td>
</tr>
<tr>
<td>Pernici, Mario</td>
<td>847</td>
</tr>
<tr>
<td>Pfeiffer, Axel</td>
<td>801</td>
</tr>
<tr>
<td>Pizzocchero, Livio</td>
<td>847</td>
</tr>
<tr>
<td>Plaza, Ramón G.</td>
<td>857</td>
</tr>
<tr>
<td>Popov, Bojan</td>
<td>119</td>
</tr>
<tr>
<td>Praveen, C.</td>
<td>733</td>
</tr>
<tr>
<td>Privat, Yannick</td>
<td>89</td>
</tr>
<tr>
<td>Puppo, Gabriella</td>
<td>923</td>
</tr>
<tr>
<td>Raynaud, Xavier</td>
<td>193</td>
</tr>
<tr>
<td>Reintjes, Moritz</td>
<td>209</td>
</tr>
<tr>
<td>Risebro, Nils Henrik</td>
<td>41, 759, 1031</td>
</tr>
<tr>
<td>Rispoli, Vittorio</td>
<td>865</td>
</tr>
<tr>
<td>Roberts, Joseph</td>
<td>873</td>
</tr>
<tr>
<td>Rodriguez-Bermudez, Panters</td>
<td>881</td>
</tr>
<tr>
<td>Rosini, Massimiliano D.</td>
<td>889</td>
</tr>
<tr>
<td>Rozanova, Olga</td>
<td>899</td>
</tr>
<tr>
<td>Russo, Giovanni</td>
<td>61, 455</td>
</tr>
<tr>
<td>Rybicki, Martin</td>
<td>907</td>
</tr>
<tr>
<td>Saye, Robert</td>
<td>81</td>
</tr>
<tr>
<td>Schütz, Jochen</td>
<td>915</td>
</tr>
<tr>
<td>Schmeiser, Christian</td>
<td>241</td>
</tr>
<tr>
<td>Schwandt, Hartmut</td>
<td>317</td>
</tr>
<tr>
<td>Seguin, Nicolas</td>
<td>631</td>
</tr>
<tr>
<td>Semplice, Matteo</td>
<td>923</td>
</tr>
<tr>
<td>Sethian, James</td>
<td>81</td>
</tr>
<tr>
<td>Sfakianakis, Nikolaos</td>
<td>931</td>
</tr>
<tr>
<td>Shao, Zhi-Qiang</td>
<td>941</td>
</tr>
<tr>
<td>Shearer, Michael</td>
<td>221</td>
</tr>
<tr>
<td>Shen, Wen</td>
<td>641</td>
</tr>
<tr>
<td>Shyue, Keh-Ming</td>
<td>839</td>
</tr>
<tr>
<td>Silva, Julio Daniel</td>
<td>949</td>
</tr>
<tr>
<td>Sjögreen, Bjorn</td>
<td>1047</td>
</tr>
<tr>
<td>Soares, Ana Jacinta</td>
<td>407</td>
</tr>
<tr>
<td>Spinolo, Laura V.</td>
<td>957</td>
</tr>
<tr>
<td>Spírito, Stefano</td>
<td>967</td>
</tr>
<tr>
<td>Strani, Marta</td>
<td>975</td>
</tr>
<tr>
<td>Suzuki, Masahiro</td>
<td>817</td>
</tr>
<tr>
<td>Terracina, Andrea</td>
<td>983</td>
</tr>
<tr>
<td>Thein, Ferdinand</td>
<td>651</td>
</tr>
<tr>
<td>Tkachev, Dmitry</td>
<td>991</td>
</tr>
<tr>
<td>Trélat, Emmanuel</td>
<td>89</td>
</tr>
<tr>
<td>Trakhinin, Yuri</td>
<td>999</td>
</tr>
<tr>
<td>Trivisa, Konstantina</td>
<td>285</td>
</tr>
<tr>
<td>Turpault, Rodolphe</td>
<td>541</td>
</tr>
<tr>
<td>Tzavaras, Athanasios E.</td>
<td>163, 599</td>
</tr>
<tr>
<td>Vauchelet, Nicolas</td>
<td>693</td>
</tr>
<tr>
<td>Villada, Luis M.</td>
<td>381</td>
</tr>
<tr>
<td>Waagan, Kmut</td>
<td>1015</td>
</tr>
<tr>
<td>Wang, Yi</td>
<td>701</td>
</tr>
<tr>
<td>Wang, Zhiqiang</td>
<td>1023</td>
</tr>
<tr>
<td>Weber, Franziska</td>
<td>1031</td>
</tr>
<tr>
<td>Wollner, Winnifried</td>
<td>907</td>
</tr>
<tr>
<td>Woopen, Michael</td>
<td>915</td>
</tr>
<tr>
<td>Wu, Kung-Chien</td>
<td>1039</td>
</tr>
<tr>
<td>Xin, Zhouping</td>
<td>701</td>
</tr>
<tr>
<td>Yee, H.C.</td>
<td>1047</td>
</tr>
<tr>
<td>Yuan, Guangwei</td>
<td>685</td>
</tr>
<tr>
<td>Yuri Trakhinin</td>
<td>1007</td>
</tr>
<tr>
<td>Zanelli, Lorenzo</td>
<td>1059</td>
</tr>
<tr>
<td>Zhou, Yijie</td>
<td>623</td>
</tr>
<tr>
<td>Zuazua, Enrique</td>
<td>89</td>
</tr>
</tbody>
</table>
The International Conference devoted to Theory, Numerics and Applications of Hyperbolic Problems, HYP2012, was held in Padova on June 24–29, 2012. The conference was the fourteenth in a highly successful series of bi-annual meetings that has become one of the most important international events in Applied Mathematics. The volume contains more than 110 contributions that were presented in this conference, including plenary presentations by C. De Lellis, E. Feireisl, N. Masmoudi, S. Mishra, G. Russo, J. Sethian, E. Zuazua, and a contribution by the keynote speaker J. Glimm. These contributions cover a wide range of topics. A very partial list includes: new methods for constructing turbulent solutions to multi-dimensional systems of conservation laws based on Baire category, transport equations with non-Lipschitz velocity fields, relative entropy functionals and the stability of fluid systems, numerical methods for hyperbolic systems with stiff relaxation terms and for multiphase flow, new advances in homogenization theory, optimal sensor location for solutions to multidimensional wave equations, singularities in general relativity. The volume should appeal to researchers, students and practitioners with general interest in time-dependent problems governed by hyperbolic equations.