

*Communications in  
Applied  
Mathematics and  
Computational  
Science*

vol. 13 no. 2 2018

# Communications in Applied Mathematics and Computational Science

msp.org/camcos

## EDITORS

### MANAGING EDITOR

John B. Bell  
Lawrence Berkeley National Laboratory, USA  
jbbell@lbl.gov

### BOARD OF EDITORS

Marsha Berger	New York University berger@cs.nyu.edu	Ahmed Ghoniem	Massachusetts Inst. of Technology, USA ghoniem@mit.edu
Alexandre Chorin	University of California, Berkeley, USA chorin@math.berkeley.edu	Raz Kupferman	The Hebrew University, Israel raz@math.huji.ac.il
Phil Colella	Lawrence Berkeley Nat. Lab., USA pcolella@lbl.gov	Randall J. LeVeque	University of Washington, USA rjl@amath.washington.edu
Peter Constantin	University of Chicago, USA const@cs.uchicago.edu	Mitchell Luskin	University of Minnesota, USA luskin@umn.edu
Maksymilian Dryja	Warsaw University, Poland maksymilian.dryja@acn.waw.pl	Yvon Maday	Université Pierre et Marie Curie, France maday@ann.jussieu.fr
M. Gregory Forest	University of North Carolina, USA forest@amath.unc.edu	James Sethian	University of California, Berkeley, USA sethian@math.berkeley.edu
Leslie Greengard	New York University, USA greengard@cims.nyu.edu	Juan Luis Vázquez	Universidad Autónoma de Madrid, Spain juanluis.vazquez@uam.es
Rupert Klein	Freie Universität Berlin, Germany rupert.klein@pik-potsdam.de	Alfio Quarteroni	Ecole Polytech. Féd. Lausanne, Switzerland alfio.quarteroni@epfl.ch
Nigel Goldenfeld	University of Illinois, USA nigel@uiuc.edu	Eitan Tadmor	University of Maryland, USA etadmor@cscamm.umd.edu
		Denis Talay	INRIA, France denis.talay@inria.fr

## PRODUCTION

production@msp.org

Silvio Levy, Scientific Editor

---

See inside back cover or [msp.org/camcos](http://msp.org/camcos) for submission instructions.


The subscription price for 2018 is US \$100/year for the electronic version, and \$150/year (+\$15, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscriber address should be sent to MSP.

Communications in Applied Mathematics and Computational Science (ISSN 2157-5452 electronic, 1559-3940 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

---

CAMCoS peer review and production are managed by EditFLOW® from MSP.

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2018 Mathematical Sciences Publishers

## A NUMERICAL STUDY OF THE EXTENDED KOHN–SHAM GROUND STATES OF ATOMS

ERIC CANCÈS AND NAHIA MOURAD

In this article, we consider the extended Kohn–Sham model for atoms subjected to cylindrically symmetric external potentials. The variational approximation of the model and the construction of appropriate discretization spaces are detailed together with the algorithm to solve the discretized Kohn–Sham equations used in our code. Using this code, we compute the occupied and unoccupied energy levels of all the atoms of the first four rows of the periodic table for the reduced Hartree–Fock (rHF) and the extended Kohn–Sham  $X\alpha$  models. These results allow us to test numerically the assumptions on the negative spectra of atomic rHF and Kohn–Sham Hamiltonians used in our previous theoretical works on density functional perturbation theory and pseudopotentials. Interestingly, we observe accidental degeneracies between s and d shells or between p and d shells at the Fermi level of some atoms. We also consider the case of an atom subjected to a uniform electric field. For various magnitudes of the electric field, we compute the response of the density of the carbon atom confined in a large ball with Dirichlet boundary conditions, and we check that, in the limit of small electric fields, the results agree with the ones obtained with first-order density functional perturbation theory.

### 1. Introduction

Since the introduction by Dirac in 1929 of a many-body nonrelativistic quantum Hamiltonian allowing a comprehensive description of the physical and chemical properties of atoms and molecules [12], countless research articles and several monographs [2; 11; 16; 17; 28; 38; 39] have been devoted to the calculation of the ground states of atoms. Some of these works aimed at computing numerically the atomic ground state energy of the helium atom — the simplest nontrivial case — with spectroscopic accuracy using relativistic corrections to Dirac’s nonrelativistic model [21]. On the other extremity of the spectrum, other works focused on proving mathematical theorems about the asymptotic limit of the nonrelativistic [25; 33; 14] or relativistic [36] ground state energy and density of neutral atoms when the

---

*MSC2010:* 35P30, 35Q40, 65Z05, 81V45.

*Keywords:* density functional theory, electronic structure of atoms, extended Kohn–Sham model, Stark effect.

nuclear charge goes to infinity. The mathematical foundation of Hund’s rule, the well known empirical recipe to fill in the occupied spin orbitals of atoms, was elucidated in the limit of weak electronic interactions [15]. Interesting articles containing new results on atomic electronic structures in the framework of density functional theory have been recently published [22; 26].

This work is concerned with extended Kohn–Sham models for atoms. Recall that extended Kohn–Sham models are (zero-temperature) Kohn–Sham models allowing fractional occupancies of the Kohn–Sham orbitals (see [13] and references therein). The exact extended Kohn–Sham model, that is, the extended Kohn–Sham model with exact exchange–correlation functional, is obtained by applying Levy’s constraint search method [23] to the mixed-state variational formulation of the electronic ground-state problem [40]. Alternatively, the exact extended Kohn–Sham density functional can be seen as the Legendre–Fenchel transform of the functional mapping external potentials onto electronic ground-state energies [24]. Among other interesting mathematical features, the exact extended Kohn–Sham model is convex in the density, which is not the case of the standard Kohn–Sham model.

The simplest instance of extended Kohn–Sham model is the reduced Hartree–Fock (rHF, also called Hartree) model [35]. It is obtained by setting to zero the exchange–correlation energy functional. Although this model is too crude to obtain accurate properties of atoms and molecules, it is extremely interesting from a mathematical point of view, since its structure is very similar to the Kohn–Sham models actually used in chemistry and physics, while being *strictly convex* in the density. As a result, the rHF ground state density of a given molecular system, if it exists, is unique, and shares the symmetry properties of the nuclear distribution. In particular, the rHF density of any neutral atom is unique and spherically symmetric. Uniqueness of the ground-state density is also key to rigorously establish mathematical results in the thermodynamic limit for perfect crystals [10], crystals with points defects [3], or disordered materials [4]. The rHF model therefore is of particular interest for mathematicians. One of the motivations of the present work is to contribute to a better understanding of the rHF model, by carefully investigating the structures of rHF atomic ground states. We will also consider extended Kohn–Sham LDA (local density approximation) models [20; 29]. We will study the case of an isolated atom, as well as the case of an atom subjected to cylindrically symmetric external potential. We notably have in mind Stark potentials, which are potentials of the form  $W(\mathbf{r}) = -\mathcal{E} \cdot \mathbf{r}$  generated by a uniform electric field  $\mathcal{E} \neq 0$ .

We first propose a method to accurately solve the extended Kohn–Sham problem for cylindrically symmetric systems, using spherical coordinates and a separation of variables. This approach is based on the fact that, for such systems, the Kohn–Sham Hamiltonian commutes with  $L_z$ , the  $z$ -component of the angular momentum

operator,  $z$  denoting the symmetry axis of the system. We obtain in this way a family of 2D elliptic eigenvalue problems in the  $r$  and  $\theta$  variables, indexed by the eigenvalue  $m \in \mathbb{Z}$  of  $L_z$ , all these problems being coupled together through the self-consistent density. To discretize the 2D eigenvalue problems, we use harmonic polynomials in  $\theta$  (or in other words, spherical harmonics  $Y_l^0$ , which only depend on  $\theta$ ) to discretize along the angular variable, and high-order finite element methods to discretize along the radial variable  $r \in [0, L_e]$ . We then apply this approach to study numerically two kinds of systems.

First, we provide accurate approximations of the extended Kohn–Sham ground states of all the atoms of the first four rows of the periodic table. These results allow us to test numerically the assumptions on the negative spectra of atomic rHF and Kohn–Sham LDA Hamiltonians that we used in previous theoretical works on density functional perturbation theory [7] and norm-conserving semilocal pseudopotentials [8]. We show in particular that for most atoms of the first four rows of the periodic table, the Fermi level is negative and is not an accidentally degenerate eigenvalue of the rHF Hamiltonian. We also observe that there seems to be no unoccupied orbitals with negative energies. On the other hand, for some chemical elements, the Fermi level seems to be an accidentally degenerate eigenvalue (for example the rHF 5s and 4d states of the palladium atom seem to be degenerate). For a few of them, this accidentally degenerate eigenvalue is so close to zero that our calculations do not allow us to know whether it is slightly negative or equal to zero. For instance, our simulations seem to show that the 5s and 3d states of the iron atom seem to be degenerate at the rHF level of theory, and the numerical value of their energy we obtain with our code is about  $-10^{-5}$  Ha.

Second, we study an atom subjected to a uniform electric field (Stark effect). In this case, the system has no ground state (the Kohn–Sham energy functional is not bounded below), but density functional perturbation theory (see [7; 8] for a mathematical analysis) can be used to compute the polarization of the electronic cloud caused by the external electric field. The polarized electronic state is not a steady state, but a resonant state, and the smaller the electric field, the longer its life time. Another way to compute the polarization of the electronic cloud is to compute the ground state for a small enough electric field in a basis set consisting of functions decaying fast enough at infinity for the electrons to stay close to the nuclei. The Gaussian basis functions commonly used in quantum chemistry satisfy this decay property. However, it is not easy to obtain very accurate results with Gaussian basis sets, since they are not systematically improvable (over-completeness issues). Here we consider instead basis functions supported in a ball  $B_{L_e}$ , where  $L_e$  is a numerical parameter chosen large enough to obtain accurate results and small enough to prevent electrons from escaping to infinity (for a given, small, value of the external electric field  $\mathcal{E}$ ). We study the ground state energy and density

as functions of the cut-off radius  $L_e$ , and observe that for a given, small enough, uniform electric field, there is a plateau  $[L_{e,\min}, L_{e,\max}]$  on which these quantities hardly vary. For  $L_e < L_{e,\min}$ , the simulated system is too much confined, which artificially increases its energy, while for  $L_e > L_{e,\max}$ , a noticeable amount of charge accumulates at the boundary of the simulation domain, in the direction of  $\mathcal{C}$  (where the potential energy is very negative). On the other hand, for  $L_{e,\min} \leq L_e \leq L_{e,\max}$ , the simulation provides a fairly accurate approximation of the polarization energy and the polarized density.

The article is organized as follows. In Section 2, we recall the mathematical formulation of the extended Kohn–Sham model, and some theoretical results about the rHF and LDA ground states of isolated atoms and of atoms subjected to an external cylindrically symmetric potential. In Section 3, we describe the discretization method and the algorithms used in this work to compute the extended Kohn–Sham ground states of atoms subjected to cylindrically symmetric external potentials. Some numerical results are presented in Section 4.

## 2. Modeling

In this article, we consider a molecular system consisting of a single nucleus of atomic charge  $Z \in \mathbb{N}^*$  and of  $N$  electrons. For  $N = Z$ , this system is the neutral atom with nuclear charge  $Z$ , which we call atom  $Z$  for convenience.

**2.1. Kohn–Sham models for atoms.** In the framework of the (extended) Kohn–Sham model [13], the ground state energy of a system with one nucleus with charge  $Z$  and  $N$  electrons is obtained by minimizing an energy functional of the form

$$E_{Z,N}(\gamma) := \text{Tr}(-\frac{1}{2}\Delta\gamma) - Z \int_{\mathbb{R}^3} \frac{\rho_\gamma}{|\cdot|} + \frac{1}{2}D(\rho_\gamma, \rho_\gamma) + E_{\text{xc}}(\rho_\gamma) \quad (1)$$

over the set

$$\mathcal{H}_N := \{\gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) \mid 0 \leq \gamma \leq 2, \text{Tr}(\gamma) = N, \text{Tr}(-\Delta\gamma) < \infty\}, \quad (2)$$

where  $\mathcal{S}(L^2(\mathbb{R}^3))$  is the space of the self-adjoint operators on  $L^2(\mathbb{R}^3) := L^2(\mathbb{R}^3, \mathbb{R})$  and  $\text{Tr}(-\Delta\gamma) := \text{Tr}(|\nabla|\gamma|\nabla|)$ . Note that  $\mathcal{H}_N$  is a closed convex subset of the space  $\mathfrak{S}_{1,1}$  defined by

$$\mathfrak{S}_{1,1} := \{T \in \mathfrak{S}_1 \mid |\nabla|T|\nabla| \in \mathfrak{S}_1\},$$

endowed with norm

$$\|T\|_{\mathfrak{S}_{1,1}} := \|T\|_{\mathfrak{S}_1} + \||\nabla|T|\nabla|\|_{\mathfrak{S}_1}.$$

The function  $-Z/|\cdot|$  is the attraction potential induced on the electrons by the nucleus, and  $\rho_\gamma$  is the density associated with the one-body density matrix  $\gamma$ . For

$\gamma \in \mathcal{H}_N$ , we have

$$\rho_\gamma \geq 0, \quad \int_{\mathbb{R}^3} \rho_\gamma = N, \quad \int_{\mathbb{R}^3} |\nabla \sqrt{\rho_\gamma}|^2 \leq \text{Tr}(-\Delta \gamma) < \infty.$$

The last result is the Hoffmann-Ostenhof inequality [19]. Therefore, we have

$$\sqrt{\rho_\gamma} \in H^1(\mathbb{R}^3),$$

and in particular,

$$\rho_\gamma \in L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3).$$

For  $\rho \in L^{6/5}(\mathbb{R}^3)$ ,  $D(\rho, \rho)$  is equal to  $\int_{\mathbb{R}^3} V^H(\rho)\rho$ , where  $V^H$  is the Coulomb, also called Hartree, potential generated by  $\rho$ :

$$V^H(\rho) = \rho \star |\cdot|^{-1}.$$

Recall that  $V^H$  can be seen as a unitary operator from the Coulomb space  $\mathcal{C}$  to its dual  $\mathcal{C}'$ , where

$$\mathcal{C} := \{\rho \in \mathcal{S}'(\mathbb{R}^3) \mid \hat{\rho} \in L^1_{\text{loc}}(\mathbb{R}^3, \mathbb{C}), |\cdot|^{-1} \hat{\rho} \in L^2(\mathbb{R}^3, \mathbb{C})\},$$

$$(\rho_1, \rho_2)_{\mathcal{C}} = 4\pi \int_{\mathbb{R}^3} \frac{\hat{\rho}_1(\mathbf{k})^* \hat{\rho}_2(\mathbf{k})}{|\mathbf{k}|^2} d\mathbf{k}, \quad (3)$$

and

$$\mathcal{C}' := \{v \in L^6(\mathbb{R}^3) \mid \nabla v \in (L^2(\mathbb{R}^3))^3\},$$

$$(v_1, v_2)_{\mathcal{C}'} = \frac{1}{4\pi} \int_{\mathbb{R}^3} \nabla v_1 \nabla v_2 = \frac{1}{4\pi} \int_{\mathbb{R}^3} |\mathbf{k}|^2 \hat{v}_1(\mathbf{k})^* \hat{v}_2(\mathbf{k}) d\mathbf{k}. \quad (4)$$

The term  $E_{\text{xc}}$  is the exchange-correlation energy. We will restrict ourselves to two kinds of Kohn–Sham models: the rHF model, for which the exchange-correlation energy is taken equal to zero,

$$E_{\text{xc}}^{\text{rHF}} = 0,$$

and the Kohn–Sham LDA (local density approximation) model, for which the exchange-correlation energy has the form

$$E_{\text{xc}}^{\text{LDA}}(\rho) = \int_{\mathbb{R}^3} \epsilon_{\text{xc}}(\rho(r)) dr,$$

where  $\epsilon_{\text{xc}}$  is the sum of the exchange and correlation energy densities of the homogeneous electron gas. As the function  $\epsilon_{\text{xc}} : \mathbb{R}_+ \rightarrow \mathbb{R}$  is not explicitly known, it is approximated in practice by an explicit function, still denoted by  $\epsilon_{\text{xc}}$  for simplicity. We assume here that the approximate function  $\epsilon_{\text{xc}}$  is a  $C^1$  function from  $\mathbb{R}_+$  into  $\mathbb{R}_-$ ,

twice differentiable on  $\mathbb{R}_+^*$ , and obeying the conditions

$$\epsilon_{xc}(0) = 0, \quad \epsilon'_{xc}(0) \leq 0, \quad (5)$$

$$\text{there exists } 0 < \beta_- \leq \beta_+ < \frac{2}{3} \quad \text{such that } \sup_{\rho \in \mathbb{R}_+} \frac{|\epsilon'_{xc}(\rho)|}{\rho^{\beta_-} + \rho^{\beta_+}} < \infty, \quad (6)$$

$$\text{there exists } 1 \leq \alpha < \frac{3}{2} \quad \text{such that } \limsup_{\rho \rightarrow 0_+} \frac{\epsilon_{xc}(\rho)}{\rho^\alpha} < 0, \quad (7)$$

$$\text{there exists } \lambda > -1 \quad \text{such that } \epsilon''_{xc}(\rho) \underset{\rho \rightarrow 0_+}{\sim} c\rho^\lambda. \quad (8)$$

Note that these properties are satisfied by the exact function  $\epsilon_{xc}$ . They are also satisfied by Slater's  $X\alpha$  model for which  $\epsilon_{xc}(\rho) = -C_D\rho^{1/3}$ , where  $C_D = \frac{3}{4}(\frac{3}{\pi})^{1/3}$  is the Dirac constant. This model is used in the simulations reported in Section 4.

**Remark.** The minimization set  $\mathcal{H}_N$  defined in (2) is the set of *real spin-unpolarized* first-order reduced density matrices. We will call its elements *nonmagnetic states*. The general (complex noncollinear spin-polarized; see, e.g., [18]) rHF model being convex in the density matrix, and strictly convex in the density, the general rHF ground state density of a given molecular system in the absence of magnetic field, if it exists, is unique, and one of the minimizers is a nonmagnetic state. Indeed, using the notation of [18], if  $\gamma$  is a complex noncollinear spin-polarized ground state, the nonmagnetic state

$$\gamma_0 := \frac{1}{4}(\gamma^{\uparrow\uparrow} + \overline{\gamma^{\uparrow\uparrow}} + \gamma^{\downarrow\downarrow} + \overline{\gamma^{\downarrow\downarrow}}),$$

where  $\overline{\gamma^{\sigma,\sigma}}$  is the complex conjugate (not the adjoint) of the operator  $\gamma^{\sigma,\sigma}$ , is a nonmagnetic ground state. The general rHF ground state energy and density of a molecular system in the absence of magnetic field can therefore be determined by minimizing the rHF energy functional over the set  $\mathcal{H}_N$ . The LDA model is not a priori strictly convex in the density, but it is convex over the set of complex noncollinear spin-polarized density matrices having a given density  $\rho$ . Therefore, the general LDA ground state energy and densities can be obtained by minimizing the LDA energy functional over the set  $\mathcal{H}_N$ . In contrast, this argument does not apply to the local spin density approximation (LSDA) model, whose ground states are, in general, spin-polarized.

To avoid ambiguity, for any  $Z$  and  $N$  in  $\mathbb{R}_+^*$ , we denote

$$\mathcal{J}_{Z,N}^{\text{rHF}} := \inf\{E_{Z,N}^{\text{rHF}}(\gamma) \mid \gamma \in \mathcal{H}_N\}, \quad (9)$$

where

$$E_{Z,N}^{\text{rHF}}(\gamma) := \text{Tr}(-\frac{1}{2}\Delta\gamma) - Z \int_{\mathbb{R}^3} \frac{\rho_\gamma}{|\cdot|} + \frac{1}{2}D(\rho_\gamma, \rho_\gamma),$$

and

$$\mathcal{J}_{Z,N}^{\text{LDA}} := \inf\{E_{Z,N}^{\text{LDA}}(\gamma) \mid \gamma \in \mathcal{H}_N\}, \quad (10)$$



where

$$E_{Z,N}^{\text{LDA}}(\gamma) := \text{Tr}(-\frac{1}{2}\Delta\gamma) - Z \int_{\mathbb{R}^3} \frac{\rho_\gamma}{|\cdot|} + \frac{1}{2}D(\rho_\gamma, \rho_\gamma) + E_{\text{xc}}^{\text{LDA}}(\rho_\gamma).$$

We recall the following two theorems which ensure the existence of ground states for neutral atoms and positive ions.

**Theorem 1** (ground state for the rHF model [7; 35]). *Let  $Z \in \mathbb{R}_+^*$  and  $N \leq Z$ . Then the minimization problem (9) has a ground state  $\gamma_{Z,N}^{0,\text{rHF}}$ , and all the ground states share the same density  $\rho_{Z,N}^{0,\text{rHF}}$ . The mean-field Hamiltonian*

$$H_{Z,N}^{0,\text{rHF}} := -\frac{1}{2}\Delta - \frac{Z}{|\cdot|} + V^{\text{H}}(\rho_{Z,N}^{0,\text{rHF}})$$

is a bounded-below self-adjoint operator on  $L^2(\mathbb{R}^3)$ ,  $\sigma_{\text{ess}}(H_{Z,N}^{0,\text{rHF}}) = \mathbb{R}_+$ , and the ground state  $\gamma_{Z,N}^{0,\text{rHF}}$  is of the form

$$\gamma_{Z,N}^{0,\text{rHF}} = 2\mathbb{1}_{(-\infty, \epsilon_{Z,N,\text{F}}^{0,\text{rHF}})}(H_{Z,N}^{0,\text{rHF}}) + \delta_{Z,N}^{0,\text{rHF}},$$

where  $\epsilon_{Z,N,\text{F}}^{0,\text{rHF}} \leq 0$  is the Fermi level,  $\text{Ran}(\delta_{Z,N}^{0,\text{rHF}}) \subset \text{Ker}(H_{Z,N}^{0,\text{rHF}} - \epsilon_{Z,N,\text{F}}^{0,\text{rHF}})$ , and  $0 \leq \delta_{Z,N}^{0,\text{rHF}} \leq 2$ . If  $\epsilon_{Z,N,\text{F}}^{0,\text{rHF}}$  is negative and is not an accidentally degenerate eigenvalue of  $H_{Z,N}^{0,\text{rHF}}$ , then the nonmagnetic ground state  $\gamma_{Z,N}^{0,\text{rHF}}$  is unique.

Our numerical results indicate that, for neutral atoms, the assumption

$$\epsilon_{Z,Z,\text{F}}^{0,\text{rHF}} \text{ is negative and is not an accidentally degenerate eigenvalue of } H_{Z,Z}^{0,\text{rHF}}$$

is satisfied for most chemical elements of the first four rows, but not for all of them. We will elaborate on this observation in Section 4.1.1.

**Theorem 2** (ground state for the LDA model [1]). *Let  $Z \in \mathbb{R}_+^*$  and  $N \leq Z$ . Suppose that (5)–(7) hold. Then the minimization problem (10) has a ground state  $\gamma_{Z,N}^{0,\text{LDA}}$ . In addition,  $\gamma_{Z,N}^{0,\text{LDA}}$  satisfies the self-consistent field equation*

$$\gamma_{Z,N}^{0,\text{LDA}} = 2\mathbb{1}_{(-\infty, \epsilon_{Z,N,\text{F}}^{0,\text{LDA}})}(H_{Z,N}^{0,\text{LDA}}) + \delta_{Z,N}^{0,\text{LDA}}, \quad (11)$$

where  $\epsilon_{Z,N,\text{F}}^{0,\text{LDA}} \leq 0$  is the Fermi level,  $\text{Ran}(\delta_{Z,N}^{0,\text{LDA}}) \subset \text{Ker}(H_{Z,N}^{0,\text{LDA}} - \epsilon_{Z,N,\text{F}}^{0,\text{LDA}})$ ,  $0 \leq \delta_{Z,N}^{0,\text{LDA}} \leq 2$ , and the mean-field Hamiltonian

$$H_{Z,N}^{0,\text{LDA}} := -\frac{1}{2}\Delta - \frac{Z}{|\cdot|} + V^{\text{H}}(\rho_{Z,N}^{0,\text{LDA}}) + v_{\text{xc}}(\rho_{Z,N}^{0,\text{LDA}}),$$

where  $\rho_{Z,N}^{0,\text{LDA}} = \rho_{\gamma_{Z,N}^{0,\text{LDA}}}$  and  $v_{\text{xc}}(\rho) = \frac{d\epsilon_{\text{xc}}}{d\rho}(\rho)$ , is a bounded-below self-adjoint operator on  $L^2(\mathbb{R}^3)$  and  $\sigma_{\text{ess}}(H_{Z,N}^{0,\text{LDA}}) = \mathbb{R}_+$ .

**2.2. Density functional perturbation theory.** We now examine the response of the ground state density matrix when an additional external potential  $\beta W$  is turned on. The energy functional to be minimized over  $\mathcal{H}_N$  now reads

$$\tilde{E}_{Z,N}^{\text{rHF/LDA}}(\gamma, \beta W) := E_{Z,N}^{\text{rHF/LDA}}(\gamma) + \int_{\mathbb{R}^3} \beta W \rho_\gamma, \quad (12)$$

and is well defined for any  $\gamma \in \mathcal{H}_N$ ,  $W \in \mathcal{C}'$ , and  $\beta \in \mathbb{R}$ . The parameter  $\beta$  is called the coupling constant in quantum mechanics. Denote by

$$\tilde{\mathcal{F}}_{Z,N}^{\text{rHF/LDA}}(\beta W) := \inf\{\tilde{E}_{Z,N}^{\text{rHF/LDA}}(\gamma, \beta W) \mid \gamma \in \mathcal{H}_N\}. \quad (13)$$

The following theorem ensures the existence of a perturbed ground state density matrix for perturbation potentials in  $\mathcal{C}'$ .

**Theorem 3** (existence of a perturbed minimizer [7]). *Let  $Z \in \mathbb{R}_+^*$ ,  $N \leq Z$ , and  $W \in \mathcal{C}'$ . Assume that the Fermi level  $\epsilon_{Z,N,F}^{0,\text{rHF}}$  is negative and is not an accidentally degenerate eigenvalue of  $H_{Z,N}^{0,\text{rHF}}$ . Then the nonmagnetic unperturbed rHF ground state, that is, the minimizer of (9), is unique, and the perturbed problem (13) has a unique nonmagnetic ground state  $\gamma_{Z,N,\beta W}^{\text{rHF}}$ , for  $\beta \in \mathbb{R}$  small enough. The Hamiltonian*

$$H_{Z,N,\beta W}^{\text{rHF}} = -\frac{1}{2}\Delta - \frac{Z}{|\cdot|} + V^{\text{H}}(\rho_{Z,N,\beta W}^{\text{rHF}}) + \beta W, \quad (14)$$

where  $\rho_{Z,N,\beta W}^{\text{rHF}} = \rho_{\gamma_{Z,N,\beta W}^{\text{rHF}}}$ , is a bounded-below self-adjoint operator on  $L^2(\mathbb{R}^3)$  with form domain  $H^1(\mathbb{R}^3)$  and  $\sigma_{\text{ess}}(H_{Z,N,\beta W}^{0,\text{rHF}}) = \mathbb{R}_+$ . Moreover,  $\gamma_{Z,N,\beta W}^{\text{rHF}}$  and  $\rho_{Z,N,\beta W}^{\text{rHF}}$  are analytic in  $\beta$ ; that is,

$$\gamma_{Z,N,\beta W}^{\text{rHF}} = \sum_{k \geq 0} \beta^k \gamma_{Z,N,W}^{(k),\text{rHF}} \quad \text{and} \quad \rho_{Z,N,\beta W}^{\text{rHF}} = \sum_{k \geq 0} \beta^k \rho_{Z,N,W}^{(k),\text{rHF}},$$

the above series being normally convergent in  $\mathfrak{S}_{1,1}$  and  $\mathcal{C}$ , respectively.

In the sequel, we will refer to  $\gamma_{Z,N,W}^{(k)}$  as the  $k$ -th-order perturbation of the density matrix.

Although we focus here on nonmagnetic states, it is convenient to consider  $H_{Z,N}^{0,\text{rHF}}$  as an operator on  $L^2(\mathbb{R}^3, \mathbb{C})$  in order to expand the angular part of the atomic orbitals on the usual complex spherical harmonics. It would of course have been possible to avoid considering complex wave functions by expanding on real spherical harmonics. However, we have chosen to work with complex wave functions to prepare for future works on magnetic systems.

The unperturbed Hamiltonian  $H_{Z,N}^{0,\text{rHF}}$  is a self-adjoint operator on  $L^2(\mathbb{R}^3, \mathbb{C})$  invariant with respect to rotations around the nucleus (assumed located at the origin). This operator is therefore block-diagonal in the decomposition of  $L^2(\mathbb{R}^3, \mathbb{C})$  as the

direct sum of the pairwise orthogonal subspaces  $\mathcal{H}_l := \text{Ker}(L^2 - l(l+1))$ :

$$L^2(\mathbb{R}^3, \mathbb{C}) = \bigoplus_{l \in \mathbb{N}} \mathcal{H}_l,$$

where  $L = \mathbf{r} \times (-i\nabla)$  is the angular momentum operator. Since we are going to consider perturbation potentials which are not spherically symmetric, but only cylindrically symmetric, or in other words independent of the azimuthal angle  $\varphi$  in spherical coordinates, the  $\mathcal{H}_l$  are no longer invariant subspaces of the perturbed Hamiltonians. The appropriate decomposition of  $L^2(\mathbb{R}^3, \mathbb{C})$  into invariant subspaces for Hamiltonians  $H_{Z,N,\beta W}^{\text{HF}}$  with  $W$  cylindrically symmetric is the following: for  $m \in \mathbb{Z}$ , we set

$$\mathcal{H}^m := \text{Ker}(L_z - m),$$

where  $L_z$  is the  $z$ -component of the angular momentum operator  $L$  ( $L_z = L \cdot \mathbf{e}_z$ ).

Note that,

$$\text{for all } l \in \mathbb{N}, \quad \mathcal{H}_l = \left\{ \phi \in L^2(\mathbb{R}^3, \mathbb{C}) \mid \phi(r, \theta, \varphi) = \sum_{-l \leq m \leq l} R^m(r) Y_l^m(\theta, \varphi) \right\},$$

and,

$$\text{for all } m \in \mathbb{Z}, \quad \mathcal{H}^m = \left\{ \phi \in L^2(\mathbb{R}^3, \mathbb{C}) \mid \phi(r, \theta, \varphi) = \sum_{l \geq |m|} R_l(r) Y_l^m(\theta, \varphi) \right\},$$

where  $Y_l^m$  are the spherical harmonics, i.e., the joint eigenfunctions of  $\Delta_S$ , the Laplace–Beltrami operator on the unit sphere  $\mathbb{S}^2$  of  $\mathbb{R}^3$ , and  $\mathcal{L}_z$  the generator of rotations about the azimuthal axis of  $\mathbb{S}^2$ . More precisely, we have

$$-\Delta_S Y_l^m = l(l+1) Y_l^m \quad \text{and} \quad \mathcal{L}_z Y_l^m = m Y_l^m,$$

where, in spherical coordinates,

$$\Delta_S = \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \varphi^2} \quad \text{and} \quad \mathcal{L}_z = -i \frac{\partial}{\partial \varphi}.$$

These functions are orthonormal, in the sense that

$$\int_{\mathbb{S}^2} Y_l^m (Y_{l'}^{m'})^* = \int_{\theta=0}^{\pi} \int_{\varphi=0}^{2\pi} Y_l^m(\theta, \varphi) (Y_{l'}^{m'}(\theta, \varphi))^* \sin \theta \, d\theta \, d\varphi = \delta_{ll'} \delta_{mm'}, \quad (15)$$

where  $\delta_{ij}$  is the Kronecker symbol and  $(Y_l^m)^* = (-1)^m Y_l^{-m}$  is the complex conjugate of  $Y_l^m$ .

We also define

$$\mathcal{V}^m := \mathcal{H}^m \cap H^1(\mathbb{R}^3, \mathbb{C}),$$

so that  $L^2(\mathbb{R}^3, \mathbb{C})$  and  $H^1(\mathbb{R}^3, \mathbb{C})$  are decomposed as the direct sums

$$L^2(\mathbb{R}^3, \mathbb{C}) = \bigoplus_{m \in \mathbb{Z}} \mathcal{H}^m \quad \text{and} \quad H^1(\mathbb{R}^3, \mathbb{C}) = \bigoplus_{m \in \mathbb{Z}} \mathcal{V}^m, \quad (16)$$

each  $\mathcal{H}^m$  being  $H_{Z,N,\beta W}^{\text{rHF}}$ -stable (in the sense of unbounded operators) for  $W$  cylindrically symmetric. This is due to the fact that, for  $W$  cylindrically symmetric, the operator  $H_{Z,N,\beta W}^{\text{rHF}}$  commutes with  $L_z$ . Note that  $\sigma(H_{Z,N,\beta W}^{\text{rHF}}) = \bigcup_{m \in \mathbb{Z}} \overline{\sigma(H_{Z,N,\beta W}^{\text{rHF}}|_{\mathcal{H}^m})}$ . Same arguments hold true for  $H_{Z,N,\beta W}^{\text{LDA}}$  under the assumption that the ground state density  $\rho_{Z,N,\beta W}^{0,\text{LDA}}$  is cylindrically symmetric (which is the case whenever it is unique).

We are interested in the Stark potential

$$W_{\text{Stark}}(\mathbf{r}) = -e_z \cdot \mathbf{r}, \quad (17)$$

which does not belong to  $\mathcal{C}'$ , and thus does not fall under the scope of Theorem 3. We therefore introduce the classes of perturbation potentials

$$\mathcal{W}_s := \left\{ W \in \mathcal{H}_{\text{loc}}^0 \mid \int_{\mathbb{R}^3} \frac{|W(\mathbf{r})|^2}{(1+|\mathbf{r}|^2)^s} d\mathbf{r} < \infty \right\},$$

where  $\mathcal{H}_{\text{loc}}^0 := \mathcal{H}^0 \cap L_{\text{loc}}^2(\mathbb{R}^3)$ , which contain the Stark potential  $W_{\text{Stark}}$  whenever  $s > \frac{5}{2}$ . For  $W \in \mathcal{W}_s \setminus \mathcal{C}'$ , the energy functional (12) is not necessarily bounded below on  $\mathcal{H}_N$  for  $\beta \neq 0$ . Thus, the solution of (13) may not exist. This is the case for the Stark potential  $W_{\text{Stark}}$ . However, the  $k$ -th-order perturbation of the ground state may exist, as this is the case when the linear Schrödinger operator of the hydrogen atom is perturbed by the Stark potential  $W_{\text{Stark}}$  (see, e.g., [31]). The following theorem ensures the existence of the first-order perturbation of the density matrix.

**Theorem 4** (first-order density functional perturbation theory [8]). *Let  $Z \in \mathbb{R}_+^*$ ,  $0 < N \leq Z$ , such that  $\epsilon_{Z,N,F}^{0,\text{rHF}}$  is negative<sup>1</sup> and is not an accidentally degenerate eigenvalue of  $H_{Z,N}^{0,\text{rHF}}$ ,  $s \in \mathbb{R}$ , and  $W \in \mathcal{W}_s$ . In the rHF framework, the first-order perturbation of the density matrix  $\gamma_{Z,N,W}^{(1),\text{rHF}}$  is well defined in  $\mathfrak{S}_{1,1}$ .*

Note that assumption (8) is used to establish the existence and uniqueness of the first-order perturbation of the density matrix  $\gamma_{Z,N,W}^{(1),\text{LDA}}$  in  $\mathfrak{S}_{1,1}$ .

### 3. Numerical method

In this section, we present the discretization method and the algorithms we used to calculate numerically the ground state density matrices for (9), (10), and (13) for cylindrically symmetric perturbation potentials  $W$ , together with the ground state energy and the lowest eigenvalues of the associated Kohn–Sham operator. From now on, we make the assumption that the ground state density of (13), if it exists,

<sup>1</sup>Note that  $\epsilon_{Z,N,F}^{0,\text{rHF}} < 0$  whenever  $0 < N < Z$  (see, e.g., [35]).

is cylindrically symmetric, which is always the case for the rHF model. Using spherical coordinates, we can write

$$W(r, \theta) = \sum_{l=0}^{+\infty} W_l(r) Y_l^0(\theta) \in \mathfrak{H}^0$$

(since  $Y_l^0$  is independent of  $\varphi$ , we use the notation  $Y_l^0(\theta)$  instead of  $Y_l^0(\theta, \varphi)$ ). As the ground state density  $\rho_{Z,N,\beta W}$  is assumed to be cylindrically symmetric as well, one has

$$\rho_{Z,N,\beta W}(r, \theta) = \sum_{l=0}^{+\infty} \rho_{Z,N,\beta W,l}(r) Y_l^0(\theta).$$

The Hartree and the exchange-correlation potentials also have the same symmetry. For  $\rho \in L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3) \cap \mathfrak{H}^0$ , we have

$$V^H(\rho)(r, \theta) = \sum_{l=0}^{+\infty} V_{\rho_l}^H(r) Y_l^0(\theta) \quad \text{and} \quad v_{\text{xc}}(\rho)(r, \theta) = \sum_{l=0}^{+\infty} (v_{\rho}^{\text{xc}})_l(r) Y_l^0(\theta),$$

where, for each  $l \geq 0$ ,  $V_{\rho_l}^H(r)$  solves the differential equation

$$-\frac{1}{r} \frac{d^2}{dr^2} (r V_{\rho_l}^H) + \frac{l(l+1)}{r^2} V_{\rho_l}^H = 4\pi \rho_l$$

with boundary conditions

$$\lim_{r \rightarrow 0^+} r V_{\rho_l}^H(r) = 0 \quad \text{and} \quad \lim_{r \rightarrow +\infty} r V_{\rho_l}^H(r) = \left( 4\pi \int_0^{+\infty} r^2 \rho_0(r) dr \right) \delta_{l0},$$

while  $(v_{\rho}^{\text{xc}})_l$  can be computed by projection on the spherical harmonics  $Y_l^0$ :

$$(v_{\rho}^{\text{xc}})_l(r) = 2\pi \int_0^{\pi} v_{\text{xc}}(\rho)(r, \theta) Y_l^0(\theta) \sin \theta d\theta.$$

**3.1. Discretization of the Kohn–Sham model.** Recall that for  $W \in \mathcal{W}_s$  and  $\beta \neq 0$ , the energy functional defined by (12) is not necessarily bounded below on  $\mathfrak{H}_N$ , which implies in particular that (13) may have no ground state. Nevertheless, one can compute approximations of (13) in finite-dimensional spaces, provided that the basis functions decay fast enough at infinity. Let  $N_h \in \mathbb{N}^*$  and  $m_h \geq m_Z^* := \max\{m \mid \text{there exists } k > 0, \epsilon_{m,k}^0 \leq \epsilon_{Z,N,F}^0\}$ , and let  $\{\mathfrak{X}_i\}_{1 \leq i \leq N_h} \in (H_0^1(0, +\infty))^{N_h}$  be a free family of real-valued basis functions. We then introduce the finite-dimensional spaces

$$\mathfrak{V}^{m,h} := \mathfrak{V}^m \cap \text{span}_{\mathbb{R}} \left( \frac{\mathfrak{X}_i(r)}{r} Y_l^m(\theta, \phi) \right)_{\substack{1 \leq i \leq N_h \\ |m| \leq l \leq m_h}} \subset H^1(\mathbb{R}^3, \mathbb{C}) \quad (18)$$

and

$$\mathcal{X}^h = \text{span}_{\mathbb{R}}(\mathcal{X}_1, \dots, \mathcal{X}_{N_h}) \subset H_0^1(0, +\infty), \quad (19)$$

and the set

$$\mathcal{H}_{N,h} := \left\{ \gamma \in \mathcal{H}_N \mid \gamma = \sum_{m=-m_h}^{m_h} \gamma^m, \gamma^m \in \mathcal{G}(\mathcal{X}^m), \text{Ran}(\gamma^m) \subset \mathcal{V}^{m,h} \right\} \subset \mathcal{H}_N.$$

Note that since our goal is to compute nonmagnetic ground states, we are allowed to limit ourselves to real linear combinations in (18) and (19).

**3.1.1. Variational approximation.** A variational approximation of (13) is obtained by minimizing the energy functional (12) over the approximation set  $\mathcal{H}_{N,h}$ :

$$\tilde{\mathcal{F}}_{Z,N,h}^{\text{rHF/LDA}}(\beta W) := \inf\{\tilde{E}_{Z,N}^{\text{rHF/LDA}}(\gamma_h, \beta W) \mid \gamma_h \in \mathcal{H}_{N,h}\}. \quad (20)$$

Any  $\gamma_h \in \mathcal{H}_{N,h}$  can be written as

$$\gamma_h = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} n_{m,k} |\Phi_{m,k,h}\rangle \langle \Phi_{m,k,h}|, \quad (21)$$

with

$$\begin{aligned} \Phi_{m,k,h} \in \mathcal{V}^{m,h}, \quad \int_{\mathbb{R}^3} \Phi_{m,k,h} \Phi_{m',k',h}^* = \delta_{kk'}, \quad \Phi_{-m,k,h} = (-1)^m \Phi_{m,k,h}^*, \\ 0 \leq n_{m,k} = n_{-m,k} \leq 2, \quad \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} n_{m,k} = N. \end{aligned}$$

The functions  $\Phi_{m,k,h}$ , being in  $\mathcal{V}^{m,h}$ , are of the form

$$\Phi_{m,k,h}(r, \theta, \varphi) = \sum_{l=|m|}^{m_h} \frac{u_l^{m,k,h}(r)}{r} Y_l^m(\theta, \varphi), \quad (22)$$

where for each  $-m_h \leq m \leq m_h$ ,  $1 \leq k \leq (m_h - |m| + 1)N_h$  and  $|m| \leq l \leq m_h$ ,  $u_l^{m,k,h} \in \mathcal{X}^h$ . Note that  $u_l^{-m,k,h} = u_l^{m,k,h}$ . Expanding the functions  $u_l^{m,k,h}$  in the basis  $(\mathcal{X}_i)_{1 \leq i \leq N_h}$  as

$$u_l^{m,k,h}(r) = \sum_{i=1}^{N_h} U_{i,l}^{m,k} \mathcal{X}_i(r), \quad (23)$$

and gathering the coefficients  $U_{i,l}^{m,k}$  for fixed  $m$  and  $k$  in a rectangular matrix  $U^{m,k} \in \mathbb{R}^{N_h \times (m_h - |m| + 1)}$ , any  $\gamma_h \in \mathcal{H}_{N,h}$  can be represented via (21)–(23) by at least one element of the set

$$\mathcal{M}_{N,h} := \mathcal{U}_h \times \mathcal{N}_{N,h}, \quad (24)$$

where

$$\mathcal{U}_h := \left\{ (U^{m,k})_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} \mid U^{m,k} = U^{-m,k} \in \mathbb{R}^{N_h \times (m_h - |m| + 1)}, \operatorname{Tr}([U^{m,k}]^T M_0 U^{m,k'}) = \delta_{kk'} \right\},$$

and

$$\mathcal{N}_{N,h} := \left\{ (n_{m,k})_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} \mid 0 \leq n_{m,k} = n_{-m,k} \leq 2, \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} n_{m,k} = N \right\}.$$

The matrix  $M_0$  appearing in the definition of  $\mathcal{U}_h$  is the mass matrix defined by

$$[M_0]_{ij} = \int_0^{+\infty} \mathcal{X}_i \mathcal{X}_j,$$

and the constraints  $\operatorname{Tr}([U^{m,k}]^T M_0 U^{m,k'}) = \delta_{kk'}$  come from the fact that

$$\begin{aligned} \int_{\mathbb{R}^3} \Phi_{m,k,h} \Phi_{m,k',h}^* &= \int_0^{+\infty} \int_{\mathbb{S}^2} \left( \sum_{l=|m|}^{m_h} \sum_{i=1}^{N_h} U_{i,l}^{m,k} \frac{\mathcal{X}_i(r)}{r} Y_l^m(\sigma) \right) \\ &\quad \times \left( \sum_{l'=|m|}^{m_h} \sum_{j=1}^{N_h} U_{j,l'}^{m,k'} \frac{\mathcal{X}_j(r)}{r} Y_{l'}^m(\sigma)^* \right) r^2 d\sigma dr \\ &= \sum_{l=|m|}^{m_h} \sum_{i,j=1}^{N_h} U_{i,l}^{m,k} [M_0]_{ij} U_{j,l}^{m,k'} = \operatorname{Tr}([U^{m,k}]^T M_0 U^{m,k'}). \end{aligned}$$

**Remark.** An interesting observation is that, if there is no accidental degeneracy in the set of the occupied energy levels of  $H_{Z,N}^{0,\text{rHF/LDA}}$ , and if the occupied orbitals are well enough approximated in the space  $\mathcal{V}^{m,h}$ , then the approximate ground state density matrix  $\gamma_{Z,N,h}^{0,\text{rHF/LDA}}$  has a unique representation of the form (21)–(23), up to the signs and the numbering of the functions  $u_l^{m,k,h}$ , that is, up to the signs and numbering of the column vectors of the matrices  $U^{m,k}$ . By continuity, this uniqueness of the representation will survive if a small-enough cylindrically symmetric perturbation is turned on. This is the reason why this representation is well suited to our study.

Let us now express each component of the energy functional  $\tilde{E}_{Z,N}^{\text{rHF,LDA}}(\gamma_h, \beta W)$  using the representation (21)–(23) of the elements of  $\mathcal{K}_{N,h}$ . For this purpose, we introduce the  $N_h \times N_h$  real symmetric matrices  $A$  and  $M_n$ ,  $n = -2, -1, 0, 1$ , with entries

$$A_{ij} = \int_0^{+\infty} \mathcal{X}_i' \mathcal{X}_j' \quad \text{and} \quad [M_n]_{ij} = \int_0^{+\infty} r^n \mathcal{X}_i(r) \mathcal{X}_j(r) dr. \quad (25)$$

The weighted mass matrices  $M_{-2}$  and  $M_{-1}$  are well defined in view of the Hardy inequality,

$$\text{for all } u \in H_0^1(0, +\infty), \quad \int_0^{+\infty} \frac{u^2(r)}{r^2} dr \leq 4\pi \int_0^{+\infty} |u'|^2.$$

We assume from now on that the basis functions  $\mathcal{X}_i$  decay fast enough at infinity for the weighted mass matrix  $M_1$  to be well defined.

In the representation (21)–(23), the kinetic energy is equal to

$$\frac{1}{2} \text{Tr}(-\Delta \gamma_h) = \frac{1}{2} \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} n_{m,k} \left( \text{Tr}([U^{m,k}]^T A U^{m,k}) + \text{Tr}(D_m [U^{m,k}]^T M_{-2} U^{m,k}) \right), \quad (26)$$

where  $D_m \in \mathbb{R}^{(m_h - |m| + 1) \times (m_h - |m| + 1)}$  is the diagonal matrix defined by

$$D_m = \text{diag}(|m|(|m| + 1), \dots, m_h(m_h + 1)). \quad (27)$$

All the other terms in the energy functional depending on the density

$$\rho_h := \rho_{\gamma_h} = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) N_h}} n_{m,k} |\Phi_{m,k,h}|^2, \quad (28)$$

we first need to express this quantity as a function of the matrices  $U^{m,k}$  and the occupation numbers  $n_{m,k}$ . As the function  $\rho_h$  is in  $\mathcal{H}^0$ , we have

$$\rho_h(r, \theta) = \sum_{l=0}^{2m_h} \rho_l^h(r) Y_l^0(\theta). \quad (29)$$

Inserting (22) in (28), we get

$$\rho_h(r, \theta) = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) N_h}} n_{m,k} \left| \sum_{l=|m|}^{m_h} \frac{u_l^{m,k,h}(r)}{r} Y_l^m(\theta, \varphi) \right|^2. \quad (30)$$

We recall the equality [32]

$$Y_{l_1}^m (Y_{l_2}^m)^* = (-1)^m Y_{l_1}^m Y_{l_2}^{-m} = \sum_{l_3=|l_1-l_2|}^{l_1+l_2} c_{l_1, l_2, l_3}^m Y_{l_3}^0, \quad (31)$$

with

$$c_{l_1, l_2, l_3}^m = (-1)^m \sqrt{\frac{(2l_1 + 1)(2l_2 + 1)(2l_3 + 1)}{4\pi}} \begin{pmatrix} l_1 & l_2 & l_3 \\ m & -m & 0 \end{pmatrix} \begin{pmatrix} l_1 & l_2 & l_3 \\ 0 & 0 & 0 \end{pmatrix},$$



where  $\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix}$  denote the Wigner 3j-symbols. Inserting the expansion (23) in (30) and using (31) and the fact that

$$\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} = 0 \quad \text{unless } |l_1 - l_2| \leq l_3 \leq l_1 + l_2,$$

we obtain

$$\rho_h(r, \theta) = \sum_{l=0}^{2m_h} \left[ \sum_{i,j=1}^{N_h} \left( \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} n_{m,k} \sum_{l',l''=|m|}^{m_h} c_{l',l'',l}^m U_{i,l'}^{m,k} U_{j,l''}^{m,k} \right) \times \frac{\mathcal{X}_i(r)}{r} \frac{\mathcal{X}_j(r)}{r} \right] Y_l^0(\theta),$$

from which we conclude that

$$\rho_l^h(r) = \sum_{i,j=1}^{N_h} \left( \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} n_{m,k} \sum_{l',l''=|m|}^{m_h} c_{l',l'',l}^m U_{i,l'}^{m,k} U_{j,l''}^{m,k} \right) \frac{\mathcal{X}_i(r)}{r} \frac{\mathcal{X}_j(r)}{r}.$$

For  $0 \leq l \leq 2m_h$ , we introduce the matrix  $R_l \in \mathbb{R}^{N_h \times N_h}$  defined by

$$R_l := \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} n_{m,k} U^{m,k} C^{l,m} [U^{m,k}]^T \quad (32)$$

where  $C^{l,m} \in \mathbb{R}^{(m_h - |m| + 1) \times (m_h - |m| + 1)}$  is the symmetric matrix<sup>2</sup> defined by,

$$\text{for all } |m| \leq l \leq 2m_h, \quad C_{l',l''}^{l,m} = \sqrt{4\pi} c_{l',l'',l}^m, \quad (33)$$

so that

$$\rho_h(r, \theta) = \frac{1}{\sqrt{4\pi}} \sum_{l=0}^{2m_h} \sum_{i,j=1}^{N_h} [R_l]_{i,j} \frac{\mathcal{X}_i(r)}{r} \frac{\mathcal{X}_j(r)}{r} Y_l^0(\theta). \quad (34)$$

Note that  $C^{0,m}$  is the identity matrix, so that

$$R_0 = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} n_{m,k} U^{m,k} [U^{m,k}]^T$$

and

$$\text{Tr}(M_0 R_0) = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} n_{m,k} \text{Tr}(M_0 U^{m,k} [U^{m,k}]^T) = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} n_{m,k} = N,$$

<sup>2</sup>The symmetry of the matrix  $C^{lm}$  comes from the symmetry properties of the 3j-symbols

$$\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} = (-1)^{l_1 + l_2 + l_3} \begin{pmatrix} l_2 & l_1 & l_3 \\ m_2 & m_1 & m_3 \end{pmatrix} = (-1)^{l_1 + l_2 + l_3} \begin{pmatrix} l_2 & l_1 & l_3 \\ -m_2 & -m_1 & -m_3 \end{pmatrix}.$$

and that  $C^{1,m}$  is a symmetric tridiagonal matrix whose diagonal elements all are equal to zero.

The Coulomb attraction energy between the nucleus and the electrons then is equal to

$$\begin{aligned} -Z \int_{\mathbb{R}^3} \frac{\rho_h}{|\cdot|} &= -Z \int_0^{+\infty} \int_{\mathbb{S}^2} \frac{1}{r} \left( \frac{1}{\sqrt{4\pi}} \sum_{l=0}^{2m_h} \sum_{i,j=1}^{N_h} [R_l]_{i,j} \frac{\mathcal{X}_i(r)}{r} \frac{\mathcal{X}_j(r)}{r} Y_l^0(\sigma) \right) r^2 dr d\sigma \\ &= -Z \int_0^{+\infty} \int_{\mathbb{S}^2} \frac{1}{r} \left( \sum_{l=0}^{2m_h} \sum_{i,j=1}^{N_h} [R_l]_{i,j} \frac{\mathcal{X}_i(r)}{r} \frac{\mathcal{X}_j(r)}{r} Y_l^0(\sigma) \right) Y_0^0(\sigma)^* r^2 dr d\sigma \\ &= -Z \sum_{i,j=1}^{N_h} [R_0]_{i,j} [M_{-1}]_{ij} = -Z \operatorname{Tr}(M_{-1} R_0), \end{aligned}$$

where we have used the orthonormality condition (15) and the fact that  $Y_0^0 = 1/\sqrt{4\pi}$ .

Likewise, since  $Y_1^0(\theta) = \sqrt{3/(4\pi)} \cos(\theta)$ , the Stark potential (17) can be written in spherical coordinates as

$$W_{\text{Stark}}(r, \theta) = -\sqrt{\frac{4\pi}{3}} r Y_1^0(\theta) = -\sqrt{\frac{4\pi}{3}} r Y_1^0(\theta)^*,$$

and the potential energy due to the external electric field is then equal to

$$\beta \int_{\mathbb{R}^3} \rho_h W_{\text{Stark}} = -\frac{1}{\sqrt{3}} \beta \sum_{i,j=1}^{N_h} [R_1]_{ij} [M_1]_{ij} = -\frac{1}{\sqrt{3}} \beta \operatorname{Tr}(M_1 R_1).$$

Let  $\mu$  be a radial, continuous function from  $\mathbb{R}^3$  to  $\mathbb{R}$  vanishing at infinity and such that  $\int_{\mathbb{R}^3} \mu = 1$ . The Coulomb interaction energy can be rewritten as

$$\frac{1}{2} D(\rho_h, \rho_h) = \frac{1}{2} D\left(\rho_h - \left(\int_{\mathbb{R}^3} \rho_h\right) \mu, \rho_h - \left(\int_{\mathbb{R}^3} \rho_h\right) \mu\right) + N D(\mu, \rho_h) - \frac{N^2}{2} D(\mu, \mu). \quad (35)$$

The reason why we introduce the charge distribution  $\mu$  is to make neutral the charge distributions  $\rho_h - \left(\int_{\mathbb{R}^3} \rho_h\right) \mu$  in the first term of the right-hand side of (35), in such a way that the physical solution  $Q_{0,R_0}$  to (38) below for  $l=0$  is in  $H_0^1(0, +\infty)$ .

Introducing the real symmetric matrix  $V_\mu \in \mathbb{R}^{N_h \times N_h}$  with entries

$$[V_\mu]_{ij} = \int_0^{+\infty} [V^H(\mu)](r\mathbf{e}) \mathcal{X}_i(r) \mathcal{X}_j(r) dr, \quad (36)$$

where  $\mathbf{e}$  is any unit vector of  $\mathbb{R}^3$  (the value of  $V^H(\mu)(r\mathbf{e})$  is independent of  $\mathbf{e}$  since  $V^H(\mu)$  is radial), the sum of the last two terms of the right-hand side of (35) can

be rewritten as

$$ND(\mu, \rho_h) - \frac{N^2}{2}D(\mu, \mu) = N \operatorname{Tr}(V_\mu R_0) - \frac{N^2}{2}D(\mu, \mu).$$

Denoting by

$$\tilde{V}^H(\rho_h) = V^H(\rho_h - \left(\int_{\mathbb{R}^3} \rho_h\right)\mu),$$

we have by symmetry  $\tilde{V}^H(\rho_h) \in \mathfrak{H}^0$  and

$$[\tilde{V}^H(\rho_h)](r, \theta) = \sum_{l=0}^{2m_h} \tilde{V}_l(\rho_l^h)(r) Y_l^0(\theta) = \sum_{l=0}^{2m_h} \frac{Q_{l, R_l}(r)}{r} Y_l^0(\theta),$$

where  $Q_{l, R_l}$  is the unique solution in  $H_0^1(0, +\infty)$  to the differential equation

$$\begin{aligned} -\frac{d^2 Q_{l, R_l}}{dr^2}(r) + \frac{l(l+1)}{r^2} Q_{l, R_l}(r) \\ = 4\pi r \left( \left( \frac{1}{\sqrt{4\pi}} \sum_{i, j=1}^{N_h} [R_l]_{ij} \frac{\mathcal{X}_i(r) \mathcal{X}_j(r)}{r^2} \right) - N\mu(r) \delta_{l0} \right). \end{aligned} \quad (37)$$

Note that the mappings  $R_l \mapsto Q_{l, R_l}$  are linear. We therefore obtain

$$\begin{aligned} \frac{1}{2}D(\rho_h, \rho_h) = \frac{1}{2} \sum_{l=0}^{2m_h} \frac{1}{4\pi} \left( \int_0^{+\infty} \left( \left( \frac{dQ_{l, R_l}}{dr}(r) \right)^2 + \frac{l(l+1)}{r^2} Q_{l, R_l}(r)^2 \right) dr \right) \\ + N \operatorname{Tr}(V_\mu R_0) - \frac{N^2}{2}D(\mu, \mu). \end{aligned} \quad (38)$$

Finally, the exchange-correlation energy is

$$E_{\text{xc}}(\rho_h) = 2\pi \int_0^{+\infty} \int_0^\pi \epsilon_{\text{xc}} \left( \frac{1}{\sqrt{4\pi}} \sum_{l=0}^{2m_h} \sum_{i, j=1}^{N_h} [R_l]_{ij} \frac{\mathcal{X}_i(r)}{r} \frac{\mathcal{X}_j(r)}{r} Y_l^0(\theta) \right) r^2 \sin \theta \, dr \, d\theta. \quad (39)$$

**3.1.2. Approximation of the Hartree term.** Except for very specific basis functions (such as Gaussian atomic orbitals), it is not possible to evaluate exactly the first contribution to the Coulomb energy (38). It is therefore necessary to approximate it. For this purpose, we use a variational approximation of (37)–(38) in an auxiliary basis set  $\{\zeta_p\}_{1 \leq p \leq N_{h,a}} \in (H_0^1(0, +\infty))^{N_{h,a}}$ , which amounts to replacing  $\frac{1}{2}D(\rho_h, \rho_h)$  by its lower bound

$$\begin{aligned} \frac{1}{2}D_h(\rho_h, \rho_h) = \frac{1}{8\pi} \left( \int_0^{+\infty} \left( \left( \frac{dQ_{l, R_l}^h}{dr}(r) \right)^2 + \frac{l(l+1)}{r^2} Q_{l, R_l}^h(r)^2 \right) dr \right) \\ + N \operatorname{Tr}(V_\mu R_0) - \frac{N^2}{2}D(\mu, \mu), \end{aligned} \quad (40)$$

where  $Q_{l,R_l}^h$  is the unique solution in  $\zeta^h = \text{span}(\zeta_1, \dots, \zeta_{N_{h,a}})$  to the problem,

$$\begin{aligned} \text{for all } v_h \in \zeta^h, \quad & \int_0^{+\infty} \left( \frac{dQ_{l,R_l}^h}{dr}(r) \frac{dv_h}{dr}(r) + \frac{l(l+1)}{r^2} Q_{l,R_l}^h(r) v_h(r) \right) dr \\ & = 4\pi \int_0^{+\infty} r \left( \left( \frac{1}{\sqrt{4\pi}} \sum_{i,j=1}^{N_h} [R_l]_{ij} \frac{\mathcal{X}_i(r)\mathcal{X}_j(r)}{r^2} \right) - N\mu(r)\delta_{l0} \right) v_h(r) dr, \end{aligned}$$

which is nothing but the variational approximation of (37) in the finite-dimensional space  $\zeta^h$ . Expanding the functions  $Q_{l,R_l}^h$  in the basis set  $\{\zeta_k\}_{1 \leq k \leq N_{h,a}}$  as

$$Q_{l,R_l}^h(r) = \sum_{p=1}^{N_{h,a}} Q_{p,l} \zeta_p(r),$$

and collecting the coefficients  $Q_{p,l}$ ,  $1 \leq k \leq N_{h,a}$ , in a vector  $Q_l \in \mathbb{R}^{N_{h,a}}$ , we obtain that the vector  $Q_l$  is the solution to the linear system

$$(A^a + l(l+1)M_{-2}^a)Q_l = 4\pi(F : R_l - N\delta_{l0}G), \quad (41)$$

where the  $N_{h,a} \times N_{h,a}$  real symmetric matrices  $A^a$  and  $M_{-2}^a$  are defined by

$$A_{pq}^a = \int_0^{+\infty} \zeta_p' \zeta_q', \quad [M_{-2}^a]_{pq} = \int_0^{+\infty} \frac{\zeta_p(r)\zeta_q(r)}{r^2} dr, \quad (42)$$

where  $F \in \mathbb{R}^{N_{h,a} \times N_h \times N_h}$  is the three-index tensor with entries

$$F_{pij} = \frac{1}{\sqrt{4\pi}} \int_0^{+\infty} \frac{\mathcal{X}_i(r)\mathcal{X}_j(r)\zeta_p(r)}{r} dr, \quad (43)$$

and where  $G \in \mathbb{R}^{N_{h,a}}$  is the vector with entries

$$G_p = \int_0^{+\infty} r\mu(r)\zeta_p(r) dr. \quad (44)$$

Note that since  $N = \text{Tr}(M_0 R_0)$ , the mappings  $R_l \mapsto Q_l$  are in fact linear. We finally get

$$\frac{1}{2}D_h(\rho_h, \rho_h) = \frac{1}{8\pi} \sum_{l=0}^{2m_h} Q_l^T (A^a + l(l+1)M_{-2}^a) Q_l + N \text{Tr}(V_\mu R_0) - \frac{N^2}{2} D(\mu, \mu), \quad (45)$$

where  $Q_l$  is the solution to (41).

**3.1.3. Final form of the discretized problem and Euler–Lagrange equations.** We therefore end up with the following approximation of problem (13):

$$\begin{aligned} \mathfrak{J}_{Z,N,h}^{\text{rHF/LDA}}(\beta W) := \inf \{ & \mathcal{E}_{Z,N,\beta}^{\text{rHF/LDA}}((U^{m,k}), (n_{m,k})) \mid (U^{m,k})_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} \in \mathcal{U}_h, \\ & (n_{m,k})_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} \in \mathcal{N}_{N,h} \}, \quad (46) \end{aligned}$$

where

$$\begin{aligned}
 \mathcal{E}_{Z,N,\beta}^{\text{rHF/LDA}}((U^{m,k}), (n_{m,k})) \\
 &:= \frac{1}{2} \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} n_{m,k} (\text{Tr}([U^{m,k}]^T A U^{m,k}) + \text{Tr}(D_m [U^{m,k}]^T M_{-2} U^{m,k})) \\
 &\quad - Z \text{Tr}(M_{-1} R_0) + \frac{1}{8\pi} \sum_{l=0}^{2m_h} Q_l^T (A^a + l(l+1)M_{-2}^a) Q_l + N \text{Tr}(V_\mu R_0) \\
 &\quad - \frac{N^2}{2} D(\mu, \mu) + E_{\text{xc}}(\rho_h) - \frac{\beta}{\sqrt{3}} \text{Tr}(M_1 R_1),
 \end{aligned}$$

where for each  $l$ , the matrix  $R_l$  and the vector  $Q_l$  are defined by (32) and (41), respectively, and where the second-to-last term in the right-hand side is given by (39).

The gradient of  $\mathcal{E}_{Z,N,\beta}^{\text{rHF/LDA}}$  with respect to  $U^{m,k}$  is

$$\begin{aligned}
 \nabla_{U^{m,k}} \mathcal{E}_{Z,N,\beta}^{\text{rHF/LDA}} &= 2n_{m,k} \left( \frac{1}{2} A U^{m,k} + \frac{1}{2} M_{-2} U^{m,k} D_m - Z M_{-1} U^{m,k} + N V_\mu U^{m,k} \right. \\
 &\quad \left. + \sum_{l=0}^{2m_h} (Q_l^T \cdot F)(U^{m,k} C^{l,m}) + \sum_{l=0}^{2m_h} V_{\text{xc}}^l U^{m,k} C^{l,m} \right. \\
 &\quad \left. - \frac{\beta}{\sqrt{3}} M_1 U^{m,k} C^{1,m} \right),
 \end{aligned}$$

where for each  $0 \leq l \leq 2m_h$ , the  $N_h \times N_h$  real matrix  $V_{\text{xc}}^l$  is defined by

$$\begin{aligned}
 [V_{\text{xc}}^l]_{ij} &= \sqrt{\pi} \int_0^{+\infty} \int_0^\pi v_{\text{xc}} \left( \frac{1}{\sqrt{4\pi}} \sum_{i,j=1}^{N_h} [R_l]_{ij} \frac{\mathcal{X}_i(r) \mathcal{X}_j(r)}{r^2} \right) \\
 &\quad \times \mathcal{X}_i(r) \mathcal{X}_j(r) Y_l^0(\theta) \sin \theta \, dr \, d\theta, \quad (47)
 \end{aligned}$$

where  $v_{\text{xc}}(\rho) := \frac{d\epsilon_{\text{xc}}}{d\rho}(\rho)$  is the exchange-correlation potential.

Diagonalizing simultaneously the Kohn–Sham Hamiltonian and the ground state density matrix in an orthonormal basis, we obtain that the ground state can be obtained by solving the following system of first-order optimality conditions, which is nothing but a reformulation of the discretized extended Kohn–Sham equations

exploiting the cylindrical symmetry of the problem:

$$\begin{aligned} \frac{1}{2}AU^{m,k} + \frac{1}{2}M_{-2}U^{m,k}D_m - ZM_{-1}U^{m,k} + NV_\mu U^{m,k} + \sum_{l=0}^{2m_h} (Q_l^T \cdot F)(U^{m,k}C^{l,m}) \\ + \sum_{l=0}^{2m_h} V_{xc}^l U^{m,k}C^{l,m} - \frac{1}{\sqrt{3}}\beta M_1 U^{m,k}C^{1,m} = \epsilon_{m,k}M_0U^{m,k}, \end{aligned} \quad (48)$$

$$\text{Tr}([U^{m,k}]^T M_0 U^{m,k'}) = \delta_{kk'}, \quad (49)$$

$$(A^a + l(l+1)M_{-2}^a)Q_l = F : R_l - \text{Tr}(M_0 R_0)\delta_{l0}G, \quad (50)$$

$$\begin{aligned} [V_{xc}^l]_{ij} = \sqrt{\pi} \int_0^{+\infty} \int_0^\pi v_{xc} \left( \frac{1}{\sqrt{4\pi}} \sum_{i,j=1}^{N_h} [R_l]_{ij} \frac{\mathcal{X}_i(r)\mathcal{X}_j(r)}{r^2} \right) \\ \times \mathcal{X}_i(r)\mathcal{X}_j(r)Y_l^0(\theta) \sin\theta dr d\theta, \end{aligned} \quad (51)$$

$$n_{m,k} = 2 \text{ if } \epsilon_{m,k} < \epsilon_F, \quad 0 \leq n_{m,k} \leq 2 \text{ if } \epsilon_{m,k} = \epsilon_F, \quad n_{m,k} = 0 \text{ if } \epsilon_{m,k} > \epsilon_F, \quad (52)$$

$$\sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} n_{m,k} = N, \quad (53)$$

$$R_l = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} n_{m,k} U^{m,k} C^{l,m} [U^{m,k}]^T, \quad (54)$$

where the matrices  $A$ ,  $M_{-2}$ ,  $M_{-1}$ ,  $M_0$ ,  $M_1$ ,  $D_m$ ,  $V_\mu$ ,  $A^a$ ,  $M_{-2}^a$ , and  $C^{l,m}$ , the 3-index tensor  $F$ , and the vector  $G$  are defined by (25), (27), (33), (36), (42), (43), and (44).

**3.1.4.  $\mathbb{P}_4$  finite element method.** In our calculations, we use the same approximation space to discretize the radial components of the Kohn–Sham orbitals and the radial Poisson equations (37), so that, in our implementation of the method,  $N_{h,a} = N_h$  and  $\mathcal{X}^h = \zeta^h$ . We choose a cut-off radius  $L_e > 0$  large enough and discretize the interval  $[0, L_e]$  using a nonuniform grid with  $N_l + 1$  points  $0 = r_1 < r_2 < \dots < r_{N_l} < r_{N_l+1} = L_e$ . The positions of the points are chosen according to the rule

$$r_k = r_{k-1} + h_k, \quad h_{N_l} = \frac{1-s}{1-s^{N_l}}L_e, \quad h_{k-1} = sh_k,$$

where  $0 < s < 1$  is a scaling parameter leading to a progressive refinement of the mesh when one gets closer to the nucleus ( $r = 0$ ). To achieve the desired accuracy, we use the  $\mathbb{P}_4$  finite element method.

All the terms in the variational discretization of the energy and of the constraints can be computed exactly (up to finite arithmetic errors), except the exchange-correlation terms (39) and (47), which require a numerical quadrature method. In

our calculation, we use Gaussian quadrature formulas [37] of the form

$$\begin{aligned} \int_0^{+\infty} \int_0^\pi f(r, \theta) \sin \theta \, dr \, d\theta &= \int_0^{+\infty} \int_{-1}^1 f(r, \arccos t_\theta) \, dr \, dt_\theta \\ &\simeq \sum_{k=1}^{N_I} \sum_{i=1}^{N_{g,r}} \sum_{j=1}^{N_{g,\theta}} h_k w_{i,r} w_{j,\theta} f(r_k + h_k t_{i,r}, \arccos(t_{j,\theta})), \end{aligned}$$

where the  $0 < t_{1,r} < \dots < t_{N_{g,r},r} < 1$  and  $-1 < t_{1,\theta} < \dots < t_{N_{g,\theta},\theta} < 1$  are Gauss points for the  $r$ -variable and  $t_\theta$ -variable with associated weights  $w_{1,r}, \dots, w_{N_{g,r},r}$  and  $w_{1,\theta}, \dots, w_{N_{g,\theta},\theta}$ , respectively.

More details about the practical implementation of the method are provided in [9].

**3.2. Description of the algorithm.** In order to solve the self-consistent equations (48)–(54), we use an iterative algorithm. For clarity, we first present this algorithm within the continuous setting. Its formulation in the discretized setting considered here is detailed below. The iterations are defined as follows with an Ansatz of the ground state density  $\rho^{[n]}$  being known.

- (1) Construct the Kohn–Sham operator

$$H^{[n]} = -\frac{1}{2}\Delta - \frac{Z}{|\cdot|} + V^H(\rho^{[n]}) + v_{xc}(\rho^{[n]}) + \beta W$$

where  $v_{xc} = 0$  for the rHF model and  $v_{xc} = v_{xc}^{\text{LDA}}$  for the Kohn–Sham LDA model.

- (2) For each  $m \in \mathbb{Z}$ , compute the negative eigenvalues of  $H_m^{[n]} := \Pi_m H^{[n]} \Pi_m$ , where  $\Pi_m$  is the orthogonal projector on the space  $\mathcal{H}^m$ :

$$H_m^{[n]} \phi_{m,k}^{[n+1]} = \epsilon_{m,k}^{[n+1]} \phi_{m,k}^{[n+1]}, \quad \int_{\mathbb{R}^3} \phi_{m,k}^{[n+1]*} \phi_{m,k'}^{[n+1]} = \delta_{kk'}.$$

- (3) Construct a new density

$$\rho_*^{[n+1]} = \sum_{m,k} n_{m,k}^{[n+1]} |\phi_{m,k}^{[n+1]}|^2,$$

where

$$\begin{cases} n_{m,k}^{[n+1]} = 2 & \text{if } \epsilon_{m,k}^{[n+1]} < \epsilon_F^{[n+1]}, \\ 0 \leq n_{m,k}^{[n+1]} \leq 2 & \text{if } \epsilon_{m,k}^{[n+1]} = \epsilon_F^{[n+1]}, \\ n_{m,k}^{[n+1]} = 0 & \text{if } \epsilon_{m,k}^{[n+1]} > \epsilon_F^{[n+1]}, \end{cases} \quad \text{and} \quad \sum_{(m,k)} n_{m,k}^{[n+1]} = N.$$

- (4) Update the density:

$$\rho^{[n+1]} = t_n \rho_*^{[n+1]} + (1 - t_n) \rho^{[n]},$$

where  $t_n \in [0, 1]$  either is a fixed parameter independent of  $n$  and chosen a priori, or is optimized using the optimal damping algorithm (see below).

- (5) If some convergence criterion is satisfied, then stop; else, replace  $n$  with  $n + 1$  and go to step (1).

In the nondegenerate case, that is, when  $\epsilon_F^{[n+1]}$  is not an eigenvalue of the Hamiltonian  $H^{[n]}$ , the occupation numbers  $n_{m,k}^{[n+1]}$  are equal to either 0 (unoccupied) or 2 (fully occupied), while in the degenerate case the occupation numbers at the Fermi level have to be determined. We distinguish two cases: if  $W = 0$ , or more generally if  $W$  is spherically symmetric, and if  $\epsilon_F^{[n+1]}$  is not an accidentally degenerate eigenvalue of  $H^{[n]}$ , then the occupation numbers at the Fermi level are all equal; otherwise, the occupation numbers are not known a priori. In our approach we select the occupation numbers at the Fermi level which provide the lowest Kohn–Sham energy. When the degenerate eigenspace at the Fermi level is of dimension 3, that is, when the highest-energy partially occupied orbitals are perturbations of a three-fold degenerate p-orbital, the optimal occupation numbers can be found by using the golden search or bisection method [30, Chapter 10] since, in this case, the search space can be parametrized by a single real-valued parameter (this is due to the fact that the sum of the three occupation numbers is fixed and that two of them are equal by cylindrical symmetry). In the general case, more generic optimization methods have to be resorted to.

In the discretization framework we have chosen, the algorithm can be formulated as follows.

#### *Initialization.*

- (1) Choose the numerical parameters  $m_h$  (cut-off in the spherical harmonics expansion),  $L_e$  (size of the simulation domain for the radial components of the Kohn–Sham orbitals and the electrostatic potential),  $N_l$  (size of the mesh for solving the radial equations),  $N_{g,r}$  (number of Gauss points for the radial quadrature formula),  $N_{g,\theta}$  (number of Gauss points for the angular quadrature formula), and  $\varepsilon > 0$  (convergence threshold).
- (2) Assemble the matrices  $A = A^a$ ,  $M_{-2} = M_{-2}^a$ ,  $M_{-1}$ ,  $M_0$ ,  $M_1$ ,  $C^{l,m}$ , and  $V_\mu$  and the vector  $G$ . The tensor  $F$  can be either computed once and for all, or the contractions  $F : R_l^{[n]}$  can be computed on the fly, depending on the size of the discretization parameters and the computational means available.
- (3) Choose an initial guess  $(R_l^{[0]})_{0 \leq l \leq 2m_h}$  for the matrices representing the discretized ground state density at iteration 0 (it is possible to take  $R_l = 0$  for all  $l$  if no other better guess is known).



*Iterations.* Assume the matrices  $(R_l^{[n]})_{0 \leq l \leq 2m_h}$  at iteration  $n$  are known.

(1) Construct the building blocks of the discretized analogues of the operators  $H_m^{[n]}$ .

For this purpose,

(a) solve, for each  $l = 0, \dots, 2m_h$ , the linear equation

$$(A^a + l(l+1)M_{-2}^a)Q_l^{[n]} = 4\pi(F : R_l^{[n]} - N\delta_{l0}G)$$

and

(b) assemble, for each  $l = 0, \dots, 2m_h$ , the matrix  $V_l^{\text{xc},[n]}$  by Gauss quadrature rules

$$[V_{\text{xc}}^{l,[n]}]_{ij} = \sqrt{\pi} \sum_{k=1}^{N_l} \sum_{p=1}^{N_{\text{g},r}} \sum_{q=1}^{N_{\text{g},\theta}} h_k w_{p,r} w_{q,\theta} f_{ij}^l(r_k + h_k t_{p,r}, t_{q,\theta}),$$

where

$$f_{ij}^l(r, t_\theta) = v_{\text{xc}} \left( \frac{1}{\sqrt{4\pi}} \sum_{l=0}^{m_h} \sum_{i,j=1}^{N_h} [R_l]_{i,j} \frac{\mathcal{X}_i(r)\mathcal{X}_j(r)}{r^2} Y_l^0(\arccos t_\theta) \right) \times \mathcal{X}_i(r)\mathcal{X}_j(r) Y_l^0(\arccos t_\theta).$$

(2) Solve, for each  $0 \leq m \leq m_h$ , the generalized eigenvalue problem

$$\begin{aligned} \frac{1}{2}AU^{m,k,[n+1]} + \frac{1}{2}M_{-2}U^{m,k,[n+1]}D_m - ZM_{-1}U^{m,k,[n+1]} + NV_\mu U^{m,k,[n+1]} \\ + \sum_{l=0}^{2m_h} (Q_l^{[n]T} \cdot F)(U^{m,k,[n+1]}C^{l,m}) + \sum_{l=0}^{2m_h} V_{\text{xc}}^{l,[n]}U^{m,k,[n+1]}C^{l,m} \\ - \frac{\beta}{\sqrt{3}}M_1U^{m,k,[n+1]}C^{1,m} = \epsilon_{m,k}^{[n+1]}M_0U^{m,k,[n+1]}, \end{aligned} \quad (55)$$

$$\text{Tr}([U^{m,k,[n+1]}]^T M_0 U^{m,k',[n+1]}) = \delta_{kk'}. \quad (56)$$

(3) Build the matrices  $R_{l,*}^{[n+1]}$  using the Aufbau principle, and, if necessary, optimizing the occupation numbers  $n_{m,k}^{[n+1]}$ , by selecting the occupation numbers at the Fermi level leading to the lowest Kohn–Sham energy:<sup>3</sup>

$$R_{l,*}^{[n+1]} = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} n_{m,k}^{[n+1]} U^{m,k,[n+1]} C^{l,m} [U^{m,k,[n+1]}]^T,$$

<sup>3</sup>In practice, this optimization problem is low-dimensional. Indeed, the degeneracy of the Fermi level is typically 3 (perturbation of p-orbitals) or 5 (perturbation of d-orbitals) for most atoms of the first four rows of the periodic table, and some of the occupation numbers are known to be equal for symmetric reasons.

where

$$\begin{cases} n_{m,k}^{[n+1]} = 2 & \text{if } \epsilon_{m,k}^{[n+1]} < \epsilon_{\mathbb{F}}^{[n+1]}, \\ 0 \leq n_{m,k}^{[n+1]} \leq 2 & \text{if } \epsilon_{m,k}^{[n+1]} = \epsilon_{\mathbb{F}}^{[n+1]}, \\ n_{m,k}^{[n+1]} = 0 & \text{if } \epsilon_{m,k}^{[n+1]} > \epsilon_{\mathbb{F}}^{[n+1]}, \end{cases} \quad \text{and} \quad \sum_{(m,k)} n_{m,k}^{[n+1]} = N.$$

(4) Update the density:

$$\text{for all } 0 \leq l \leq 2m_h, \quad R_l^{[n+1]} = t_n R_{l,*}^{[n+1]} + (1 - t_n) R_l^{[n]},$$

where  $t_n \in [0, 1]$  either is a fixed parameter independent of  $n$  and chosen a priori, or is optimized using the optimal damping algorithm (see below).

(5) If (for instance)  $\max_{0 \leq l \leq 2m_h} \|R_l^{[n+1]} - R_l^{[n]}\| \leq \epsilon$  or  $|E^{[n+1]} - E^{[n]}| \leq \epsilon$ , then stop; else go to step (1).

Note that the generalized eigenvalue problem (55)–(56) can be rewritten as a standard generalized eigenvalue problem of the form

$$\mathbb{H}^m \mathbb{V}_k = \epsilon_{m,k}^{[n+1]} \mathbb{M} \mathbb{V}_k, \quad \mathbb{V}_k^T \mathbb{M} \mathbb{V}_{k'} = \delta_{kk'}, \quad (57)$$

where the unknowns are vectors (and not matrices) by introducing the column vectors  $\mathbb{V}_k \in \mathbb{R}^{(m_h+1-|m|)N_h}$  and the block matrices

$$\mathbb{H}^m \in \mathbb{R}^{(m_h+1-|m|)N_h \times (m_h+1-|m|)N_h} \quad \text{and} \quad \mathbb{M} \in \mathbb{R}^{(m_h+1-|m|)N_h \times (m_h+1-|m|)N_h}$$

defined as

$$\mathbb{V}_k = \begin{pmatrix} U_{\cdot, |m|}^{m,k,[n+1]} \\ \vdots \\ U_{\cdot, m_h}^{m,k,[n+1]} \end{pmatrix},$$

$$\mathbb{H}^m = \begin{pmatrix} \mathbb{H}_{|m|, |m|}^m & \mathbb{H}_{|m|, |m|+1}^m & \cdots & \mathbb{H}_{|m|, m_h-1}^m & \mathbb{H}_{|m|, m_h}^m \\ \mathbb{H}_{|m|+1, |m|}^m & \mathbb{H}_{|m|+1, |m|+1}^m & \cdots & \mathbb{H}_{|m|+1, m_h-1}^m & \mathbb{H}_{|m|+1, m_h}^m \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbb{H}_{m_h-1, |m|}^m & \mathbb{H}_{m_h-1, |m|+1}^m & \cdots & \mathbb{H}_{m_h-1, m_h-1}^m & \mathbb{H}_{m_h-1, m_h}^m \\ \mathbb{H}_{m_h, |m|}^m & \mathbb{H}_{m_h, |m|+1}^m & \cdots & \mathbb{H}_{m_h, m_h-1}^m & \mathbb{H}_{m_h, m_h}^m \end{pmatrix},$$

and

$$\mathbb{M} = \text{block diag}(M_0, \dots, M_0),$$

where each of the  $(m_h - |m| + 1)$  block  $\mathbb{H}_{l,l'}^m$  is of size  $N_h \times N_h$  with

for all  $|m| \leq l \leq m_h$ ,

$$\mathbb{H}_{l,l}^m = \frac{1}{2}A + \frac{l(l+1)}{2}M_{-2} - ZM_{-1} + NV_\mu + \sum_{l''=0}^{2m_h} C_{l,l''}^{l,m} ([Q_{l''}^{[n]}]^T \cdot F + V_{xc}^{l'',[n]}),$$

for all  $|m| \leq l \neq l' \leq m_h$ ,

$$\mathbb{H}_{l,l'}^m = \sum_{l''=0}^{2m_h} C_{l',l''}^{l,m} ([Q_{l''}^{[n]}]^T \cdot F + V_{xc}^{l'',[n]}) - \frac{\beta}{\sqrt{3}} C^{1,m} M_1 \delta_{|l-l'|,1}.$$

If  $\beta = 0$  and if the density  $\rho_h^{[n]}$  is radial, then  $R_l^{[n]} = 0$  for all  $l \in \mathbb{N}^*$ , and the matrix  $\mathbb{H}^m$  is block diagonal. The generalized eigenvalue problem (57) can then be decoupled in  $(m_h - |m| + 1)$  independent generalized eigenvalue problems of size  $N_h$ . This comes from the fact that, the problem being spherically symmetric, the Kohn–Sham Hamiltonian is block diagonal in the two decompositions

$$L^2(\mathbb{R}^3) = \bigoplus_{l \in \mathbb{N}} \mathcal{H}_l \quad \text{and} \quad L^2(\mathbb{R}^3) = \bigoplus_{m \in \mathbb{Z}} \mathcal{H}^m.$$

Let us conclude this section with some remarks on the optimal damping algorithm (ODA) [5; 6], used to find an optimal step length  $t_n$  to mix the matrices  $R_{l,*}^{[n+1]}$  and  $R_l^{[n]}$  in step (4) of the iterative algorithm. This step length is obtained by minimizing on the range  $t \in [0, 1]$  the one-dimensional function

$$t \mapsto \tilde{E}_{Z,N}^{\text{rHF/LDA}}((1-t)\gamma_*^{[n+1]} + t\gamma^{[n]}, \beta W),$$

where  $\gamma^{[n]}$  is the current approximation of the ground state density matrix at iteration  $n$  and

$$\gamma_*^{[n+1]} = \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1)N_h}} n_{m,k}^{[n+1]} |\Phi_{m,k,h}^{[n+1]}\rangle \langle \Phi_{m,k,h}^{[n+1]}|,$$

with

$$\Phi_{m,k,h}^{[n+1]}(r, \theta, \varphi) = \sum_{l=|m|}^{m_h} \sum_{i=1}^{N_h} U_{i,l}^{m,k,[n+1]} \frac{\mathcal{X}_i(r)}{r} Y_l^m(\theta, \varphi).$$

A key observation is that this optimization problem can be solved without storing density matrices, but only the two sets of matrices  $R^{[n]} := (R_l^{[n]})_{0 \leq l \leq 2m_h}$  and  $R_*^{[n+1]} := (R_{l,*}^{[n+1]})_{0 \leq l \leq 2m_h}$ , and the scalars

$$E_{\text{kin}}^{[n]} := \text{Tr}(-\frac{1}{2} \Delta \gamma^{[n]})$$

and

$$\begin{aligned} E_{\text{kin},*}^{[n+1]} &:= \text{Tr}\left(-\frac{1}{2}\Delta\gamma_*^{[n+1]}\right) \\ &= \frac{1}{2} \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} n_{m,k}^{[n+1]} \left( \text{Tr}([U^{m,k,[n+1]}]^T A U^{m,k,[n+1]}) \right. \\ &\quad \left. + \text{Tr}(D_m[U^{m,k,[n+1]}]^T M_{-2} U^{m,k,[n+1]}) \right). \end{aligned}$$

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

$$\begin{aligned} \tilde{E}_{Z,N}^{\text{rHF/LDA}}((1-t)\gamma_*^{[n+1]} + t\gamma_*^{[n]}, \beta W) \\ = (1-t)E_{\text{kin},*}^{[n+1]} + tE_{\text{kin}}^{[n]} + \mathcal{F}^{\text{rHF/LDA}}((1-t)R_*^{[n+1]} + tR_*^{[n]}, \beta W), \end{aligned}$$

where the functional  $\mathcal{F}^{\text{rHF/LDA}}$  collects all the terms of the Kohn–Sham functional depending on the density only. When  $E_{\text{xc}} = 0$  (rHF model), the function

$$t \mapsto \tilde{E}_{Z,N}^{\text{rHF/LDA}}((1-t)\gamma_*^{[n+1]} + t\gamma_*^{[n]}, \beta W)$$

is a convex polynomial of degree two, and its minimizer on  $[0, 1]$  can therefore be easily computed explicitly. In the LDA case, the minimum on  $[0, 1]$  of the above function of  $t$  can be obtained using any line search method. We use here the golden search method. Once the minimizer  $t_n$  is found, the quantity  $E_{\text{kin}}^{[n]}$  is updated using the relation

$$E_{\text{kin}}^{[n+1]} = (1-t_n)E_{\text{kin},*}^{[n+1]} + t_n E_{\text{kin}}^n.$$

The source code of a Fortran 95 implementation of the method is available on GitHub [27].

#### 4. Numerical results

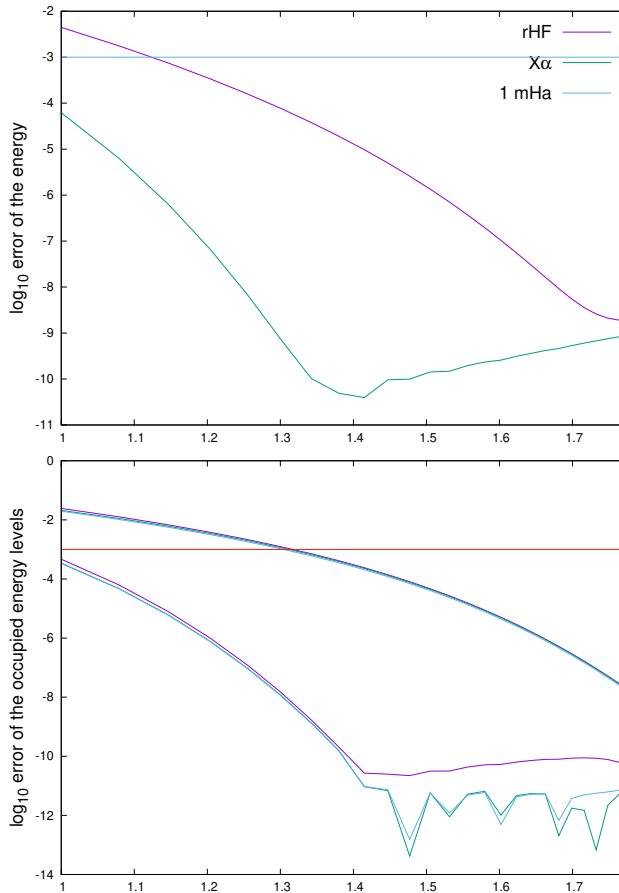
As previously mentioned, we use in our code, written in Fortran 95 and linked to the BLAS, LAPACK, and ARPACK libraries, the same basis to discretize the radial components of the Kohn–Sham orbitals and of the Hartree potential, that is,  $(\mathcal{X}_i)_{1 \leq i \leq N_h} = (\zeta_i)_{1 \leq i \leq N_h}$ , and the  $\mathbb{P}_4$  finite elements method to construct the discretization basis.

In order to test our methodology on LDA-type models, we have chosen to work with the  $X\alpha$  model [34], which has a simple analytic expression

$$E_{\text{xc}}(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int_{\mathbb{R}^3} \rho^{4/3} \quad \text{and} \quad v_{\text{xc}}(\rho) = -\left(\frac{3}{\pi}\right)^{1/3} \rho^{1/3}.$$

The exchange–correlation contributions must be computed by numerical quadratures. We use here the Gauss quadrature method with  $N_{g,r} = 15$  and  $N_{g,\theta} = 30$  (see Section 3.1.4).

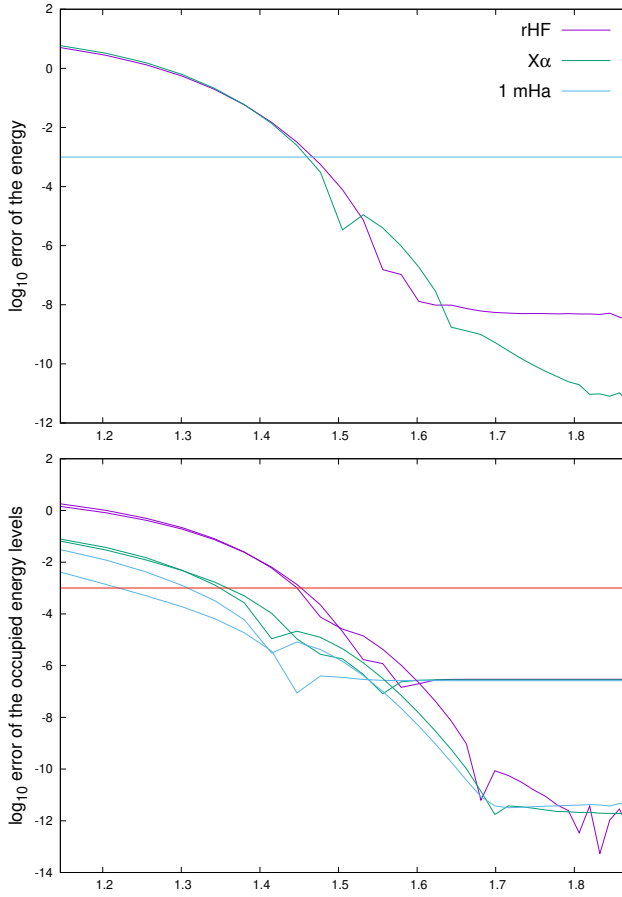
We start this section by studying the convergence rate of the ground state energy and of the occupied energy levels of the carbon atom ( $Z = 6$ ) as functions of the



**Figure 1.** Log-log plot of the error on the total energy (top) and the three occupied energy levels (bottom) of the carbon atom for the rHF (solid lines) and  $X\alpha$  (dashed lines) models as a function of the cut-off radius  $L_e$  for a fixed mesh size  $N_I = 50$  (the reference calculation corresponds to  $L_e = 100$  and  $N_I = 100$ ).

cut-off radius  $L_e$  and the mesh size  $N_I$  (see Section 3.1.4). The errors on the total energy and on the occupied energy levels for the rHF and  $X\alpha$  models are plotted in Figure 1 (for  $L_e = 50$  and different values of  $N_I$ ) and Figure 2 (for  $N_I = 50$  and different values of  $L_e$ ), the reference calculation corresponding to  $L_e = 100$  and  $N_I = 100$ . We can see that the choice  $L_e = 50$  and  $N_I = 50$  provides accuracies of about  $1 \mu\text{Ha}$  (recall that chemical accuracy corresponds to  $1 \text{ mHa}$ ).

Note that the convergence of the ground state energy and occupied energy levels with respect to the cut-off radius  $L_e$  is much faster for  $X\alpha$  than for rHF. This is due to the fact that the energies of the highest occupied orbitals are closer to zero for the rHF model, leading to a slower asymptotic decay at infinity of the ground state



**Figure 2.** Log-log plot of the error on the total energy (top) and the three occupied energy levels (bottom) of the carbon atom for the rHF (solid lines) and  $X\alpha$  (dashed lines) models as a function of the mesh size  $N_I$ , for a fixed cut-off radius  $L_e = 50$  (the reference calculation corresponds to  $L_e = 100$  and  $N_I = 100$ ).

density. In contrast, the convergence rates with respect to the mesh size are almost the same for the two models.

**4.1. Electronic structures of isolated atoms.** We report here calculations on all the atoms of the first four rows of the periodic table obtained with the rHF (Section 4.1.1) and  $X\alpha$  (Section 4.1.2) models.

**4.1.1. Occupied energy levels in the rHF model.** The numerical results presented in this section indicate that, for neutral atoms, the assumption

$$\epsilon_{Z,Z,F}^{0,\text{rHF}} \text{ is negative and is not an accidentally degenerate eigenvalue of } H_{Z,Z}^{0,\text{rHF}},$$

which guarantees the uniqueness of the nonmagnetic rHF ground state density matrix (Theorem 1), is satisfied for all the chemical elements of the first two rows of the periodic table, and for most of the elements of the third and four rows. Surprisingly, we observe accidental degeneracies at the Fermi level for Sc and Ti (4p and 3d shells), for V, Cr, Mn, and Fe (5s and 3d shells), for Zr (5p and 4d shells), for Nb and Mo (6s and 4d shells), and for Pd and Ag (5s and 4d shells). For some of these elements, the Fermi level is clearly negative, and we can conclude that (see Appendix C)

- if the Fermi level contains an s and a d shell, then the nonmagnetic rHF ground state is unique and
- if the Fermi level contains a p and a d shell, and if both shells are partially occupied (which is suggested by our numerical simulations), then the nonmagnetic rHF ground state is not unique.

For other chemical elements, such as iron ( $Z = 26$ ), the Fermi level is so close to zero that the numerical accuracy of our numerical method does not allow us to know whether it is slightly negative or equal to zero.

The negative eigenvalues of  $H_{\rho^0}^{\text{rHF}}$  for all  $1 \leq Z \leq 54$  (first four rows of the periodic table) are listed in Appendix A. The results for  $1 \leq Z \leq 20$ ,  $27 \leq Z \leq 39$ ,  $43 \leq Z \leq 45$ , and  $48 \leq Z \leq 54$  correspond to  $N_l$  increasing from 35 to 75 as  $Z$  increases and  $L_e$  increasing from 30 to 100 as  $|\epsilon_{Z,Z,F}^{0,\text{rHF}}|$  decreases, which were sufficient to obtain an accuracy of  $1 \mu\text{Ha}$ . The remaining atoms are more difficult to deal with because the Fermi level seems to be an accidentally degenerate eigenvalue of  $H_{\rho^0}^{\text{rHF}}$  associated with

- the 4p and 3d shells for  $Z = 21$  and  $Z = 22$ ,
- the 5s and 3d shells for  $23 \leq Z \leq 26$ , with a Fermi level very close (or possibly equal) to zero,
- the 5p and 4d shells for  $Z = 40$ , with a Fermi level very close (or possibly equal) to zero,
- the 6s and 4d shells for  $Z = 41$  and  $Z = 42$ , with a Fermi level very close (or possibly equal) to zero, and
- the 5s and 4d shells for  $Z = 46$  and  $Z = 47$ .

Since the radial component of the highest occupied orbital typically vanishes as  $\exp(-\sqrt{2|\epsilon_{Z,Z,F}^{0,\text{rHF}}|}r)$  if  $\epsilon_{Z,N,F}^{0,\text{rHF}} < 0$  and algebraically if  $\epsilon_{Z,Z,F}^{0,\text{rHF}} = 0$ , a very large value of  $L_e$  is needed for the atoms for which the Fermi level is very close or possibly equal to zero. For that case, we use a nonuniform grid with  $N_l' = 80$  and  $L_e' = 100$  as explained in Section 3.1.4 and glue to it a uniform one with 10 points and length  $L_e - L_e'$  varying from 70 to 700. Lastly, we add to the basis a function with an unbounded support, equal to  $L_e/r$  on  $[L_e, +\infty)$  (see [9] for details). This was sufficient to obtain an accuracy of  $10 \mu\text{Ha}$ .

When the accidental degeneracy involves an s shell and since the density is radial, the problem of finding the occupation numbers at the Fermi level reduces to finding a single parameter  $t_0 \in [0, 1]$ , which encodes the amount of electrons on the upper s shell. In other words, one can write

$$\rho_{Z,Z}^{0,\text{rHF}} = \rho_f + t_0 \rho_s + (1 - t_0) \rho_d,$$

where  $\rho_f$  is the density corresponding to the fully occupied shells, and where  $\rho_s$  and  $\rho_d$  are densities corresponding to the accidentally degenerate s and d shells. Using the same trick for accidentally degenerate p and d shells, we manage to obtain a self-consistent solution to the rHF equations, which is necessarily a ground state since the rHF model is convex in the density matrix.

In the following tables, we report the rHF occupied energy levels (in Ha) of all the atoms of the first four rows of the periodic table, for which the Fermi level seems to be an accidentally degenerate eigenvalue:

- the 4p and 3d orbitals have the same energy for  $Z = 21, 22$ ,
- the 5s and 3d orbitals have the same energy for  $23 \leq Z \leq 26$ ,
- the 5p and 4d orbitals have the same energy for  $Z = 40$ ,
- the 6s and 4d orbitals have the same energy for  $Z = 41, 42$ , and
- the 5s and 4d orbitals have the same energy for  $Z = 46, 47$ .

In all these cases, the occupation number  $0 \leq n \leq 2$  of the partially occupied d orbitals is also given.

*Third row.*

Atom	Sc	Ti	V	Cr	Mn	Fe
$Z$	21	22	23	24	25	26
1s	-154.35864	-171.13186	-188.77080	-207.27457	-226.64207	-246.87446
2s	-15.78538	-17.95490	-20.24077	-22.64280	-25.15938	-27.79250
2p	-12.74151	-14.69008	-16.75392	-18.93275	-21.22503	-23.63263
3s	-1.69002	-1.91684	-2.15109	-2.39225	-2.63884	-2.89238
3p	-0.96964	-1.11529	-1.26708	-1.42402	-1.58451	-1.74975
4s	-0.08646	-0.08224	-0.07796	-0.07027	-0.06385	-0.05831
4p	-0.00262	-0.00056				
5s			-0.00044	-0.00021	-0.00008	-0.00001
3d	-0.00262	-0.00056	-0.00044	-0.00021	-0.00008	-0.00001
$n(3d)$	0.0056	0.3076	0.5662	0.7794	0.9886	1.1957



Fourth row.

Atom	Zr	Nb	Mo	Pd	Ag
$Z$	40	41	42	46	47
1s	-627.17364	-661.38533	-696.51265	-846.21733	-885.91821
2s	-83.77963	-89.25420	-94.89727	-119.19036	-125.66905
2p	-76.25111	-81.47185	-86.85999	-110.12295	-116.34159
3s	-12.93681	-14.14588	-15.39104	-20.77177	-22.19282
3p	-10.23529	-11.31538	-12.43032	-17.27895	-18.56438
3d	-5.37710	-6.20667	-7.06960	-10.89439	-11.91967
4s	-1.58204	-1.76423	-1.94236	-2.66438	-2.82492
4p	-0.89248	-1.01284	-1.13016	-1.61336	-1.71460
5s	-0.07367	-0.06267	-0.04957	-0.03846	-0.03379
5p	-0.00048				
6s		-0.00014	-0.000002		
4d	-0.00048	-0.00014	-0.000002	-0.03846	-0.03379
$n(4d)$	0.3207	0.5840	0.7983	1.6655	1.9293

The occupied energy levels in the rHF model of all the atoms of the first four rows of the periodic table are given in Appendix A.

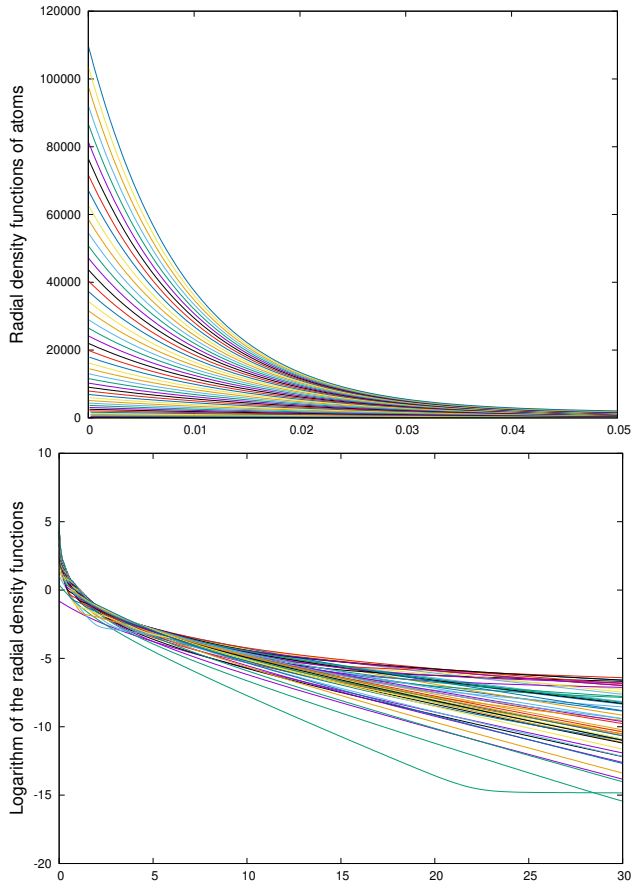
**Remark.** Our numerical simulations seem to show that for all  $1 \leq Z \leq 54$ , there are no unoccupied negative eigenvalues in the rHF ground states of neutral atoms. Thus, the negative spectrum of the rHF Hamiltonian coincides with the set of occupied energy levels.

We end this section with Figure 3, which backs up the conjecture that rHF atomic densities are decreasing radial functions of the distance to the nucleus.

**4.1.2. Occupied energy levels in the  $X\alpha$  model.** The tables below provide the negative eigenvalues of the Kohn–Sham  $X\alpha$  Hamiltonian (in Ha) for all the atoms of the first four rows of the periodic table with accidentally degenerate Fermi levels, the degeneracy occurring in all cases between an s shell and a d shell (4s–3d for  $23 \leq Z \leq 28$ , 5s–4d for  $41 \leq Z \leq 44$ ). All the results of this section are obtained for  $L_e = 30$  and  $N_I$  increasing from 30 to 75 as  $Z$  increases.

Third row.

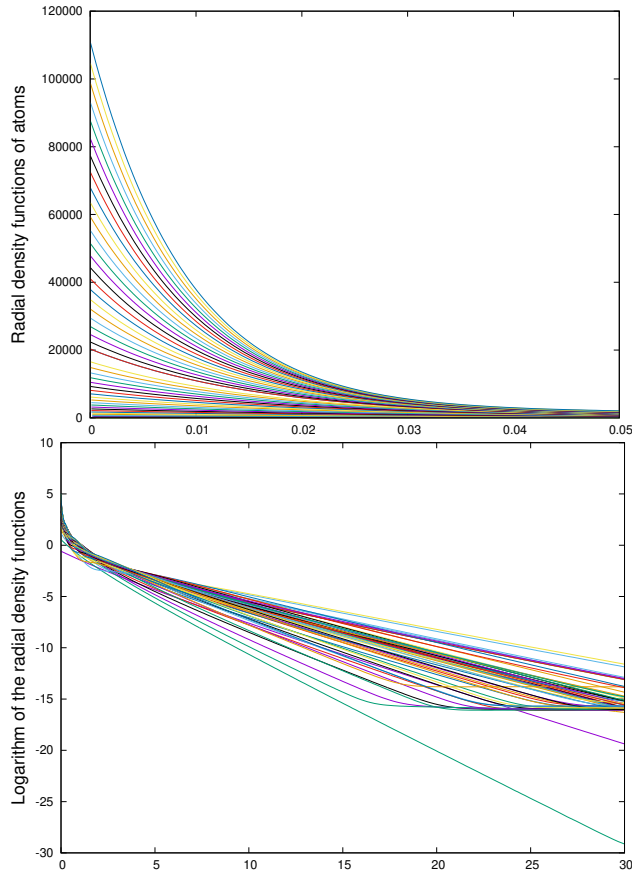
Atom	V	Cr	Mn	Fe	Co	Ni
$Z$	23	24	25	26	27	28
1s	-195.11079	-213.87746	-233.50875	-254.00470	-275.36535	-297.59075
2s	-21.72028	-24.14440	-26.68762	-29.35014	-32.13212	-35.03372
2p	-18.33888	-20.55424	-22.88690	-25.33699	-27.90468	-30.59009
3s	-2.44810	-2.68033	-2.92165	-3.17214	-3.43191	-3.70102
3p	-1.53340	-1.68342	-1.83995	-2.00304	-2.17274	-2.34907
4s	-0.13684	-0.13575	-0.13474	-0.13379	-0.13292	-0.13212
3d	-0.13684	-0.13575	-0.13474	-0.13379	-0.13292	-0.13212
$n(3d)$	0.6393	0.8873	1.1278	1.3622	1.5918	1.8174



**Figure 3.** Top: the plot of the densities of all the atoms  $1 \leq Z \leq 54$  obtained with our code as a function of the distance to the nucleus, on the interval  $[0, 0.05]$ . Bottom: the plot of the logarithms of those densities on the interval  $[0, 50]$ .

*Fourth row.*

Atom	Nb	Mo	Tc	Ru
Z	41	42	43	44
1s	-673.74149	-709.15136	-745.48044	-782.72787
2s	-92.74707	-98.44597	-104.31826	-110.36286
2p	-85.27606	-90.73190	-96.35989	-102.15896
3s	-15.40918	-16.63439	-17.90004	-19.20531
3p	-12.56830	-13.66757	-14.80629	-15.98365
3d	-7.35588	-8.21062	-9.10349	-10.03361
4s	-2.05942	-2.19877	-2.34006	-2.48279
4p	-1.27048	-1.35425	-1.43939	-1.52544
5s	-0.13172	-0.11937	-0.10617	-0.09183
4d	-0.13172	-0.11937	-0.10617	-0.09183
$n(4d)$	0.6535	0.9847	1.2956	1.5896



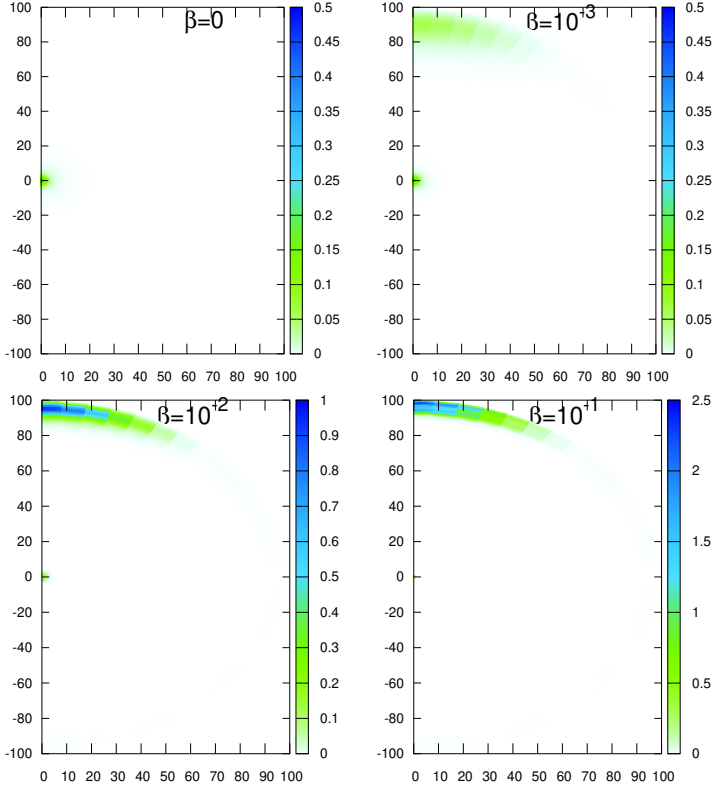
**Figure 4.** Top: the plot of the  $X\alpha$  densities of all the atoms  $1 \leq Z \leq 54$  obtained with our code as a function of the distance to the nucleus, on the interval  $[0, 0.05]$ . Bottom: the plot of the logarithms of those densities on the interval  $[0, 50]$ .

The occupied energy levels in the  $X\alpha$  model of all the atoms of the first four rows of the periodic table are given in Appendix B.

We end this section with Figure 4, which shows that as in the rHF case, the  $X\alpha$  atomic densities seem to be decreasing radial functions of the distance to the nucleus.

**4.2. Perturbation by a uniform electric field (Stark effect).** In this section, we consider atoms subjected to a uniform electric field, that is, to an external potential  $\beta W_{\text{Stark}}$  with

$$W_{\text{Stark}}(\mathbf{r}) = -e_z \cdot \mathbf{r}$$



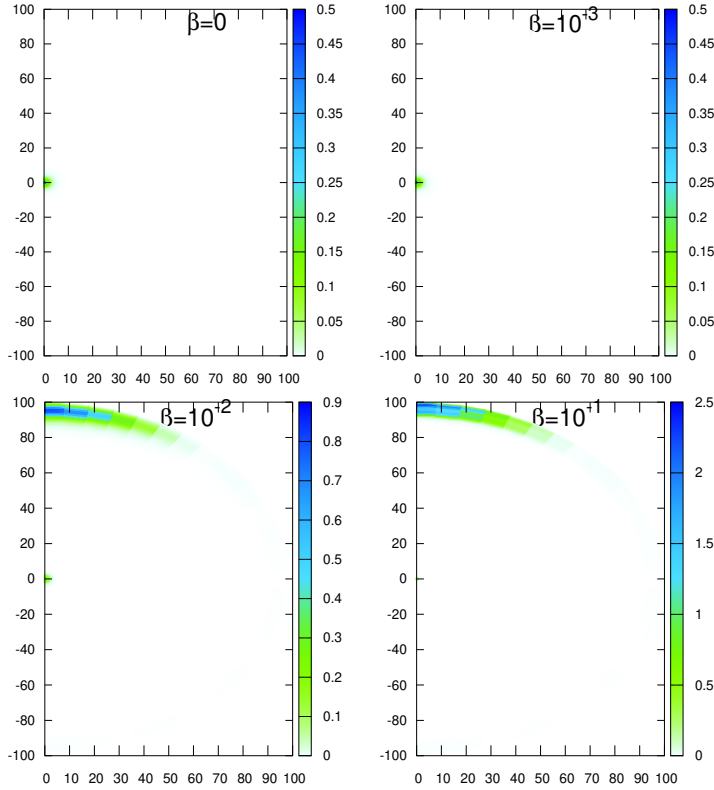
**Figure 5.** rHF case: at the top left is a plot in the  $\rho z$ -half-plane of the ground state density (multiplied by  $r^2$ ) of an isolated carbon atom. The others are plots of the ground state densities (multiplied by  $r^2$ ) of the carbon atom in a sphere of radius  $L_e = 100$ , subjected to a uniform external electric field, with coupling constants  $\beta = 10^{-3}, 10^{-2}, 10^{-1}$ .

or, in spherical coordinates,

$$W_{\text{Stark}}(r, \theta, \varphi) = -\sqrt{\frac{4\pi}{3}} r Y_1^0(\theta, \varphi).$$

As already mentioned in Section 2.2,  $\tilde{\mathcal{F}}_{Z,N}^{\text{rHF/LDA}}(\beta W_{\text{Stark}}) = -\infty$  whenever  $\beta \neq 0$ , and the corresponding variational problem has no minimizer. However, one can find a minimizer  $\gamma_h \in \mathcal{K}_{N,h}$  to the approximated problem  $\tilde{\mathcal{F}}_{Z,N,h}^{\text{rHF/LDA}}(\beta W_{\text{Stark}})$ . Hereafter we consider the carbon atom ( $Z = 6$ ). Even though the cutoff  $m_h$  is set equal to 6, all the terms corresponding to a magnetic number  $m > 1$  are in fact equal to zero.

Recall that the perturbed ground state density is cylindrically symmetric about the  $z$ -axis. Figures 5 and 6 are plots in the  $\rho z$ -half-plane ( $\rho = r \sin \theta$  and  $z = r \cos \theta$ ) of the densities  $\rho_h$  (multiplied by  $r^2$  in order to emphasize what is going on at large distances from the nucleus), for the carbon atom ( $Z = 6$ ), obtained for different values of  $\beta$ .

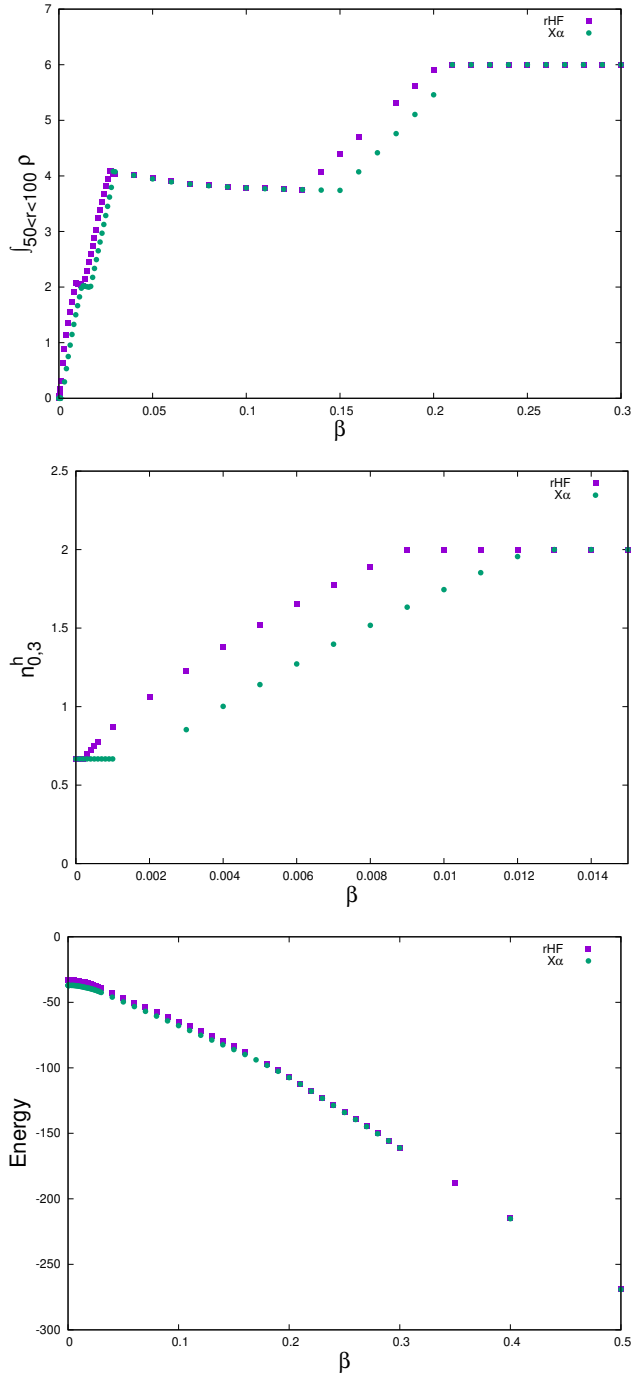


**Figure 6.**  $X\alpha$  case: at the top left is a plot in the  $qz$ -half-plane of the ground state density (multiplied by  $r^2$ ) of an isolated carbon atom. The others are plots of the ground state densities (multiplied by  $r^2$ ) of the carbon atom in a sphere of radius  $L_e = 100$ , subjected to a uniform external electric field, with coupling constants  $\beta = 10^{-3}, 10^{-2}, 10^{-1}$ .

For  $\beta = 10^{-2}$  and  $\beta = 10^{-1}$ , we clearly see spurious boundary effects: part of the electronic cloud is localized in the region where the external potential takes highly negative values. This result is obviously not physical. On the other hand, for the  $X\alpha$  model and for  $\beta = 10^{-3}$  we simply observe a polarization of the electronic cloud. The perturbation potential being not spherically symmetric, it breaks the symmetry of the density. This numerical solution can probably be interpreted as a (nonlinear) resonant state. We will come back to the analysis of this interesting case in a following work.

Figure 7 shows the amount of electrons of the carbon atom which escape to infinity as a function of the coupling constant  $\beta$  (for  $L_e = 100$  and  $N_I = 50$ ), in the rHF and  $X\alpha$  cases.

In general, the standard ODA is used to achieve convergence (see Section 3). However, for  $\beta$  small or large enough, the occupation numbers are selected as follows:  $n_{0,1}^{[n]} = n_{0,2}^{[n]} = 2$ ,  $n_{0,3} = 2(1 - t_0)$ , and  $n_{1,1}^{[n]} = 2 - n_{0,3} = 2t_0$ ,  $t_0$  being the



**Figure 7.** Top: the plot of the integral on  $B_{100} \setminus B_{50}$  of the density  $\rho_h$ . Middle: the plot of the occupation number  $n_{0,3}^h$ . Bottom: the plot of the total energy, for  $L_e = 100$  and  $N_I = 50$  as a function of  $\beta$  in the rHF and  $X\alpha$  cases.

minimizer of

$$t \mapsto \widetilde{E}_{6,6}^{\text{rHF/LDA}}((1-t)\gamma_{0,*}^{[n]} + t\gamma_{1,*}^{[n]}, \beta W),$$

where

$$\gamma_{0,*}^{[n]} = 2 \sum_{1 \leq k \leq 3} |\Phi_{0,k,h}\rangle \langle \Phi_{0,k,h}|, \quad \gamma_{1,*}^{[n]} = 2 \sum_{1 \leq k \leq 2} |\Phi_{0,k,h}\rangle \langle \Phi_{0,k,h}| + 2|\Phi_{1,1,h}\rangle \langle \Phi_{1,1,h}|.$$

This modification of ODA significantly increases the rate of convergence for  $\beta$  small or large, but does not converge for all intermediate values of  $\beta$ .

While  $\widetilde{\mathcal{J}}_{Z,N}^{\text{rHF/LDA}}(\beta W_{\text{Stark}}) = -\infty$  and the corresponding variational problem has no minimizer, the first-order perturbation  $\gamma_{Z,N,W_{\text{Stark}}}^{(1),\text{rHF}}$  of the ground state density matrix does exist (see Theorem 4). If we consider the carbon atom, it can be expressed as a function of the unperturbed occupied Kohn–Sham orbitals and of their first-order perturbations. We indeed have

$$\gamma_{6,6,W_{\text{Stark}}}^{(1),\text{rHF}} = \sum_{\substack{(m,k) \in \mathbb{C}_{6,6} \\ i_1 \geq 0, i_2 \geq 0, i_3 \geq 0 \\ i_1 + i_2 + i_3 = 1}} n_{m,k}^{(i_1)} |\Phi_{m,k}^{(i_2)}\rangle \langle \Phi_{m,k}^{(i_3)}|,$$

where  $\mathbb{C}_{6,6} = \{(0, 1), (0, 2), (0, 3), (1, 1)\}$ , where  $\epsilon_{m,k}^{(0)}$  is the  $k$ -th lowest eigenvalue of  $H_{6,6}^{0,\text{rHF}}$  in the subspace  $\mathcal{H}^m$ ,  $\Phi_{m,k}^{(0)}$  is an associated normalized eigenfunction, and

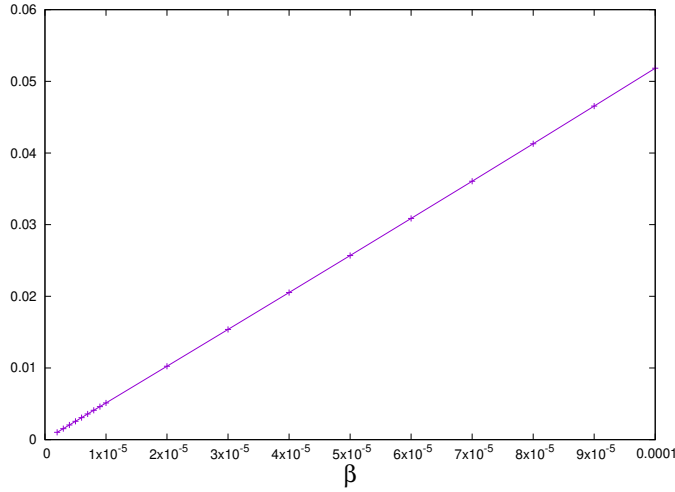
$$n_{0,1}^{(0)} = n_{0,2}^{(0)} = 2, \quad n_{0,3}^{(0)} = \frac{2}{3}, \quad \text{and} \quad n_{1,1}^{(0)} = \frac{4}{3},$$

while  $\epsilon_{m,k}^{(1)}$ ,  $\Phi_{m,k}^{(1)}$ , and  $n_{m,k}^{(1)}$  satisfy the self-consistent equation

$$\begin{aligned} (H_{6,6}^{0,\text{rHF}} - \epsilon_{m,k}^{(0)})\Phi_{m,k}^{(1)} + (\rho^{(1)} \star |\cdot|^{-1})\Phi_{m,k}^{(0)} + W_{\text{Stark}}\Phi_{m,k}^{(0)} &= \epsilon_{m,k}^{(1)}\Phi_{m,k}^{(0)}, \\ \rho^{(1)} &= \sum_{(m,k) \in \mathbb{C}_{6,6}} 2n_{m,k}^{(0)}\Phi_{m,k}^{(0)}\Phi_{m,k}^{(1)} + n_{m,k}^{(1)}\Phi_{m,k}^{(0)}\Phi_{m,k}^{(0)}, \\ \int_{\mathbb{R}^3} \Phi_{m,k}^{(1)}\Phi_{m,k}^{(0)} &= 0, \quad \text{and} \quad \sum_{(m,k) \in \mathbb{C}_{6,6}} n_{m,k}^{(1)} = 0. \end{aligned}$$

We denote by  $\epsilon_{m,k,h}^{(0)}$ ,  $\epsilon_{m,k,h}^{(1)}$ ,  $\Phi_{m,k,h}^{(0)}$ ,  $\Phi_{m,k,h}^{(1)}$ , and  $n_{m,k,h}^{(1)}$  the approximations of  $\epsilon_{m,k}^{(0)}$ ,  $\epsilon_{m,k}^{(1)}$ ,  $\Phi_{m,k}^{(0)}$ ,  $\Phi_{m,k}^{(1)}$ , and  $n_{m,k}^{(1)}$ , respectively. For each  $(m, k) \in \mathbb{C}_{6,6}$ , define

$$\begin{aligned} \widetilde{\epsilon}_{m,k,h}^{(1)}(\beta) &:= \frac{1}{\beta}(\epsilon_{m,k,h}(\beta) - \epsilon_{m,k,h}^{(0)}), \\ \widetilde{\Phi}_{m,k,h}^{(1)}(\beta) &:= \frac{1}{\beta}(\Phi_{m,k,h}(\beta) - \Phi_{m,k,h}^{(0)}), \\ \widetilde{n}_{m,k,h}^{(1)}(\beta) &:= \frac{1}{\beta}(n_{m,k,h}(\beta) - n_{m,k,h}^{(0)}). \end{aligned}$$



**Figure 8.** Plot of the function  $\beta \mapsto \max_{(m,k) \in \mathbb{O}_{6,6}} \max_{l \geq |m|} \|V_{l,(m,k)}(\beta)\|_\infty$  where  $V_{l,(m,k)}(\beta)$  is the vector in the left-hand side of (58).

Recall that  $(\Phi_{m,k,h}(\beta))_{(m,k) \in \mathbb{O}_{6,6}}$  and  $n_{m,k,h}^{(1)}(\beta)$  are the eigenfunctions and eigenvalues, respectively, of the density matrix  $\gamma_h$ , the minimizer of the approximated problem  $\tilde{\mathcal{F}}_{Z,N,h}^{\text{HF}}(\beta W_{\text{Stark}})$ .

Let  $U^{m,k}$  and  $\tilde{U}^{m,k}(\beta)$  be such that

$$\begin{aligned} \Phi_{m,k,h}^{(0)}(r, \theta, \varphi) &= \sum_{l=|m|}^{m_h} \left( \sum_{i=1}^{N_h} U_{i,l}^{m,k}(\beta) \mathcal{X}_i(r)/r \right) Y_l^m(\theta, \varphi), \\ \tilde{\Phi}_{m,k,h}^{(1)}(\beta)(r, \theta, \varphi) &= \sum_{l=|m|}^{m_h} \left( \sum_{i=1}^{N_h} \tilde{U}_{i,l}^{m,k}(\beta) \mathcal{X}_i(r)/r \right) Y_l^m(\theta, \varphi). \end{aligned}$$

To show that  $\tilde{\Phi}_{m,k,h}^{(1)}(\beta) \rightarrow \Phi_{m,k,h}^{(1)}$  when  $\beta \rightarrow 0$ , it is enough to show that for each  $l \geq 0$

$$\begin{aligned} &\left( \frac{1}{2}A + \frac{l(l+1)}{2}M_{-2} - ZM_{-1} + NV_\mu - \epsilon^{(0)}M_0 \right) \tilde{U}_{\cdot,l}(\beta) - \frac{1}{\sqrt{3}}C^{1,m}M_1U_{\cdot,l-1} \\ &- \frac{1}{\sqrt{3}}C^{1,m}M_1U_{\cdot,l+1} + \sum_{l'=|m|}^{m_h} \sum_{l''=0}^{2m_h} C_{l',l''}^{l,m} ([Q_{l''}]^T \cdot F) \tilde{U}_{\cdot,l'}(\beta) \\ &+ C_{l',l''}^{l,m} ([\tilde{Q}_{l''}(\beta)]^T \cdot F) U_{\cdot,l'} - \epsilon^{(1)}M_0U_{\cdot,l} \xrightarrow{\beta \rightarrow 0} 0. \quad (58) \end{aligned}$$

The index  $(m, k)$  is omitted for simplicity, and the vector  $\tilde{Q}_l(\beta)$  is the solution to the linear system

$$(A^a + l(l+1)M_{-2}^a) \tilde{Q}_l = 4\pi F : \tilde{R}_l,$$



with

$$\tilde{R}_l := \sum_{\substack{-m_h \leq m \leq m_h \\ 1 \leq k \leq (m_h - |m| + 1) \times N_h}} 2n_{m,k}^{(0)} \tilde{U}^{m,k} C^{l,m} [\tilde{U}^{m,k}]^T + n_{m,k}^{(1)} \tilde{U}^{m,k} C^{l,m} [U^{m,k}]^T.$$

Our numerical results show that, as expected by symmetry,  $n_{m,k,h}^{(1)} = \epsilon_{m,k,h}^{(1)} = 0$  for all  $(m, k) \in \mathbb{O}_{6,6}$ , and that the left-hand side of (58) converges to zero linearly in  $\beta$  (see Figure 8).

### Appendix A: Occupied energy levels in the rHF model

In the following tables, we report the rHF occupied energy levels (in Ha) of all the atoms of the first four rows of the periodic table. In case the Fermi level seems to be an accidentally degenerate eigenvalue, the occupation number  $0 \leq n \leq 2$  of the partially occupied d orbitals is also given.

*Hydrogen and helium.*

Atom	Z	1s
H	1	-0.046222
He	2	-0.184889

*First row.*

Atom	Z	1s	2s	2p
Li	3	-1.202701	-0.013221	
Be	4	-2.902437	-0.043722	
B	5	-5.407212	-0.164961	-0.002389
C	6	-8.555732	-0.265682	-0.012046
N	7	-12.390177	-0.384699	-0.027312
O	8	-16.912538	-0.522883	-0.047280
F	9	-22.123525	-0.680479	-0.071663
Ne	10	-28.023481	-0.857597	-0.100342

*Second row.*

Atom	Z	1s	2s	2p	3s	3p
Na	11	-35.065314	-1.453872	-0.514340	-0.012474	
Mg	12	-42.963178	-2.169348	-1.037891	-0.034036	
Al	13	-51.833760	-3.118983	-1.789953	-0.135543	-0.002486
Si	14	-61.532179	-4.160128	-2.629056	-0.208803	-0.010768
P	15	-72.083951	-5.319528	-3.582422	-0.284199	-0.023431
S	16	-83.489746	-6.598489	-4.651551	-0.363585	-0.039746
Cl	17	-95.749535	-7.997404	-5.836930	-0.447628	-0.059401
Ar	18	-108.863191	-9.516434	-7.138772	-0.536669	-0.082233

*Third row.*

Atom	K	Ca	Sc	Ti
Z	19	20	21	22
1s	-123.093717	-138.233855	-154.35864	-171.13186
2s	-11.413369	-13.478564	-15.78538	-17.95490
2p	-8.815789	-10.658837	-12.74151	-14.69008
3s	-0.866180	-1.225936	-1.69002	-1.91684
3p	-0.326113	-0.596554	-0.96964	-1.11529
4s	-0.009500	-0.024275	-0.08646	-0.08224
4p			-0.00262	-0.00056
3d			-0.00262	-0.00056
$n(3d)$			0.0056	0.3076

Atom	V	Cr	Mn	Fe
Z	23	24	25	26
1s	-188.77080	-207.27457	-226.64207	-246.87446
2s	-20.24077	-22.64280	-25.15938	-27.79250
2p	-16.75392	-18.93275	-21.22503	-23.63263
3s	-2.15109	-2.39225	-2.63884	-2.89238
3p	-1.26708	-1.42402	-1.58451	-1.74975
4s	-0.07796	-0.07027	-0.06385	-0.05831
5s	-0.00044	-0.00021	-0.00008	-0.00001
3d	-0.00044	-0.00021	-0.00008	-0.00001
$n(3d)$	0.5662	0.7794	0.9886	1.1957

Atom	Co	Ni	Cu	Zn
Z	27	28	29	30
1s	-267.97363	-289.94364	-312.78019	-336.48301
2s	-30.54468	-33.42047	-36.41574	-39.53045
2p	-26.15798	-28.80557	-31.57124	-34.45491
3s	-3.15502	-3.43107	-3.71624	-4.01038
3p	-1.92172	-2.10456	-2.29392	-2.48957
4s	-0.05438	-0.05459	-0.05539	-0.05646
3d	-0.00121	-0.00722	-0.01370	-0.02026

Atom	Ga	Ge	As	Se	Br	Kr
Z	31	32	33	33	35	36
1s	-361.309461	-387.039855	-413.704397	-441.297733	-469.815876	-499.256211
2s	-43.037010	-46.711685	-50.583856	-54.647174	-58.896767	-63.329305
2p	-37.727020	-41.164308	-44.796323	-48.616891	-52.621294	-56.806329
3s	-4.576035	-5.182760	-5.856750	-6.590128	-7.377307	-8.214637
3p	-2.951273	-3.449483	-4.011096	-4.628856	-5.297678	-6.014298
3d	-0.264266	-0.533749	-0.860725	-1.240224	-1.668313	-2.142323
4s	-0.165288	-0.229337	-0.293291	-0.358794	-0.426192	-0.495638
4p	-0.002386	-0.010542	-0.022574	-0.037413	-0.054625	-0.073991

Fourth row.

Atom	Rb	Sr	Y	Zr	Nb	Mo
Z	37	38	39	40	41	42
1s	-529.827018	-561.340511	-593.866153	-627.17364	-661.38533	-696.51265
2s	-68.150675	-73.171957	-78.461974	-83.77963	-89.25420	-94.89727
2p	-61.378353	-66.148672	-71.186183	-76.25111	-81.47185	-86.85999
3s	-9.306434	-10.462839	-11.752114	-12.93681	-14.14588	-15.39104
3p	-6.983328	-8.015158	-9.178245	-10.23529	-11.31538	-12.43032
3d	-2.867015	-3.653051	-4.569112	-5.37710	-6.20667	-7.06960
4s	-0.760103	-1.032665	-1.383317	-1.58204	-1.76423	-1.94236
4p	-0.271916	-0.475893	-0.757307	-0.89248	-1.01284	-1.13016
5s	-0.008742	-0.021586	-0.076589	-0.07367	-0.06267	-0.04957
5p			-0.002707	-0.00048		
6s					-0.00014	-0.000002
4d				-0.00048	-0.00014	-0.000002
n (4d)				0.3207	0.5840	0.7983

Atom	Tc	Ru	Rh	Pd	Ag
Z	43	44	45	46	47
1s	-732.565071	-769.539351	-807.43252	-846.21733	-885.91821
2s	-100.718115	-106.713582	-112.88067	-119.19036	-125.66905
2p	-92.424856	-98.163269	-104.07224	-110.12295	-116.34159
3s	-16.681758	-18.014957	-19.3877	-20.77177	-22.19282
3p	-13.589644	-14.790336	-16.02956	-17.27895	-18.56438
3d	-7.975384	-8.920979	-9.90351	-10.89439	-11.91967
4s	-2.126544	-2.314092	-2.50245	-2.66438	-2.82492
4p	-1.254159	-1.381847	-1.51046	-1.61336	-1.71460
5s	-0.044554	-0.043203	-0.04269	-0.03846	-0.03379
4d	-0.009444	-0.024185	-0.04081	-0.03846	-0.03379
n (4d)				1.6655	1.9293

Atom	Cd	In	Sn	Sb
Z	48	49	50	51
1s	-926.623485	-968.415517	-1011.130388	-1054.799726
2s	-132.409803	-139.493172	-146.755434	-154.228103
2p	-122.820764	-129.641175	-136.639081	-143.846051
3s	-23.742846	-25.501835	-27.305190	-29.184036
3p	-19.977847	-21.599353	-23.264351	-25.004010
3d	-13.071777	-14.430695	-15.832028	-17.307019
4s	-3.073669	-3.487846	-3.900956	-4.343505
4p	-1.901671	-2.251916	-2.599067	-2.973930
4d	-0.096713	-0.310885	-0.517562	-0.749756
5s	-0.042861	-0.131665	-0.181855	-0.230820
5p		-0.002570	-0.010599	-0.021622

Atom	Te	I	Xe
$Z$	52	53	54
1s	-1099.421697	-1144.994552	-1191.517037
2s	-161.909103	-169.796463	-177.888737
2p	-151.260063	-158.879194	-166.702039
3s	-31.136080	-33.159211	-35.251891
3p	-26.816070	-28.698453	-30.649647
3d	-18.853453	-20.469283	-22.153023
4s	-4.812921	-5.307002	-5.824212
4p	-3.374212	-3.797932	-4.243727
4d	-1.006149	-1.285246	-1.585928
5s	-0.280095	-0.330100	-0.381026
5p	-0.034651	-0.049319	-0.065446

### Appendix B: Occupied energy levels in the $X\alpha$ model

The tables below provide the negative eigenvalues of the Kohn–Sham  $X\alpha$  Hamiltonian (in Ha) for all the atoms of the first four rows of the periodic table. We observe that atoms  $Z$ , with  $23 \leq Z \leq 28$  and  $41 \leq Z \leq 44$ , have accidentally degenerate Fermi levels, the degeneracy occurring in all cases between an s shell and a d shell (4s–3d for  $23 \leq Z \leq 28$  and 5s–4d for  $41 \leq Z \leq 44$ ). All the results of this section are obtained for  $L_e = 30$  and  $N_I$  increasing from 30 to 75 as  $Z$  increases.

*Hydrogen and helium.*

Atom	$Z$	1s
H	1	-0.194250
He	2	-0.516968

*First row.*

Atom	$Z$	1s	2s	2p
Li	3	-1.820596	-0.079032	-0.019804
Be	4	-3.793182	-0.170028	-0.045681
B	5	-6.502185	-0.305377	-0.100041
C	6	-9.884111	-0.457382	-0.157952
N	7	-13.946008	-0.628841	-0.221004
O	8	-18.690815	-0.820599	-0.289512
F	9	-24.120075	-1.032963	-0.363534
Ne	10	-30.234733	-1.266049	-0.443056

*Second row.*

Atom	Z	1s	2s	2p	3s	3p
Na	11	-37.647581	-2.007737	-1.006028	-0.077016	
Mg	12	-45.897000	-2.845567	-1.661300	-0.142129	
Al	13	-55.080562	-3.877978	-2.507293	-0.251340	-0.071775
Si	14	-65.107293	-5.017013	-3.456703	-0.359121	-0.117813
P	15	-75.982880	-6.269749	-4.516571	-0.470070	-0.166674
S	16	-87.709076	-7.638741	-5.689399	-0.585627	-0.218875
Cl	17	-100.286615	-9.125221	-6.976378	-0.706438	-0.274567
Ar	18	-113.715864	-10.729883	-8.378170	-0.832845	-0.333798

*Third row.*

Atom	K	Ca	Sc	Ti
Z	19	20	21	22
1s	-128.330888	-143.848557	-160.10133	-177.19446
2s	-12.775422	-14.981138	-17.14580	-19.39840
2p	-10.219106	-12.218289	-14.17782	-16.22419
3s	-1.233137	-1.655845	-1.94114	-2.21070
3p	-0.646636	-0.981391	-1.18677	-1.37630
4s	-0.064460	-0.111359	-0.12562	-0.13516
3d			-0.08993	-0.12742

Atom	V	Cr	Mn	Fe	Co	Ni
Z	23	24	25	26	27	28
1s	-195.11079	-213.87746	-233.50875	-254.00470	-275.36535	-297.59075
2s	-21.72028	-24.14440	-26.68762	-29.35014	-32.13212	-35.03372
2p	-18.33888	-20.55424	-22.88690	-25.33699	-27.90468	-30.59009
3s	-2.44810	-2.68033	-2.92165	-3.17214	-3.43191	-3.70102
3p	-1.53340	-1.68342	-1.83995	-2.00304	-2.17274	-2.34907
4s	-0.13684	-0.13575	-0.13474	-0.13379	-0.13292	-0.13212
3d	-0.13684	-0.13575	-0.13474	-0.13379	-0.13292	-0.13212
<i>n</i> (3d)	0.6393	0.8873	1.1278	1.3622	1.5918	1.8174

Atom	Cu	Zn	Ga	Ge
Z	29	30	31	32
1s	-320.711183	-344.885966	-370.087065	-396.206872
2s	-38.088382	-41.471174	-45.140343	-48.991790
2p	-33.426318	-36.586685	-40.030943	-43.654803
3s	-4.010749	-4.519851	-5.188704	-5.906101
3p	-2.562693	-2.969457	-3.532081	-4.139819
3d	-0.157720	-0.348234	-0.685727	-1.064181
4s	-0.138533	-0.185366	-0.290872	-0.386783
4p			-0.070624	-0.114696

Atom	As	Se	Br	Kr
Z	33	34	35	36
1s	-423.248196	-451.209748	-480.090322	-509.889039
2s	-53.026929	-57.243491	-61.639549	-66.213681
2p	-47.459904	-51.444139	-55.605706	-59.943283
3s	-6.673183	-7.487710	-8.347907	-9.252538
3p	-4.794502	-5.494354	-6.237921	-7.024197
3d	-1.487148	-1.953579	-2.462342	-3.012574
4s	-0.481338	-0.576513	-0.673116	-0.771572
4p	-0.158885	-0.20426	-0.251199	-0.299874

*Fourth row.*

Atom	Rb	Sr	Y	Zr
Z	37	38	39	40
1s	-540.863861	-572.774871	-605.539841	-639.200123
2s	-71.219637	-76.418197	-81.718973	-87.167101
2p	-64.711316	-69.670502	-74.731216	-79.938205
3s	-10.452293	-11.708284	-12.932519	-14.171025
3p	-8.104015	-9.238678	-10.340292	-11.455022
3d	-3.854833	-4.750868	-5.612293	-6.485549
4s	-1.088064	-1.407019	-1.651693	-1.873159
4p	-0.547366	-0.798079	-0.980422	-1.141874
5s	-0.061487	-0.102737	-0.120721	-0.131037
4d			-0.071919	-0.111534

Atom	Nb	Mo	Tc	Ru
Z	41	42	43	44
1s	-673.74149	-709.15136	-745.48044	-782.72787
2s	-92.74707	-98.44597	-104.31826	-110.36286
2p	-85.27606	-90.73190	-96.35989	-102.15896
3s	-15.40918	-16.63439	-17.90004	-19.20531
3p	-12.56830	-13.66757	-14.80629	-15.98365
3d	-7.35588	-8.21062	-9.10349	-10.03361
4s	-2.05942	-2.19877	-2.34006	-2.48279
4p	-1.27048	-1.35425	-1.43939	-1.52544
5s	-0.13172	-0.11937	-0.10617	-0.09183
4d	-0.13172	-0.11937	-0.10617	-0.09183
n (4d)	0.6535	0.9847	1.2956	1.5896

Atom	Rh	Pd	Ag	Cd	In
Z	45	46	47	48	49
1s	-820.927173	-860.048546	-900.232540	-941.381019	-983.552576
2s	-116.614569	-123.041777	-129.790427	-136.759252	-144.005647
2p	-108.163817	-114.343011	-120.842024	-127.559951	-134.554225
3s	-20.585170	-22.008434	-23.620128	-25.317963	-27.159345
3p	-17.234646	-18.528092	-20.009041	-21.575259	-23.284171
3d	-11.035987	-12.079263	-13.308869	-14.622541	-16.077676
4s	-2.661143	-2.845456	-3.173860	-3.543470	-4.010922
4p	-1.645733	-1.771555	-2.037653	-2.343065	-2.744597
4d	-0.103288	-0.118970	-0.252103	-0.420723	-0.681578
5s			-0.124136	-0.167825	-0.253924
5p					-0.071162

Atom	Sn	Sb	Te	I	Xe
Z	50	51	52	53	54
1s	-1026.665599	-1070.725180	-1115.731902	-1161.685673	-1208.586286
2s	-151.449408	-159.095276	-166.943588	-174.994060	-183.246330
2p	-141.744613	-149.135914	-156.728514	-164.522166	-172.516543
3s	-29.062993	-31.033521	-33.071174	-35.175601	-37.346393
3p	-25.054553	-26.891049	-28.793930	-30.762871	-32.797483
3d	-17.593291	-19.174056	-20.820270	-22.531629	-24.307764
4s	-4.493043	-4.994724	-5.516439	-6.058048	-6.619330
4p	-3.159222	-3.592188	-4.044198	-4.515264	-5.005277
4d	-0.954355	-1.244953	-1.554330	-1.882595	-2.229668
5s	-0.330583	-0.404626	-0.477952	-0.551382	-0.625352
5p	-0.110212	-0.148390	-0.186783	-0.225814	-0.265689

### Appendix C: Accidental degeneracies and nonuniqueness of the rHF ground state density matrix

When the Fermi level is negative and contains a pair of accidentally degenerate s and d shells, any nonmagnetic rHF ground state density matrix is of the form

$$\gamma_{Z,Z}^{0,\text{rHF}} = 2\mathbb{1}_{(-\infty, \epsilon_{Z,Z,\text{F}}^{0,\text{rHF}})}(H_{Z,Z}^{0,\text{rHF}}) + \alpha|\phi_s\rangle\langle\phi_s| + \sum_{m,m'=-2}^2 \beta_{m,m'}|\phi_{d,m}\rangle\langle\phi_{d,m'}| \\ + \sum_{m=-2}^2 \gamma_m(|\phi_s\rangle\langle\phi_{d,m}| + |\phi_{d,m}\rangle\langle\phi_s|),$$

where  $\alpha \in \mathbb{R}$ ,  $\beta \in \mathbb{R}_{\text{sym}}^{5 \times 5}$  and  $\gamma \in \mathbb{R}^5$  are matrices such that  $0 \leq \begin{pmatrix} \alpha & \gamma^T \\ \gamma & \beta \end{pmatrix} \leq 2$ , and

$$\phi_s(\mathbf{r}) = f_{ns}(r), \quad \phi_{d,m}(\mathbf{r}) = r^2 f_{n'd}(r) \tilde{Y}_2^m(\theta, \varphi).$$

Here, the  $\tilde{Y}_l^m$  are the real spherical harmonics, and  $f_{ns}$  and  $f_{n'd}$  are radial functions with  $(n-1)$  and  $(n'-3)$  nodes in  $(0, +\infty)$ , respectively. Since all the ground state density matrices share the same density, the function

$$\begin{aligned} \alpha^2 f_{ns}(r)^2 + \frac{\sqrt{15}}{\pi} f_{ns}(r) f_{n'd}(r) & \left( \gamma_{-2}xy + \gamma_{-1}yz + \gamma_0 \frac{2z^2 - x^2 - y^2}{\sqrt{3}} \right. \\ & \left. + \gamma_1xz + \gamma_2 \frac{x^2 - y^2}{2} \right) \\ + \frac{15}{4\pi} f_{n'd}(r)^2 & \left( \beta_{-2,-2}x^2y^2 + \beta_{-1,-1}y^2z^2 + \beta_{0,0} \frac{(2z^2 - x^2 - y^2)^2}{\sqrt{3}} + \beta_{1,1}x^2z^2 \right. \\ & + \beta_{2,2} \frac{(x^2 - y^2)^2}{4} + 2\beta_{-2,-1}xy^2z + \beta_{-2,0} \frac{xy(2z^2 - x^2 - y^2)}{\sqrt{3}} \\ & + 2\beta_{-2,1}x^2yz + \beta_{-2,2}xy(x^2 - y^2) + \beta_{-1,0} \frac{yz(2z^2 - x^2 - y^2)}{12} \\ & + 2\beta_{-1,1}xyz^2 + \beta_{-1,2}yz(x^2 - y^2) + \beta_{0,1} \frac{xz(2z^2 - x^2 - y^2)}{\sqrt{3}} \\ & \left. + \beta_{0,2} \frac{(x^2 - y^2)(2z^2 - x^2 - y^2)}{2\sqrt{3}} + \beta_{1,2}xz(x^2 - y^2) \right) \end{aligned}$$

where  $r = (x^2 + y^2 + z^2)^{1/2}$ , must be a function of  $r$ , independent of the chosen ground state density matrix. Since  $f_{ns}$  has more nodes than  $f_{n'd}$  (we have seen above that  $n = 5$  or  $6$  and  $n' = 3$  or  $4$ ), this implies that  $\beta$  is a scalar matrix, that  $\gamma = 0$ , and that only one value for the pair  $(\alpha, \beta)$  is possible. This demonstrates the uniqueness of the nonmagnetic ground state when the Fermi level is negative and contains a pair of accidentally degenerate s and d shells.

When the Fermi level is negative and contains a pair of accidentally degenerate p and d shells, any nonmagnetic ground state density matrix is of the form

$$\begin{aligned} \gamma_{Z,Z}^{0,\text{rHF}} &= 2\mathbb{1}_{(-\infty, \epsilon_{Z,Z,\text{F}}^{0,\text{rHF}})}(H_{Z,Z}^{0,\text{rHF}}) + \sum_{m,m'=-1}^1 \alpha_{m,m'} |\phi_{p,m}\rangle \langle \phi_{p,m'}| \\ + \sum_{m,m'=-2}^2 \beta_{m,m'} |\phi_{d,m}\rangle \langle \phi_{d,m'}| &+ \sum_{m=-1}^1 \sum_{m'=-2}^2 \gamma_{m,m'} (|\phi_{p,m}\rangle \langle \phi_{d,m'}| + |\phi_{d,m'}\rangle \langle \phi_{p,m}|) \quad (59) \end{aligned}$$

where  $\alpha \in \mathbb{R}_{\text{sym}}^{3 \times 3}$ ,  $\beta \in \mathbb{R}_{\text{sym}}^{5 \times 5}$  and  $\gamma \in \mathbb{R}^{3 \times 5}$  are matrices such that  $0 \leq \begin{pmatrix} \alpha & \gamma \\ \gamma^T & \beta \end{pmatrix} \leq 2$ , and

$$\phi_{p,m}(\mathbf{r}) = r f_{np}(r) \tilde{Y}_1^m(\theta, \phi), \quad \phi_{d,m}(\mathbf{r}) = r^2 f_{n'd}(r) \tilde{Y}_2^m(\theta, \varphi).$$

Here,  $f_{np}$  and  $f_{n'd}$  are radial functions with  $(n-2)$  and  $(n'-3)$  nodes in  $(0, +\infty)$ , respectively. Since all the ground state density matrices share the same density, the



function

$$\begin{aligned}
 & \frac{3}{4\pi} f_{np}(r)^2 (\alpha_{-1,-1}y^2 + \alpha_{0,0}z^2 + \alpha_{1,1}x^2 + 2\alpha_{-1,0}yz + 2\alpha_{-1,1}xy + 2\alpha_{0,1}xz) \\
 & + \frac{3\sqrt{5}}{2\pi} f_{np}(r) f_{n'd}(r) \left( \gamma_{-1,-2}xy^2 + \gamma_{-1,-1}y^2z + \gamma_{-1,0} \frac{y(2z^2 - x^2 - y^2)}{2\sqrt{3}} \right. \\
 & \quad + \gamma_{-1,1}xyz + \gamma_{-1,2} \frac{y(x^2 - y^2)}{2} + \gamma_{0,-2}xyz + \gamma_{0,-1}yz^2 \\
 & \quad + \gamma_{0,0} \frac{z(2z^2 - x^2 - y^2)}{2\sqrt{3}} + \gamma_{0,1}xz^2 + \gamma_{0,2} \frac{z(x^2 - y^2)}{2} + \gamma_{1,-2}x^2y \\
 & \quad \left. + \gamma_{1,-1}xyz + \gamma_{1,0} \frac{x(2z^2 - x^2 - y^2)}{2\sqrt{3}} + \gamma_{1,1}x^2z + \gamma_{1,2} \frac{x(x^2 - y^2)}{2} \right) \\
 & + \frac{15}{4\pi} f_{n'd}(r)^2 \left( \beta_{-2,-2}x^2y^2 + \beta_{-1,-1}y^2z^2 + \beta_{0,0} \frac{(2z^2 - x^2 - y^2)^2}{12} \right. \\
 & \quad + \beta_{1,1}x^2z^2 + \beta_{2,2} \frac{(x^2 - y^2)^2}{4} + 2\beta_{-2,-1}xy^2z + \beta_{-2,0} \frac{xy(2z^2 - x^2 - y^2)}{\sqrt{3}} \\
 & \quad + 2\beta_{-2,1}x^2yz + \beta_{-2,2}xy(x^2 - y^2) + \beta_{-1,0} \frac{yz(2z^2 - x^2 - y^2)}{\sqrt{3}} \\
 & \quad + 2\beta_{-1,1}xyz^2 + \beta_{-1,2}yz(x^2 - y^2) + \beta_{0,1} \frac{xz(2z^2 - x^2 - y^2)}{\sqrt{3}} \\
 & \quad \left. + \beta_{0,2} \frac{(x^2 - y^2)(2z^2 - x^2 - y^2)}{2\sqrt{3}} + \beta_{1,2}xz(x^2 - y^2) \right),
 \end{aligned}$$

where  $r = (x^2 + y^2 + z^2)^{1/2}$ , must be a function of  $r$ , independent of the chosen ground state density matrix. Since  $f_{np}$  has more nodes than  $f_{n'd}$ , this implies that  $\alpha$  and  $\beta$  are scalar matrices and that, for  $\alpha$  and  $\beta$  given, the function

$$\begin{aligned}
 & \gamma_{-1,-2}xy^2 + \gamma_{-1,-1}y^2z + \gamma_{-1,0} \frac{y(2z^2 - x^2 - y^2)}{2\sqrt{3}} + \gamma_{-1,1}xyz + \gamma_{-1,2} \frac{y(x^2 - y^2)}{2} \\
 & + \gamma_{0,-2}xyz + \gamma_{0,-1}yz^2 + \gamma_{0,0} \frac{z(2z^2 - x^2 - y^2)}{2\sqrt{3}} + \gamma_{0,1}xz^2 + \gamma_{0,2} \frac{z(x^2 - y^2)}{2} \\
 & + \gamma_{1,-2}x^2y + \gamma_{1,-1}xyz + \gamma_{1,0} \frac{x(2z^2 - x^2 - y^2)}{2\sqrt{3}} + \gamma_{1,1}x^2z + \gamma_{1,2} \frac{x(x^2 - y^2)}{2}
 \end{aligned}$$

is a given function of  $r$ . The vector space of homogeneous polynomials in  $x, y, z$  of total degree equal to 3 is of dimension 10, and the matrix  $\gamma$  has 15 independent entries. Provided  $\alpha$  and  $\beta$  are not equal to 0 (which is suggested by our numerical simulations), an infinity of density matrices of the form (59) satisfy the rHF equations, and are therefore admissible nonmagnetic ground states.

### Acknowledgements

The authors are grateful to Carlos García-Cervera and Vikram Gavini for valuable discussions.

### References

- [1] A. Anantharaman and E. Cancès, *Existence of minimizers for Kohn–Sham models in quantum chemistry*, Ann. Inst. H. Poincaré Anal. Non Linéaire **26** (2009), no. 6, 2425–2455. MR Zbl
- [2] H. A. Bethe and E. E. Salpeter, *Quantum mechanics of one- and two-electron atoms*, Springer, 1957. MR Zbl
- [3] E. Cancès, A. Deleurence, and M. Lewin, *A new approach to the modeling of local defects in crystals: the reduced Hartree–Fock case*, Comm. Math. Phys. **281** (2008), no. 1, 129–177. MR Zbl
- [4] E. Cancès, S. Lahbabi, and M. Lewin, *Mean-field models for disordered crystals*, J. Math. Pures Appl. (9) **100** (2013), no. 2, 241–274. MR Zbl
- [5] E. Cancès and C. Le Bris, *Can we outperform the DIIS approach for electronic structure calculations?*, Int. J. Quantum Chem. **79** (2000), no. 2, 82–90.
- [6] E. Cancès and C. Le Bris, *On the convergence of SCF algorithms for the Hartree–Fock equations*, M2AN Math. Model. Numer. Anal. **34** (2000), no. 4, 749–774. MR Zbl
- [7] E. Cancès and N. Mourad, *A mathematical perspective on density functional perturbation theory*, Nonlinearity **27** (2014), no. 9, 1999–2033. MR Zbl
- [8] ———, *Existence of a type of optimal norm-conserving pseudopotentials for Kohn–Sham models*, Commun. Math. Sci. **14** (2016), no. 5, 1315–1352. MR Zbl
- [9] ———, *A numerical study of the extended Kohn–Sham ground states of atoms*, preprint, 2017. arXiv
- [10] I. Catto, C. Le Bris, and P.-L. Lions, *On the thermodynamic limit for Hartree–Fock type models*, Ann. Inst. H. Poincaré Anal. Non Linéaire **18** (2001), no. 6, 687–760. MR Zbl
- [11] E. U. Condon and H. Odabaşı, *Atomic structure*, Cambridge University, 1980.
- [12] P. A. M. Dirac, *Quantum mechanics of many-electron systems*, P. Roy. Soc. Lond. A **123** (1929), no. 792, 714–733. Zbl
- [13] R. M. Dreizler and E. K. U. Gross, *Density functional theory: an approach to the quantum many-body problem*, Springer, 1990. Zbl
- [14] C. Fefferman and L. A. Seco, *On the Dirac and Schwinger corrections to the ground-state energy of an atom*, Adv. Math. **107** (1994), no. 1, 1–185. MR Zbl
- [15] G. Friesecke and B. D. Goddard, *Explicit large nuclear charge limit of electronic ground states for Li, Be, B, C, N, O, F, Ne and basic aspects of the periodic table*, SIAM J. Math. Anal. **41** (2009), no. 2, 631–664. MR Zbl
- [16] C. Froese Fischer, *The Hartree–Fock method for atoms: a numerical approach*, Wiley, 1977.
- [17] C. Froese Fischer, T. Brage, and P. Jönsson, *Computational atomic structure: an MCHF approach*, Institute of Physics, 1997. Zbl
- [18] D. Gontier, *N-representability in noncollinear spin-polarized density-functional theory*, Phys. Rev. Lett. **111** (2013), no. 15, 153001.
- [19] M. Hoffmann-Ostenhof and T. Hoffmann-Ostenhof, *“Schrödinger inequalities” and asymptotic behavior of the electron density of atoms and molecules*, Phys. Rev. A (3) **16** (1977), no. 5, 1782–1785. MR

- [20] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev. (2) **140** (1965), A1133–A1138. MR
- [21] V. Korobov and A. Yelkhovsky, *Ionization potential of the helium atom*, Phys. Rev. Lett. **87** (2001), no. 19, 193003.
- [22] E. S. Kryachko and E. V. Ludeña, *Density functional theory: foundations reviewed*, Phys. Rep. **544** (2014), no. 2, 123–239. MR
- [23] M. Levy, *Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the  $v$ -representability problem*, Proc. Nat. Acad. Sci. U.S.A. **76** (1979), no. 12, 6062–6065. MR
- [24] E. H. Lieb, *Density functionals for Coulomb systems*, Int. J. Quantum Chem. **24** (1983), no. 3, 243–277.
- [25] E. H. Lieb and B. Simon, *The Hartree–Fock theory for Coulomb systems*, Comm. Math. Phys. **53** (1977), no. 3, 185–194. MR
- [26] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, *Density functional theory is straying from the path toward the exact functional*, Science **355** (2017), no. 6320, 49–52.
- [27] N. Mourad, *Numerical computation of the extended Kohn–Sham ground states of atoms for cylindrically symmetric systems*, source code, 2017, version 1.0.
- [28] R. G. Parr and Y. Weitao, *Density-functional theory of atoms and molecules*, International Series of Monographs on Chemistry, no. 16, Oxford University, 1989.
- [29] J. P. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electron systems*, Phys. Rev. B **23** (1981), no. 10, 5048–5079.
- [30] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in FORTRAN: the art of scientific computing*, 2nd ed., Cambridge University, 1992. MR Zbl
- [31] M. Reed and B. Simon, *Methods of modern mathematical physics*, vol. IV: Analysis of operators, Academic, 1978. MR Zbl
- [32] M. E. Rose, *Elementary theory of angular momentum*, Wiley, 1957. MR Zbl
- [33] H. Siedentop and R. Weikard, *On the leading energy correction for the statistical model of the atom: interacting case*, Comm. Math. Phys. **112** (1987), no. 3, 471–490. MR Zbl
- [34] J. C. Slater, *A simplification of the Hartree–Fock method*, Phys. Rev. **81** (1951), no. 3, 385–390. Zbl
- [35] J. P. Solovej, *Proof of the ionization conjecture in a reduced Hartree–Fock model*, Invent. Math. **104** (1991), no. 2, 291–311. MR Zbl
- [36] J. P. Solovej, T. O. Sørensen, and W. L. Spitzer, *Relativistic Scott correction for atoms and molecules*, Comm. Pure Appl. Math. **63** (2010), no. 1, 39–118. MR Zbl
- [37] A. H. Stroud and D. Secrest, *Gaussian quadrature formulas*, Prentice-Hall, 1966. MR Zbl
- [38] L. Szasz, *The electronic structure of atoms*, Wiley, 1992.
- [39] M. Trsic and A. B. da Silva, *Electronic, atomic and molecular calculations: applying the generator coordinate method*, Elsevier, 2007.
- [40] S. M. Valone, *Consequences of extending 1 matrix energy functionals from pure-state representable to all ensemble representable 1 matrices*, J. Chem. Phys. **73** (1980), no. 3, 1344–1349. MR

Received February 7, 2017. Revised February 25, 2018.

ERIC CANCÈS: [eric.cances@enpc.fr](mailto:eric.cances@enpc.fr)

*Centre d'Enseignement et de Recherche en Mathématiques et Calcul Scientifique,  
Ecole des Ponts ParisTech and Institut National de Recherche en Informatique et en Automatique Paris,  
Université Paris-Est, Marne-la-Vallée, France*

NAHIA MOURAD: [nahia.mourad@gmail.com](mailto:nahia.mourad@gmail.com)

*Centre d'Enseignement et de Recherche en Mathématiques et Calcul Scientifique,  
Ecole des Ponts ParisTech, Université Paris-Est, Marne-la-Vallée, France*

## AN EQUATION-BY-EQUATION METHOD FOR SOLVING THE MULTIDIMENSIONAL MOMENT CONSTRAINED MAXIMUM ENTROPY PROBLEM

WENRUI HAO AND JOHN HARLIM

An equation-by-equation (EBE) method is proposed to solve a system of nonlinear equations arising from the moment constrained maximum entropy problem of multidimensional variables. The design of the EBE method combines ideas from homotopy continuation and Newton's iterative methods. Theoretically, we establish the local convergence under appropriate conditions and show that the proposed method, geometrically, finds the solution by searching along the surface corresponding to one component of the nonlinear problem. We will demonstrate the robustness of the method on various numerical examples, including (1) a six-moment one-dimensional entropy problem with an explicit solution that contains components of order  $10^0$ – $10^3$  in magnitude, (2) four-moment multidimensional entropy problems with explicit solutions where the resulting systems to be solved range from 70–310 equations, and (3) four- to eight-moment of a two-dimensional entropy problem, whose solutions correspond to the densities of the two leading EOFs of the wind stress-driven large-scale oceanic model. In this case, we find that the EBE method is more accurate compared to the classical Newton's method, the MATLAB generic solver, and the previously developed BFGS-based method, which was also tested on this problem. The fourth example is four-moment constrained up to five-dimensional entropy problems whose solutions correspond to multidimensional densities of the components of the solutions of the Kuramoto–Sivashinsky equation. For the higher-dimensional cases of this example, the EBE method is superior because it automatically selects a subset of the prescribed moment constraints from which the maximum entropy solution can be estimated within the desired tolerance. This selection feature is particularly important since the moment constrained maximum entropy problems do not necessarily have solutions in general.

---

Hao's research was partially supported by the American Heart Association (Grant 17SDG33660722), the National Science Foundation (Grant DMS-1818769) and an Institute for CyberScience Seed Grant. Harlim's research was partially supported by the Office of Naval Research (Grant N00014-16-1-2888) and the National Science Foundation (Grant DMS-1619661).

*MSC2010:* 65H10, 65H20, 94A17, 49M15.

*Keywords:* homotopy continuation, moment constrained, maximum entropy, equation-by-equation method.

## 1. Introduction

The maximum entropy principle provides a natural criterion for estimating the least biased density function subjected to the given moments [14]. This density estimation approach has a wide range of applications, such as the harmonic solid and quantum spin systems [20], econometrics [26], and geophysical applications [5; 13]. In a nutshell, this moment constrained method is a parametric estimation technique where the resulting density function is in the form of an exponential of polynomials. This is a consequence of maximizing the Shannon entropy subjected to the polynomial moment constraints, which is usually transformed into an unconstrained minimization problem of a Lagrangian function [27]. Standard approaches for solving this unconstrained minimization problem are based on Newton's iterative method [1; 27] or a quasi-Newton-based method such as the BFGS method [3; 4].

In the last two papers [3; 4], where the BFGS-based method was introduced and reviewed, Abramov considered minimization problems that involve 44–83 equations, resulting from a two-dimensional problem with moment constraints of up to order eight, a three-dimensional problem with moment constraints of up to order six, and a four-dimensional problem with moment constraints of up to order four. In this paper, we introduce a novel equation solver that can be used to find density functions of moderately high-dimensional problems (e.g., systems of 70–310 equations resulting from moments up to order four of four- to seven-dimensional density functions) provided that the solutions exist. The proposed method, which we called the equation-by-equation (EBE) method, is an iterative method that solves a one-dimensional problem at the first iterate, a two-dimensional problem at the second iterate, a three-dimensional problem at the third iterate, and eventually solves the full system of nonlinear equations corresponding to the maximum entropy problem at the last iterate. Technically, this method combines Newton's method with ideas from homotopy continuation. We will show that the EBE method is locally convergent under appropriate conditions. Furthermore, we will provide sufficient conditions for global convergence. Through the convergence analysis, we will show that, geometrically, the proposed method finds the solution of the nonlinear system of equations by tracking along the surface corresponding to one component of the system of nonlinear equations. The EBE method automatically selects a subset of the prescribed constraints from which the maximum entropy solution can be estimated within the desired tolerance. This is an important feature since the maximum entropy problems do not necessarily have solutions for general sets of moment constraints.

We shall find that the EBE method produces more accurate solutions (smaller error in the moments) compared to the classical Newton's method, MATLAB's built-in `fsolve`, and BFGS method on the test problem in [3; 4] and on test problems

based on the solutions of the Kuramoto–Sivashinsky equation. Numerically, we will demonstrate that the EBE method is able to solve problems where the true solutions consist of components of order  $10^0$ – $10^3$ . We shall also see that the EBE method can solve a system of hundreds of equations in various examples, including those with explicit solutions as well as those with densities estimated based on solutions of complex spatially extended dynamical systems.

The remaining part of the paper is organized as follows. In Section 2, we give a brief overview of the multidimensional maximum entropy problem. In Section 3, we introduce the EBE algorithm. In Section 4, we provide the local convergence analysis. In Section 5, we discuss the practical issues with the proposed method and provide remedies. In Section 6, we demonstrate the robustness of the EBE method on various numerical examples. In Section 7, we conclude the paper with a brief summary and discussion. We include an Appendix to show some computational details that are left out in the main text. Interested readers and users can access the EBE codes (written in MATLAB) at [10].

## 2. An overview of the maximum entropy problem

We consider the Hausdorff moment-constrained maximum entropy problem [1; 4; 8]. That is, find the optimal probability density  $\rho^*(\mathbf{x})$  which maximizes the Shannon entropy

$$S(\rho) := - \int_{\Omega} \log(\rho(\mathbf{x}))\rho(\mathbf{x}) d\mathbf{x}, \tag{1}$$

where  $\mathbf{x} \in \Omega = [-1, 1]^d$  satisfies the linear constraints

$$\mathcal{F}_j := \int_{\Omega} c_j(\mathbf{x})\rho(\mathbf{x}) d\mathbf{x} = f_j, \quad |\mathbf{j}| = 0, 1, 2, \dots, p. \tag{2}$$

In applications, one usually computes the statistics  $f_j$  from samples of data. For arbitrary finite domain, one can rescale the data to the domain  $\Omega$ .

While  $c_j(\mathbf{x})$  can be arbitrary functions in  $L^1(\Omega, \rho)$ , we will focus on the usual uncentered statistical moments with monomial basis functions,  $c_j(\mathbf{x}) = \mathbf{x}^j$  in this article, where we have adopted the notations  $\mathbf{x} = (x_1, \dots, x_d) \in \Omega$ ,  $\mathbf{j} = (j_1, \dots, j_d) \in \mathbb{Z}_+^d$  with  $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ , and  $\mathbf{x}^j = \prod_{i=1}^d x_i^{j_i}$ . In (2), the quantities  $f_j$  are the given  $\mathbf{j}$ -th moments that can be computed from the data. Since the total number of monomials  $\mathbf{x}^j$  where  $|\mathbf{j}| = j$  is  $C_{d-1}^{j+d-1}$ , then the total number of constraints in (2) for moments up to order  $p$  is

$$n = \sum_{j=1}^p C_{d-1}^{j+d-1},$$

excluding the normalization factor corresponding to  $c_0(\mathbf{x}) = 1$ . For example, in a two-dimensional problem, the total number of moments up to order  $p = 4$  is  $n = 14$ .

To simplify the notation below, we will use a single index notation and understand that the total number of constraints to be satisfied is  $n$ , excluding the zeroth moment. The exclusion of the zeroth moment will be clear as we discuss below.

By introducing Lagrange multipliers, the above constrained optimization problem can be transformed into the unconstrained problem

$$\mathcal{L}(\rho(\mathbf{x}), \lambda_0, \dots, \lambda_n) = S(\rho) + \sum_{j=0}^n \lambda_j (\mathcal{F}_j - f_j). \quad (3)$$

In order to find a solution of (3), we set  $\frac{\partial \mathcal{L}}{\partial \rho} = 0$ , which gives

$$\rho(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{j=1}^n \lambda_j c_j(\mathbf{x})\right), \quad (4)$$

where we have defined  $Z = \exp(1 - \lambda_0)$ . Since  $\int_{\Omega} \rho(\mathbf{x}) d\mathbf{x} = 1$ , we have

$$Z(\lambda_1, \dots, \lambda_n) = \int_{\Omega} \exp\left(\sum_{j=1}^n \lambda_j c_j(\mathbf{x})\right) d\mathbf{x}, \quad (5)$$

which indicates that  $Z$  (or implicitly  $\lambda_0$ ) is a function of  $\lambda_1, \dots, \lambda_n$ . Therefore, the normalization factor  $Z$  can be computed via (5) once  $\lambda_1, \dots, \lambda_n$  are estimated. Therefore, we can just concentrate on finding the Lagrange multipliers  $\lambda_1, \dots, \lambda_n$  which satisfy  $n$  constraints in (2), excluding the case  $c_0(\mathbf{x}) = 1$ . In particular, the constrained maximum entropy problem is to solve the nonlinear system of integral equations

$$\begin{aligned} F_j(\lambda_1, \dots, \lambda_n) &:= \mathcal{F}_j(\lambda_1, \dots, \lambda_n) - f_j \\ &= \int_{\Omega} (c_j(\mathbf{x}) - f_j) \exp\left(\sum_{k=1}^n \lambda_k c_k(\mathbf{x})\right) d\mathbf{x} = 0, \quad j = 1, \dots, n, \end{aligned} \quad (6)$$

for  $\lambda_1, \dots, \lambda_n$ .

In our numerical implementation, the integral in system (6) will be approximated with a nested sparse grid quadrature rule [9]

$$\int_{\Omega} f(\mathbf{x}) d\mathbf{x} \approx \sum_i f(\mathbf{x}_i) w_i,$$

where  $\mathbf{x}_i$  are the nested sparse grid nodes, and  $w_i$  are the corresponding weights based on the nested Clenshaw–Curtis quadrature rule [25]. The number of nodes depends on the dimension of the problem  $d$ , and the number of the nested set (based on the Smolyak construction [23]) is denoted with the parameter  $\ell$  (referred to as the level). In the numerical implementation, we need to specify the parameter  $\ell$ .



### 3. An equation-by-equation algorithm

In this section, we describe the new equation-by-equation (EBE) technique to solve the system of equations in (6),

$$\mathbf{F}_n(\boldsymbol{\lambda}_n) = \mathbf{0}, \tag{7}$$

where we have defined

$$\mathbf{F}_n(\boldsymbol{\lambda}_n) := (F_1(\boldsymbol{\lambda}_n), \dots, F_n(\boldsymbol{\lambda}_n)),$$

and  $\boldsymbol{\lambda}_n = (\lambda_1, \dots, \lambda_n)$ . In the following iterative scheme, we start the iteration with an initial condition  $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ . We define  $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^i$  as the exact solution to the  $i$ -dimensional system

$$\mathbf{F}_i(\boldsymbol{\lambda}_i, \alpha_{i+1}, \dots, \alpha_n) = \mathbf{0}, \quad i = 1, \dots, n, \tag{8}$$

where we have fixed the last  $n - i$  coefficients,  $\lambda_{i+1} = \alpha_{i+1}, \dots, \lambda_n = \alpha_n$ . With this notation, the exact solution for (7) is  $\boldsymbol{\mu}^{(n)} \in \mathbb{R}^n$ . We also define  $\hat{\boldsymbol{\mu}}^{(i)}$  to be the numerical estimate of  $\boldsymbol{\mu}^{(i)}$ . With these notations, we now describe the algorithm.

Generally speaking, at each iteration  $i$ , where  $i = 1, \dots, n$ , the EBE algorithm solves  $i$ -dimensional system in (8). At each step  $i$ , given the numerical solution at the previous step  $\hat{\boldsymbol{\mu}}^{(i-1)} \in \mathbb{R}^{i-1}$  and initial condition  $\alpha_i$ , we apply an idea from homotopy continuation to find the solution  $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^i$  that solves the  $i$ -dimensional system of equations (8). Notice that we do not only add a new equation  $F_i(\boldsymbol{\lambda}_i, \alpha_{i+1}, \dots, \alpha_n) = 0$  but we also estimate the  $i$ -th variable in the previous  $i - 1$  equations  $\mathbf{F}_{i-1}(\boldsymbol{\lambda}_i, \alpha_{i+1}, \dots, \alpha_n) = \mathbf{0}$ . The scheme proceeds by solving the larger systems one by one until  $i = n$  so we eventually solve (7).

Now let us describe how to numerically estimate  $\boldsymbol{\mu}^{(i)}$  at every step  $i$ . For the first step  $i = 1$ , we solve the one-dimensional problem

$$F_1(\boldsymbol{\lambda}_1, \alpha_2, \dots, \alpha_n) = 0$$

for  $\lambda_1$  with Newton's method. For the steps  $i = 2, \dots, n$ , we have  $\hat{\boldsymbol{\mu}}^{(i-1)}$  which are the numerical estimates of  $\mathbf{F}_{i-1}(\boldsymbol{\lambda}_{i-1}, \alpha_i, \dots, \alpha_n) = \mathbf{0}$ . To simplify the expression below, let us use  $F_i(\boldsymbol{\lambda}_{i-1}, \lambda_i)$  as a short-hand notation for  $F_i(\boldsymbol{\lambda}_{i-1}, \lambda_i, \alpha_i, \dots, \alpha_n)$  to emphasize the independent variables.

We proceed to estimate  $\lambda_i$  using Newton's method with  $\text{Tol}_1$  on the  $i$ -th equation. That is, we iterate

$$\begin{aligned} \lambda_i^{m+1} &= \lambda_i^m - \left( \frac{\partial F_i}{\partial \lambda_i}(\boldsymbol{\lambda}_{i-1}^m, \lambda_i^m) \right)^{-1} F_i(\boldsymbol{\lambda}_{i-1}^m, \lambda_i^m), \quad m = 0, 1, \dots, \\ \lambda_i^0 &= \alpha_i, \quad \boldsymbol{\lambda}_{i-1}^0 = \hat{\boldsymbol{\mu}}^{(i-1)} \end{aligned} \tag{9}$$

assuming that  $\frac{\partial F_i}{\partial \lambda_i}(\lambda_{i-1}^m, \lambda_i^m) \neq 0$ . Here, the partial derivative of  $F_i$  with respect to  $\lambda_i$  evaluated at  $\lambda_i^m$  is defined as

$$\frac{\partial F_i}{\partial \lambda_i}(\lambda_{i-1}^m, \lambda_i^m) = \int_{\Omega} (c_i(\mathbf{x}) - f_i)c_i(\mathbf{x}) \exp\left(\sum_{j=1}^{i-1} \lambda_j^m c_j(\mathbf{x}) + \lambda_i^m c_i(\mathbf{x})\right) d\mathbf{x}, \quad (10)$$

where we have denoted  $\lambda_{i-1}^m = (\lambda_1^m, \dots, \lambda_{i-1}^m)$ . Notice that to proceed the iteration in (9), we need to update  $\lambda_{i-1}^m$  for  $m > 0$ . We propose to follow the homotopy continuation method for this update. In particular, we are looking for  $\lambda_{i-1}^{m+1}$  that solves  $F_{i-1}(\lambda_{i-1}^{m+1}, \lambda_i^{m+1}) = \mathbf{0}$ , given the current estimate  $\lambda_i^{m+1}$  from (9) as well as  $F_{i-1}(\lambda_{i-1}^m, \lambda_i^m) = \mathbf{0}$ . At  $m = 0$ , this last constraint is numerically estimated by  $F_{i-1}(\hat{\boldsymbol{\mu}}^{(i-1)}, \alpha_i) \approx \mathbf{0}$ .

One way to solve this problem is through the following predictor-corrector step which is usually used in the homotopy continuation method [7; 24]. In particular, we apply Taylor's expansion to

$$F_{i-1}(\lambda_{i-1}^{m+1}, \lambda_i^{m+1}) = F_{i-1}(\lambda_{i-1}^m + \Delta\lambda, \lambda_i^m + (\lambda_i^{m+1} - \lambda_i^m)) = \mathbf{0}$$

at  $(\lambda_{i-1}^m, \lambda_i^m)$ , which gives

$$F_{i-1}(\lambda_{i-1}^m, \lambda_i^m) + F_{i-1, \lambda_{i-1}}(\lambda_{i-1}^m, \lambda_i^m)\Delta\lambda + F_{i-1, \lambda_i}(\lambda_{i-1}^m, \lambda_i^m)(\lambda_i^{m+1} - \lambda_i^m) = \mathbf{0},$$

which means that

$$\Delta\lambda = -F_{i-1, \lambda_{i-1}}^{-1}(\lambda_{i-1}^m, \lambda_i^m)F_{i-1, \lambda_i}(\lambda_{i-1}^m, \lambda_i^m)(\lambda_i^{m+1} - \lambda_i^m),$$

assuming that  $F_{i-1, \lambda_{i-1}}(\lambda_{i-1}^m, \lambda_i^m)$  is invertible. Based on this linear prediction,  $\lambda_{i-1}^{m+1}$  is approximated by

$$\begin{aligned} \tilde{\lambda}_{i-1}^{m+1} &= \lambda_{i-1}^m + \Delta\lambda \\ &= \lambda_{i-1}^m - F_{i-1, \lambda_{i-1}}^{-1}(\lambda_{i-1}^m, \lambda_i^m)F_{i-1, \lambda_i}(\lambda_{i-1}^m, \lambda_i^m)(\lambda_i^{m+1} - \lambda_i^m). \end{aligned} \quad (11)$$

Subsequently, when  $\|F_i(\tilde{\lambda}_{i-1}^{m+1}, \lambda_i^{m+1})\| \geq \text{Tol}_2$ , apply a correction using Newton's method by expanding

$$\mathbf{0} = F_{i-1}(\lambda_{i-1}^{m+1}, \lambda_i^{m+1}) = F_{i-1}(\tilde{\lambda}_{i-1}^{m+1}, \lambda_i^{m+1}) + F_{i-1, \lambda_{i-1}}(\tilde{\lambda}_{i-1}^{m+1}, \lambda_i^{m+1})\Delta\tilde{\lambda},$$

assuming that  $\lambda_{i-1}^{m+1} = \tilde{\lambda}_{i-1}^{m+1} + \Delta\tilde{\lambda}$ , to find that

$$\lambda_{i-1}^{m+1} = \tilde{\lambda}_{i-1}^{m+1} - F_{i-1, \lambda_{i-1}}(\tilde{\lambda}_{i-1}^{m+1}, \lambda_i^{m+1})^{-1}F_{i-1}(\tilde{\lambda}_{i-1}^{m+1}, \lambda_i^{m+1}). \quad (12)$$

This expression assumes that  $F_{i-1, \lambda_{i-1}}(\tilde{\lambda}_{i-1}^{m+1}, \lambda_i^{m+1})$  is invertible.

In summary, at each step  $i$ , we iterate (9), (11), and (12). So the outer loop  $i$  corresponds to adding one equation to the system at the time, and for each  $i$ , we apply an inner loop, indexed with  $m$ , to find the solution  $\boldsymbol{\mu}^{(i)}$  for  $F_i(\lambda_i, \alpha_{i+1}, \dots, \alpha_n) = \mathbf{0}$ .

We denote the approximate solution as  $\hat{\boldsymbol{\mu}}^{(i)}$ . An adaptive tolerance technique is employed to compute the initial guess of  $\mathbf{F}_i$  by using Newton's method. In particular, when the current tolerance  $\text{Tol}_2$  is not satisfied after executing (12), then we divide  $\text{Tol}_1$  by ten until  $\text{Tol}_2$  is met.

Recall that the standard Newton's method assumes that the Jacobian  $\mathbf{F}_{n,\lambda_n} \in \mathbb{R}^{n \times n}$  is nonsingular at the root of the full system in (6) to guarantee the local convergence. In the next section, we will show that the EBE method requires the following conditions for local convergence.

**Assumption 1.** Let  $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^i$  be a solution of  $\mathbf{F}_i(\boldsymbol{\lambda}_i, \alpha_{i+1}, \dots, \alpha_n) = \mathbf{0}$ , for each  $i = 1, \dots, n$ . The EBE method assumes the conditions

- (1)  $\frac{\partial \mathbf{F}_i}{\partial \boldsymbol{\lambda}_i}(\boldsymbol{\mu}^{(i)}, \alpha_{i+1}, \dots, \alpha_n) \neq 0$ ,
- (2)  $\mathbf{F}_{i,\lambda_i}(\boldsymbol{\mu}^{(i)}, \alpha_{i+1}, \dots, \alpha_n)$  are nonsingular, and
- (3) each component of  $\mathbf{F}_i$  is twice differentiable in a close region whose interior contains the solution  $\boldsymbol{\mu}^{(i)}$ .

These conditions are similar to the standard Newton's assumptions on each system of  $i$  equations. The smoothness condition will be used in the proof of the local convergence in the next section. Of course if one can specify initial conditions that are sufficiently close to the true solution, then one can simply apply Newton's method directly. With the EBE method, we can start with any arbitrary initial condition. Theoretically, this will require an additional condition beyond Assumption 1 for global convergence as we shall discuss in Section 4. In Section 5, we will provide several remedies when the initial condition is not close to the solution. In fact, we will always set the initial condition to zero in our numerical implementation in Section 6,  $\alpha_i = 0$  for all  $i = 1, \dots, n$ , and demonstrate that the EBE method is numerically accurate in the test problems with solutions that are far away from zero.

#### 4. Convergence analysis

In this section, we study the convergence of this method. First, let's concentrate on the convergence of the iteration (9), (11), and (12) for solving the  $i$ -dimensional system  $\mathbf{F}_i(\boldsymbol{\lambda}_{i-1}, \lambda_i, \alpha_{i+1}, \dots, \alpha_n) := \mathbf{F}_i(\boldsymbol{\lambda}_{i-1}, \lambda_i) = \mathbf{0}$  for  $\boldsymbol{\lambda}_{i-1}$  and  $\lambda_i$ . In compact form, these three steps can be written as an iterative map

$$(\boldsymbol{\lambda}_{i-1}^{m+1}, \lambda_{i+1}^{m+1}) = \mathbf{H}_i(\boldsymbol{\lambda}_{i-1}^m, \lambda_i^m), \quad (13)$$

where the map  $\mathbf{H}_i : \mathbb{R}^i \rightarrow \mathbb{R}^i$  is defined as

$$\mathbf{H}_i(\boldsymbol{\lambda}_{i-1}, \lambda_i) := \begin{pmatrix} \mathbf{g}_i - \mathbf{F}_{i-1,\lambda_{i-1}}(\mathbf{g}_i, H_{i,2})^{-1} \mathbf{F}_{i-1}(\mathbf{g}_i, H_{i,2}) \\ \lambda_i - \left( \frac{\partial \mathbf{F}_i}{\partial \lambda_i}(\boldsymbol{\lambda}_{i-1}, \lambda_i) \right)^{-1} \mathbf{F}_i(\boldsymbol{\lambda}_{i-1}, \lambda_i) \end{pmatrix}. \quad (14)$$

In (14), the notation  $H_{i,2}$  denotes the second component of (14) and

$$\mathbf{g}_i := \lambda_{i-1} - \mathbf{F}_{i-1, \lambda_{i-1}}(\lambda_{i-1}, \lambda_i)^{-1} \mathbf{F}_{i-1, \lambda_i}(\lambda_{i-1}, \lambda_i)(H_{i,2} - \lambda_i) \quad (15)$$

is defined exactly as in (11).

For notational convenience in the discussion below, we let the components of the exact solution of (8) be defined as  $\boldsymbol{\mu}^{(i)} := (\mu_{i-1}^{(i)}, \mu_i^{(i)}) \in \mathbb{R}^i$ . Here, we denote the first  $i-1$  components as  $\boldsymbol{\mu}_{i-1}^{(i)} = (\mu_1^{(i)}, \dots, \mu_{i-1}^{(i)}) \in \mathbb{R}^{i-1}$ . Similarly, we also denote  $\mathbf{H}_i = (H_{i,1}, H_{i,2})$ . First, we can deduce:

**Theorem 4.1.** *Let  $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^i$  be a fixed point of (13). Assume that  $\mathbf{F}_{i-1, \lambda_{i-1}}^* := \mathbf{F}_{i-1, \lambda_{i-1}}(\boldsymbol{\mu}^{(i)})$  is nonsingular and  $\frac{\partial F_i^*}{\partial \lambda_i} := \frac{\partial F_i}{\partial \lambda_i}(\boldsymbol{\mu}^{(i)}) \neq 0$ ; then  $\mathbf{F}_i^* := \mathbf{F}_i(\boldsymbol{\mu}^{(i)}) = \mathbf{0}$ .*

*Proof.* Evaluating the second equation in (14) at the fixed point, we obtain

$$\mu_i^{(i)} = \mu_i^{(i)} - \left( \frac{\partial F_i^*}{\partial \lambda_i} \right)^{-1} F_i^*,$$

which means that  $F_i^* := F_i(\boldsymbol{\mu}^{(i)}) = 0$ . This also implies that  $H_{i,2}^* = \mu_i^{(i)}$ , where  $H_{i,2}^*$  denotes the second component of (14) evaluated at the fixed point. Subsequently,

$$\mathbf{g}_i^* := \mathbf{g}_i(\boldsymbol{\mu}_{i-1}^{(i)}, \mu_i^{(i)}) = \boldsymbol{\mu}_{i-1}^{(i)}.$$

Substituting  $H_{i,2}^* = \mu_i^{(i)}$  and  $\mathbf{g}_i^* = \boldsymbol{\mu}_{i-1}^{(i)}$  into  $\boldsymbol{\mu}_{i-1}^{(i)} = \mathbf{H}_{i,1}^*$ , where  $\mathbf{H}_{i,1}^*$  denotes the first equation in (14) evaluated at the fixed point  $\boldsymbol{\mu}^{(i)}$ , we immediately obtain  $\mathbf{F}_{i-1}^* := \mathbf{F}_{i-1}(\boldsymbol{\mu}^{(i)}) = \mathbf{0}$ .  $\square$

This theorem says that the fixed points of (13) are indeed the solutions of

$$\mathbf{F}_i(\lambda_{i-1}, \lambda_i, \alpha_{i+1}, \dots, \alpha_n) = \mathbf{0},$$

which is what we intend to solve on each iteration  $i = 2, \dots, n$ . Next, we will establish the condition for the fixed point to be locally attracting. This condition will ensure that if we iterate the map in (14) with an initial condition that is close to the solution, then we will obtain the solution.

For local convergence, we want to show that eigenvalues of the Jacobian matrix  $D\mathbf{H}_i^* := D\mathbf{H}_i(\boldsymbol{\mu}^{(i)})$  are in the interior of the unit ball of the complex plane. One can verify that the components of the Jacobian matrix  $D\mathbf{H}_i^*$  are given by

$$\frac{\partial H_{i,1}^*}{\partial \lambda_j} = -(\mathbf{F}_{i-1, \lambda_{i-1}}^*)^{-1} \mathbf{F}_{i-1, \lambda_i}^* \frac{\partial H_{i,2}^*}{\partial \lambda_j}, \quad (16)$$

$$\frac{\partial H_{i,2}^*}{\partial \lambda_j} = \delta_{j,i} - \left( \frac{\partial F_i^*}{\partial \lambda_i} \right)^{-1} \frac{\partial F_i^*}{\partial \lambda_j}, \quad (17)$$

for  $j = 1, \dots, i$ , where we have used all three conditions in Assumption 1 (see the Appendix for the detailed derivation). Here,  $\delta_{j,i}$  is one only if  $j = i$  and zero

otherwise. To simplify the discussion below, let's define the notations

$$\begin{aligned} J &:= \mathbf{F}_{i-1, \lambda_{i-1}}^*, \\ \mathbf{v} &:= \mathbf{F}_{i-1, \lambda_i}^*, \\ \mathbf{c} &:= \left( \frac{\partial H_{i,2}^*}{\partial \lambda_1}, \dots, \frac{\partial H_{i,2}^*}{\partial \lambda_{i-1}} \right)^\top \end{aligned} \quad (18)$$

such that

$$D\mathbf{H}_{i+1}^* = \begin{pmatrix} J^{-1} \mathbf{v} \mathbf{c}^\top & \mathbf{0} \\ \mathbf{c}^\top & 0 \end{pmatrix} \in \mathbb{R}^{i \times i}. \quad (19)$$

We can now obtain the following result.

**Theorem 4.2.** *Let  $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^i$  be a fixed point of (13) such that the conditions in Assumption 1 are satisfied. Let  $\sigma_j(\mathbf{F}_{i-1, \lambda_{i-1}}^*)$  be the eigenvalues of  $\mathbf{F}_{i-1, \lambda_{i-1}}^*$ , and assume that they satisfy the order  $|\sigma_1| \geq |\sigma_2| \geq \dots \geq |\sigma_{i-1}|$ . If*

$$\left| \left( \frac{\partial F_i^*}{\partial \lambda_i} \right)^{-1} \sum_{j=1}^{i-1} \frac{\partial F_j^*}{\partial \lambda_i} \frac{\partial F_i^*}{\partial \lambda_j} \right| < |\sigma_{i-1}(\mathbf{F}_{i-1, \lambda_{i-1}}^*)|, \quad (20)$$

then  $\boldsymbol{\mu}^{(i)}$  is locally attracting.

*Proof.* From (19), we only need to analyze the eigenvalues of  $J^{-1} \mathbf{v} \mathbf{c}^\top$ . From basic matrix theory, recall that the magnitude of the largest eigenvalue can be bounded above as

$$|\sigma_1(J^{-1} \mathbf{v} \mathbf{c}^\top)| = \|J^{-1} \mathbf{v} \mathbf{c}^\top\|_2 \leq \|J^{-1}\|_2 \|\mathbf{v} \mathbf{c}^\top\|_2,$$

where  $\|\cdot\|_2$  denotes the matrix  $\ell_2$ -norm. For the fixed point to be locally attracting, all of the eigenvalues of  $J^{-1} \mathbf{v} \mathbf{c}^\top$  have to be in the interior of the unit ball in the complex plane. This means that we only need to show that  $\|J^{-1}\|_2 \|\mathbf{v} \mathbf{c}^\top\|_2 < 1$  or  $\|\mathbf{v} \mathbf{c}^\top\|_2 < |\sigma_{i-1}(J)|$ , where  $\sigma_{i-1}(J)$  denotes the smallest eigenvalue of the  $(i-1) \times (i-1)$  matrix  $J$  following the ordering in the hypothesis.

Since  $\text{Tr}(\mathbf{v} \mathbf{c}^\top) = \sum_{j=1}^i \sigma_j(\mathbf{v} \mathbf{c}^\top)$  and  $\mathbf{v} \mathbf{c}^\top$  is a rank-one matrix, then its nontrivial eigenvalue is given by

$$\sigma(\mathbf{v} \mathbf{c}^\top) = \text{Tr}(\mathbf{v} \mathbf{c}^\top) = \sum_{j=1}^{i-1} \frac{\partial F_j^*}{\partial \lambda_i} \frac{\partial H_{i,2}^*}{\partial \lambda_j} = - \sum_{j=1}^{i-1} \frac{\partial F_j^*}{\partial \lambda_i} \frac{\partial F_i^*}{\partial \lambda_j} \left( \frac{\partial F_i^*}{\partial \lambda_i} \right)^{-1},$$

where we have used the definitions in (18) and the second component in (17). From the assumption in (20), we have

$$\|\mathbf{v} \mathbf{c}^\top\|_2 = |\sigma(\mathbf{v} \mathbf{c}^\top)| = \left| \left( \frac{\partial F_i^*}{\partial \lambda_i} \right)^{-1} \sum_{j=1}^{i-1} \frac{\partial F_j^*}{\partial \lambda_i} \frac{\partial F_i^*}{\partial \lambda_j} \right| < |\sigma_{i-1}(J)|. \quad \square$$

This theorem provides the conditions for local convergence on each iteration  $i$ . In particular, if the hypothesis in Theorem 4.2 is satisfied, we will find the solutions to (8) by iterating (13) provided that we start with a sufficiently close initial condition. Notice also that this condition suggests that in practice the local convergence will be difficult to satisfy if the Jacobian matrix  $F_{i-1, \lambda_{i-1}}$  is close to singular. With these two theorems, we can now establish:

**Theorem 4.3.** *Let  $\mu^{(n)} \in \mathbb{R}^n$  be the solution of the  $n$ -dimensional system of equations in (7). We assume the hypothesis in Theorem 4.2; then the EBE method is locally convergent.*

*Proof.* Choose an initial condition  $(\alpha_1, \dots, \alpha_n)$  that is sufficiently close to the solution  $\mu^{(n)}$  of  $F_n(\lambda_n) = \mathbf{0}$ . First, let us define the surface  $F_1(\lambda_1, \dots, \lambda_n) = 0$  as  $\mathcal{M}_n$ ; here, the dimension of  $\mathcal{M}_n$  is at most  $n - 1$ . Subsequently, we define the surfaces  $F_2(\lambda_n) = \mathbf{0}$  as  $\mathcal{M}_{n-1}$ ,  $F_3(\lambda_n) = \mathbf{0}$  as  $\mathcal{M}_{n-2}$ , and so on. The dimension of  $\mathcal{M}_j$  is at most  $j - 1$ . We assume that  $F_n(\lambda_n) = \mathbf{0}$  has at least one solution; then  $\mathcal{M}_1$  contains the solution  $\mu^{(n)}$ . It is clear that  $\mathcal{M}_n \supset \mathcal{M}_{n-1} \supset \dots \supset \mathcal{M}_1$ .

For  $i = 1$ , we solve  $F_1(\lambda_1, \alpha_2, \dots, \alpha_n) = 0$  for  $\lambda_1$ . Geometrically, we look for the first coordinate on the surface  $\mathcal{M}_n$ . From Assumption 1(2), we have the local convergence of the usual Newton’s iteration. If  $\alpha_1$  is sufficiently close to the solution  $\mu^{(1)} = \mu_1^{(1)} \in \mathbb{R}$ , as  $m \rightarrow \infty$  we obtain the solution  $(\mu_1^{(1)}, \alpha_2, \dots, \alpha_n) \in \mathcal{M}_n$ . By the smoothness assumption,  $(\mu_1^{(1)}, \alpha_2, \dots, \alpha_n)$  is also close to  $\mu^{(n)}$ .

Continuing with  $i > 1$ , we want to solve  $F_i(\lambda_i, \alpha_{i+1}, \dots, \alpha_n) = \mathbf{0}$  for  $\lambda_i$ . Numerically, we will apply the iterative map  $H_i$  in (13) starting from  $(\mu^{(i-1)}, \alpha_i, \dots, \alpha_n) \in \mathcal{M}_{n-i+2}$ . By Assumption 1(2), the Jacobian  $F_{i-1, \lambda_{i-1}}(\mu^{(i-1)}, \alpha_i, \dots, \alpha_n)$  is non-singular, so by the implicit function theorem, for any local neighborhood  $V$  of  $\mu^{(i-1)}$ , there exists a neighborhood  $U$  of  $\alpha_i$  and a  $C^1$  function  $h_{i-1} : U \rightarrow V$  such that  $\mu^{(i-1)} = h_{i-1}(\alpha_i)$  and  $F_{i-1}(h_{i-1}(\lambda_i), \lambda_i, \alpha_{i+1}, \dots, \alpha_n) = 0$  for all  $\lambda_i \in U$ . Since the initial condition  $\alpha_i$  is close to  $\mu_i^{(n)}$ , by the smoothness assumption it is also close to  $\mu_i^{(i)}$  that solves  $F_i(\lambda_i, \alpha_{i+1}, \dots, \alpha_n) = 0$ . The continuity of  $h_{i-1}$  on  $U$  means that  $(\mu_{i-1}^{(i)}, \mu_i^{(i)}) \in V \times U$ . Geometrically, this means the surface  $F_i(\lambda_i, \alpha_{i+1}, \dots, \alpha_n) = 0$  intersects with the curve  $\lambda_{i-1} = h_{i-1}(\lambda_i)$  at  $\mu^{(i)} = (\mu_{i-1}^{(i)}, \mu_i^{(i)})$ . Therefore, we can find the solution for this  $i$ -dimensional system by tracking along the curve  $\lambda_{i-1} = h_{i-1}(\lambda_i)$  where we consider  $\lambda_i$  as an independent parameter. The iterative map  $H_i$  in (14) is to facilitate this tracking, and the conditions in Theorem 4.2 guarantee convergence to the solution. Notice that during this iteration, the solution remains on  $\mathcal{M}_{n-i+2}$ . The solution for this  $i$ -dimensional problem is  $(\mu^{(i)}, \alpha_{i+1}, \dots, \alpha_n) \in \mathcal{M}_{n-(i+1)+2} \subset \mathcal{M}_{n-i+2} \subset \dots \subset \mathcal{M}_n$ . Continuing with the same argument, we find that for  $i = n$ ,  $\mu^{(n)} \in \mathcal{M}_1 \subset \mathcal{M}_n$ .  $\square$

This iterative procedure finds the solution by searching along the manifold  $\mathcal{M}_n$  in the direction of the hypersurfaces of a single parameter at a time, whose

local existence is guaranteed by Assumption 1. It is clear that after each step  $i$ , the estimated solution may not necessarily be closer to the true solution since the estimates do not minimize the closest path to the true solution along the manifold  $\mathcal{M}_n$  (or the geodesic distance). This means that, locally,

$$\|(\boldsymbol{\mu}^{(i+1)}, \alpha_{i+2}, \dots, \alpha_n) - \boldsymbol{\mu}^{(n)}\| \leq \|(\boldsymbol{\mu}^{(i)}, \alpha_{i+1}, \dots, \alpha_n) - \boldsymbol{\mu}^{(n)}\|$$

for  $i < n - 1$  is not true.

In practice, when initial conditions are not close to the solution, the (global) convergence of EBE requires the additional condition that, for every  $i$ , there exists a nonempty connected set that contains  $(\boldsymbol{\mu}^{(i)}, \alpha_{i+1})$  and  $\boldsymbol{\mu}^{(i+1)}$  such that  $\mathbf{F}_{i,\lambda_i}$  evaluated at any point in this set is nonsingular. The existence of this set will allow us to build a path to connect these two points that are far apart. If this condition is not met, we need an additional treatment to overcome this issue which will be discussed in the next section.

### 5. Practical challenges

In this section, we will discuss several practical challenges related to our algorithm with remedies. They include nonlocality of the initial condition, mistracking due to multiple solutions, nonexistence of solutions within the desired numerical tolerance, and the computational complexity.

**Adaptive tracking.** As we mentioned in the previous section, the EBE method only converges locally, which means that it requires an adequate initial condition which is practically challenging. In our numerical simulations below, in fact, we always start from zero initial condition,  $\alpha_i = 0$  for all  $i = 1, \dots, n$ . In this case, notice that even when we obtain an accurate solution at step  $i$ , that is,  $\mathbf{F}_i(\hat{\boldsymbol{\mu}}^{(i)}) \approx \mathbf{0}$ , as we proceed to the next iteration,  $|F_{i+1}(\hat{\boldsymbol{\mu}}^{(i)}, \alpha_{i+1})| \gg 0$ , meaning that  $(\hat{\boldsymbol{\mu}}^{(i)}, \alpha_{i+1})$  is not close to the solution  $\boldsymbol{\mu}^{(i+1)}$ . Even when  $\frac{\partial F_{i+1}}{\partial \lambda_{i+1}}(\hat{\boldsymbol{\mu}}^{(i)}, \alpha_{i+1})$  is not singular, according to (9),  $\lambda_i^{m+1}$  could be very far away from  $\lambda_i^m$ . In this case, Newton’s method could fail in (12) because the initial guess could be very far from the solution.

As a remedy, we employ an adaptive tracking on  $\lambda_i$  to guarantee that the application of Newton’s method is within its zone of convergence for each predictor-corrector step. The idea of the adaptive tracking is that we cut the tracking step,  $\Delta\lambda_i := \lambda_{i+1} - \lambda_i$ , by half until the prediction-correction step in (11)–(12) converges. The algorithm is outlined in Algorithm 1.

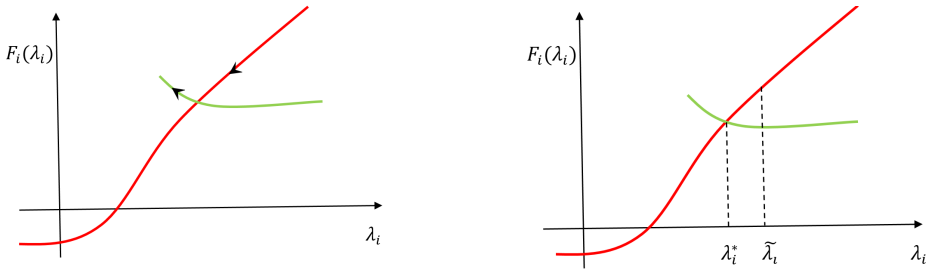
**Bifurcation.** In order to solve  $F_i(\lambda_1, \lambda_2, \dots, \lambda_i) = 0$ , we track  $\mathbf{F}_{i-1}(\boldsymbol{\lambda}_{i-1}, \lambda_i) = \mathbf{0}$  along  $\lambda_i$  as a parameter. During this parameter tracking, we may have some bifurcation points of  $\lambda_i$  for the nonlinear system  $\mathbf{F}_{i-1}(\boldsymbol{\lambda}_{i-1}, \lambda_i) = \mathbf{0}$ . This means that the Jacobian  $\mathbf{F}_{i-1,\lambda_{i-1}}(\boldsymbol{\lambda}_{i-1}, \lambda_i)$  is rank deficient such that  $\mathbf{F}_{i-1}(\boldsymbol{\lambda}_{i-1}, \lambda_i) = \mathbf{0}$  has

---

**Input** minimum step size  $\lambda_{\min}$  and threshold value of Tol.  
 Compute  $\Delta\lambda_i$  by using Newton's method to solve  $F_i = 0$ .  
 Set Final =  $\Delta\lambda_i$ .  
**while** |Final| > 0 **do**  
   Solve  $F_{i-1}(\lambda_{i-1}, \lambda_i + \Delta\lambda_i) = 0$  by using Newton's method.  
   **if** Newton's method fails **then**  
      $\Delta\lambda_i = \Delta\lambda_i/2$   
     **if**  $\Delta\lambda_i < \lambda_{\min}$  **then**  
       Discard the  $i$ -th equation.  
     **end**  
   **else**  
     Final = Final -  $\Delta\lambda_i$   
      $\Delta\lambda_i = \min\{\Delta\lambda_i, \text{Final}\}$   
   **end**  
**end**

---

**Algorithm 1.** Summary of adaptive tracking algorithm.



**Figure 1.** Plot of  $F_i(\lambda_i)$  versus  $\lambda_i$ . There are two bifurcation branches for the nonlinear system  $F_{i-1}(\lambda_{i-1}, \lambda_i) = 0$ . The left part is a mistracking example; the right part is the illustration of a numerical method to avoid the bifurcation point.

multiple solutions  $\lambda_{i-1}$  for a given  $\lambda_i$ . In this situation,  $F_i$  has multiple realization functions of  $\lambda_i$ . See the illustration in Figure 1 where the bifurcation point is the intersection of the two possible realizations of  $F_i$ . In this illustration, the goal is to track along the red branch to find the root,  $F_i(\lambda_i) = 0$ . As we get closer to the bifurcation point, the Jacobian  $F_{i-1, \lambda_{i-1}}(\lambda_{i-1}, \lambda_i)$  is singular such that we can't evaluate (11). Intuitively, the existence of multiple solutions near the bifurcation point induces a possibility of mistracking from the red curve to the green curve (as shown by the arrows), which prohibits one from finding the solution.

To avoid such mistracking, we apply the deflation technique to compute the bifurcation point directly [12; 16]. Once the bifurcation point is estimated, we approximate the correct branches using Richardson extrapolation to avoid mistracking. Denoting the bifurcation point as  $\lambda_i^*$ , the nonlinear system  $F_{i-1}(\lambda_{i-1}, \lambda_i) = 0$



is difficult to solve when  $\lambda_i$  is close to  $\lambda_i^*$  since the Jacobian of  $F_{i-1}(\lambda_{i-1}, \lambda_i)$  becomes near singular. If the last attempt is  $(\tilde{\lambda}_{i-1}, \tilde{\lambda}_i)$ , we compute  $(\lambda_{i-1}^*, \lambda_i^*)$  by solving the deflated system

$$G(\lambda_{i-1}^*, \lambda_i^*, \mathbf{v}) = \begin{pmatrix} F_{i-1}(\lambda_{i-1}, \lambda_i) \\ F_{i-1, \lambda_{i-1}}(\lambda_{i-1}, \lambda_i) \mathbf{v} \\ \boldsymbol{\xi}^\top \mathbf{v} - 1 \end{pmatrix} = \mathbf{0},$$

where  $\mathbf{v}$  is the kernel of  $F_{i-1, \lambda_{i-1}}(\lambda_{i-1}, \lambda_i)$  and  $\boldsymbol{\xi}$  is a random vector to guarantee that  $\mathbf{v}$  is not a zero eigenvector. In this case,  $G(\lambda_{i-1}^*, \lambda_i^*, \mathbf{v})$  is well conditioned [12; 16]. Once the bifurcation point  $(\lambda_{i-1}^*, \lambda_i^*)$  is estimated, we can avoid mistracking by setting  $\lambda_i = 2\lambda_i^* - \tilde{\lambda}_i$  and solve  $F_{i-1}(\lambda_{i-1}, \lambda_i) = \mathbf{0}$  by using Newton's method with an initial guess  $2\lambda_{i-1}^* - \tilde{\lambda}_{i-1}$  (which is a Richardson extrapolation).

**Nonexistence of solutions.** In general, the moment constrained maximum entropy problems may not necessarily have solutions. Even when the solutions exist theoretically, they could be difficult to find numerically due to the noisy dataset, error in the numerical integration, etc. In this case, we simply discard the equation  $F_i$  when the minimum is larger than the desired tolerance. This feature (discarding the constraints that give no solutions) is only feasible in the EBE algorithm. However, some theories are needed to preserve the convexity of the polynomials in the exponential term of (4) while discarding some of these constraints. In our numerical simulations below, we handle this issue by reordering the constraints. In particular, for a problem with moment constraints up to order four, we include the constraints corresponding to  $\mathbb{E}[x_i^4]$ ,  $i = 1, \dots, d$ , in the earlier step of the EBE iterations to avoid these constraints being discarded. Note that this method is sensitive to ordering, that is, different ordering of constraints yields different paths to compute the solution. Therefore, a systematic ordering technique that simultaneously preserves the convexity of the polynomial in the exponential term of (4) is an important problem to be addressed in the future.

**Computational complexity.** The most expensive computational part in EBE is the numerical evaluation of (6). For a fast numerical integration, we store the monomial basis  $c_j(\mathbf{x})$  as a matrix of size  $N_\ell \times n$ , where  $N_\ell$  is the number of sparse grid points and  $n$  is number of monomial basis. In this case, the computational cost in evaluating  $F_j$  is  $(2j+1)N_\ell$  ( $j-1$  additions,  $j+1$  multiplications, and 1 subtraction for each grid point), excluding the computational cost for exponential function evaluation, which is on the order of  $\log^2 m$  to obtain an error of resolution  $2^{-m}$  [6]. For the  $i$ -th iteration of the EBE algorithm, the computational cost to evaluate the  $i$ -dimensional system  $F_i$  is  $\sum_{j=1}^i (2j+1)N_\ell = \frac{1}{2}(i^2 + i)N_\ell$ , excluding the exponentiation.

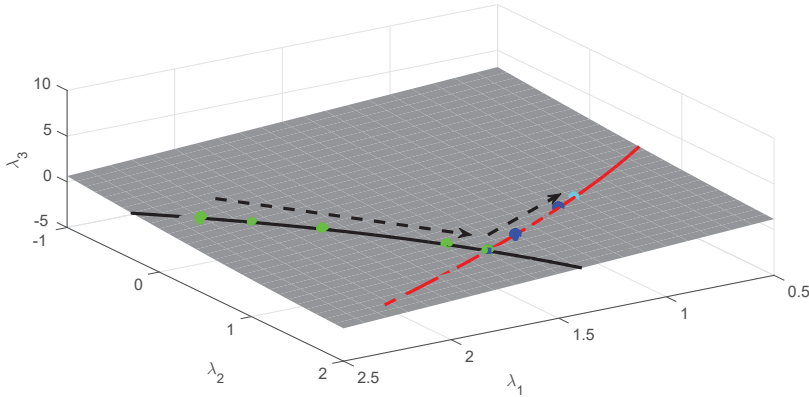
## 6. Numerical results

In this section, we show numerical results of the EBE method on five examples. In all of the simulations below, unless stated, we set the Newton's tolerance  $\text{Tol}_1 = 10^{-1}$  and the predictor tolerance  $\text{Tol}_2 = 10^{-10}$ . In the first test example, we will describe how the EBE method works on each iteration. The goal of the second example is to demonstrate the global convergence with solutions that are far away from initial condition,  $\alpha_j = 0$ . In particular, we will test the EBE method on a problem with solutions,  $\lambda_j$ , that have magnitudes ranging from orders  $10^0$ – $10^3$ . In this example, we will show the robustness of the estimate as a function of the number of integration points (or the sparse grid level  $\ell$ ). The third example demonstrates the performance on high-dimensional problems (with  $70 \leq n \leq 310$  of order one hundred), induced from order-four moments of four- to seven-dimensional density functions. While these first three examples involve estimating densities of the form (4), in the next two examples, we also test the EBE method to estimate densities from a given data set where the maximum entropy solutions may or may not exist. The first data-driven problem is to estimate densities of the first two leading EOFs of the wind stress-driven large-scale oceanic model [3; 4]. The second data-driven problem is to estimate two- to five-dimensional densities arising from solutions of the Kuramoto–Sivashinsky equation. In these two problems, we compare our method with the classical Newton's method, the MATLAB built-in solver `fsolve`, and the previously developed BFGS-based method [3; 4].

**Example 1.** We consider a simple example  $\rho(x) \propto \exp(x + x^2 + x^3)$  for  $x \in [-1, 1]$  so that the exact solution is  $\lambda = (1, 1, 1)$ . Here, the moments  $f_j$  can be computed numerically by

$$f_j = \frac{\int_{-1}^1 x^j \rho(x) dx}{\int_{-1}^1 \rho(x) dx} \quad \text{for } i = 1, 2, 3.$$

In order to numerically integrate both the denominator and numerator, we used a regular one-dimensional sparse grid of level  $\ell = 7$  (the number of nodes is 65). Our goal here is to illustrate the method and to show the trajectory of the solutions after each iteration of the inner loop  $m$  and outer loop  $i$ . In Figure 2, we show the surface of  $F_1(\lambda_1, \lambda_2, \lambda_3) = 0$  (gray). For  $i = 1$ , we solve  $F_1(\lambda_1, 0, 0) = 0$ ; after three iterations ( $m = 3$ ) the solution converges to  $\lambda_1 = 2.3$  (see Table 1). For  $i = 2$ , we start with this solution and introduce the second variable  $\lambda_2$  for solving the second equation  $F_2(\lambda_1, \lambda_2, 0) = 0$  with constraint  $F_1(\lambda_1, \lambda_2, 0) = 0$ . Here, the solution follows the path  $\lambda_1 = h_1(\lambda_2)$  thanks to the implicit function theorem (black curve). Numerically, a sequence of (green) points following this path converges to a point that satisfies  $F_1(\lambda_1, \lambda_2, 0) = F_2(\lambda_1, \lambda_2, 0) = 0$  (the green point in the intersection between black and red curves in Figure 2). In the next iteration  $i = 3$ , we



**Figure 2.** The illustration of Example 1. The black curve is  $\lambda_1 = h_1(\lambda_2)$ , the green points are the iterations when we solved  $F_1(\lambda_1, \lambda_2, 0) = 0$ , the red curve is  $(\lambda_1, \lambda_2) = h_2(\lambda_3)$ , the blue points are the iterations when we solved  $F_1(\lambda_1, \lambda_2, \lambda_3) = F_2(\lambda_1, \lambda_2, \lambda_3) = 0$ , and the cyan point is the numerical solution.

$m \downarrow i \rightarrow$	1	2	3
0	(0, 0, 0)	(2.30, 0, 0)	(1.58, 1.43, 0)
1	(1.76, 0, 0)	(2.23, 0.22, 0)	(1.52, 1.38, 0.26)
2	(2.23, 0, 0)	(1.87, 0.57, 0)	(1.12, 1.09, 0.76)
3	(2.30, 0, 0)	(1.67, 1.21, 0)	(1, 1, 1)
4		(1.58, 1.43, 0)	

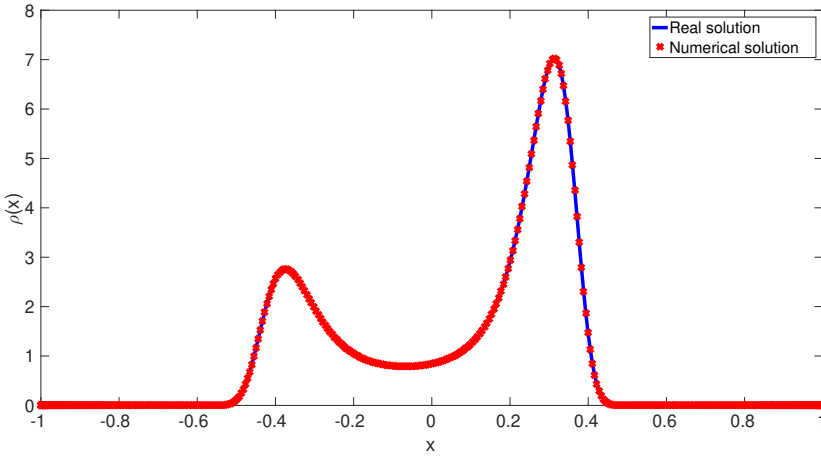
**Table 1.** The coordinate of the solutions of Example 1 for each iteration, starting from (0, 0, 0). For each outer loop  $i$ , the EBE takes few iterates ( $m$ ) to find the  $i$ -dimensional solution, fixing  $\lambda_j = \alpha_j = 0$  for  $j > i$ .

introduce the third variable  $\lambda_3$  for solving the third equation  $F_3(\lambda_1, \lambda_2, \lambda_3) = 0$  with constraints  $F_1(\lambda_1, \lambda_2, \lambda_3) = F_2(\lambda_1, \lambda_2, \lambda_3) = 0$ . By the implicit function theorem, we have  $(\lambda_1, \lambda_2) = h_2(\lambda_3)$  that satisfies  $F_1(h_2(\lambda_3), \lambda_3) = F_2(h_2(\lambda_3), \lambda_3) = 0$ , which is shown by the red curve in Figure 2. On this red curve, we have a sequence of (blue) points which converges to the solution of the full system (cyan point shown in Figure 2). The coordinate of the solution on each iteration is shown in Table 1. Notice that the solutions always lie on the surface  $F_1(\lambda_1, \lambda_2, \lambda_3) = 0$ .

**Example 2.** We consider a one-dimensional example with up to order-six moment constraints with explicit solution given by

$$\rho(x) \propto \exp(2x + 16x^2 + 24x^3 + 96x^4 - 256x^5 - 1024x^6),$$

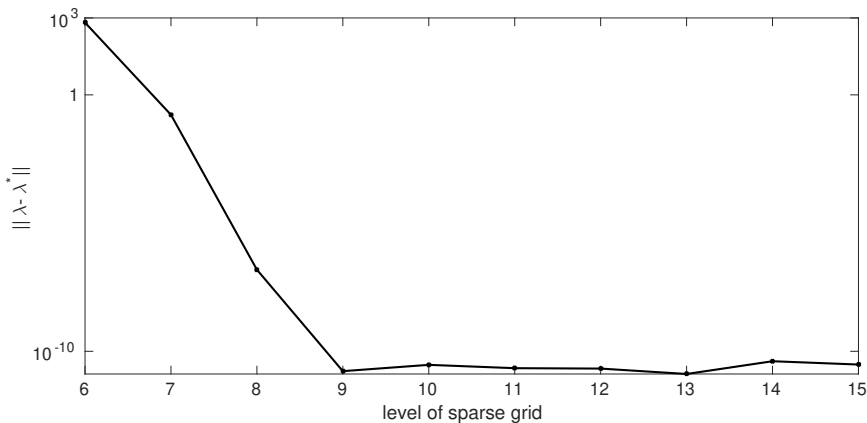
as shown in Figure 3. This example is a tough test problem since the solution,  $\lambda = (2, 16, 24, 96, -256, 1024)$ , has components of order  $10^0$ – $10^3$ . Following



**Figure 3.** The unnormalized density  $\rho(x)$  in Example 2.

Example 1, we compute the moments  $f_i$  by using a one-dimensional sparse grid of level  $\ell = 7$  (65 nodes). The EBE algorithm converges to the exact solution with error  $\|\lambda - \lambda^*\| = 5.44 \times 10^{-13}$ . Since the numerical experiment is performed with an initial condition  $\alpha_j = 0$  that is far from the solution, this result demonstrates a global convergence of the EBE method.

Next, we investigate the sensitivity of the estimates to the number of sparse grid points used in approximating the integral. In our numerical experiments, we estimate the true moments  $f_i$  using a one-dimensional sparse grid of level  $\ell = 20$  (524 289 nodes) and feed these moment estimates into the EBE algorithm. In Figure 4, we show the error in  $\lambda$  (with  $\ell_2$ -metric) for different levels of the sparse grid from 6 to 15 that are used in the EBE method. Notice that the error decreases as a function of  $\ell$  and the improvement becomes negligible for  $\ell > 8$ .



**Figure 4.** The solution error as a function of the number of sparse grid.

Methods	Order		
	4	6	8
BFGS algorithm with uniform grid	$4.07 \times 10^{-2}$	$1.45 \times 10^{-4}$	$1.14 \times 10^{-2}$
EBE algorithm with uniform grid	$1.27 \times 10^{-11}$	$9.84 \times 10^{-15}$	$7.75 \times 10^{-13}$
EBE algorithm with sparse grid	$7.54 \times 10^{-12}$	$8.12 \times 10^{-15}$	$2.43 \times 10^{-13}$
MATLAB fsolve with sparse grid	$4.70 \times 10^{-7}$	$1.19 \times 10^{-4}$	$1.74 \times 10^{-4}$
Newton with sparse grid	$5.12 \times 10^{-11}$	divergence	divergence

**Table 2.** Summary of solutions for Example 4: moment errors for different algorithms with different grids.

**Example 3.** In this example, we consider a  $d$ -dimensional example with an explicit solution,

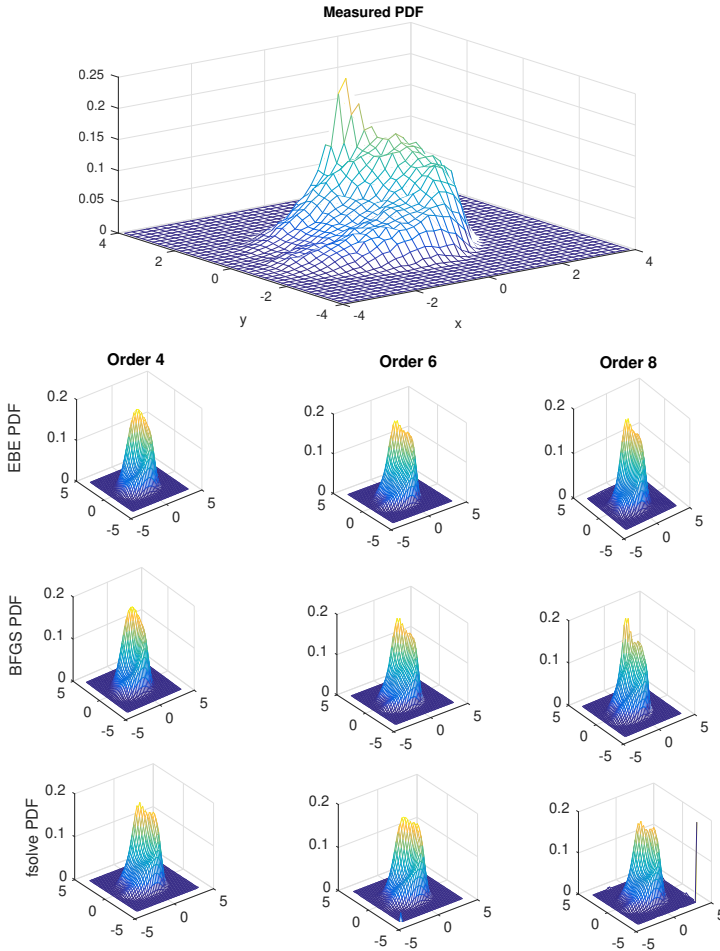
$$\rho(x) \propto \exp(-2x_1^4 + x_2^3 - x_2^4 - x_3^4 - 1.8x_4^4),$$

on domain  $\Omega = [-1, 1]^d$  where we will vary  $d = 4, \dots, 7$ . For these simulations, we consider up to order-four moment constraints and fix the sparse grid level  $\ell = 8$  to compute the integration.

Here, the EBE method is able to estimate  $\lambda$  with  $\ell_2$ -errors of order  $10^{-13}$  (the error in  $\lambda$  is  $1.11 \times 10^{-13}$  and moments error is  $3.15 \times 10^{-15}$ ). In this computation, the dimensions of the nonlinear system are 70 for  $d = 4$ , 126 for  $d = 5$ , 210 for  $d = 6$ , and 310 for  $d = 7$ . Here, the EBE method is able to recover the true density even if we prescribe more constraints, corresponding to  $d$  larger than four.

**Example 4.** Next, we consider estimating a two-dimensional probability density of the two leading empirical orthogonal functions of a geophysical model for a wind stress-driven large-scale oceanic model [18; 19]. This is exactly the same test example as in the previously developed BFGS-based method [3; 4]. In fact, the two-dimensional density that we used here was supplied by Rafail Abramov. First, we compare the EBE method with the BFGS algorithm of [3], whose code can be downloaded from [2]. In this comparison, we use the same uniformly distributed grid points where the total number of nodes is  $85 \times 85 = 7\,225$ . We set the Newton’s tolerance of the EBE algorithm to be  $10^{-10}$ . In Table 2 notice that the moment errors of the EBE are much smaller compared to those of the BFGS method.

While the EBE is superior compared to BFGS, we should note that the BFGS method does not use the Hessian of  $F_i$  whereas the EBE does. For a fair comparison, we include results using the MATLAB built-in function fsolve, whose default algorithm is the trust-region-dogleg (see the documentation for detail [17]). In our numerical implementation, we apply fsolve with a specified Hessian function  $F_n$ . We also include the classical Newton’s method with a specified Hessian function  $F_n$ . In this comparison, we use the same sparse grid of level  $\ell = 11$  (or 7 169 nodes) to



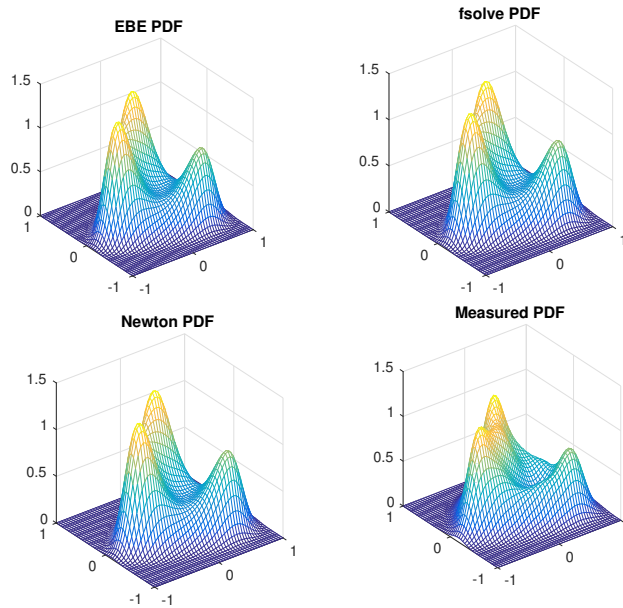
**Figure 5.** The 2D measured probability density functions supplied by R. Abramov (first row), and PDFs computed by the EBE method (second row), BFGS algorithm (third row), and the MATLAB fsolve function (fourth row).

compute the two-dimensional integral. Notice that the EBE method is still superior compared to these two schemes as reported in Table 2. In fact, Newton's method does not converge for higher-order moment constraints. The joint two-dimensional PDFs are shown in Figure 5. The first row is the two-dimensional density function provided by R. Abramov. The second row shows the EBE estimates using up to order-four, -six, and -eight moment constraints. The third and fourth rows show the BFGS and MATLAB fsolve estimates, respectively.

**Example 5.** In this example, we consider estimating multidimensional densities of the solutions of the Kuramoto–Sivashinsky equation. Here, the solutions are integrated with a fourth-order time-differencing method on 128 equally spaced grid

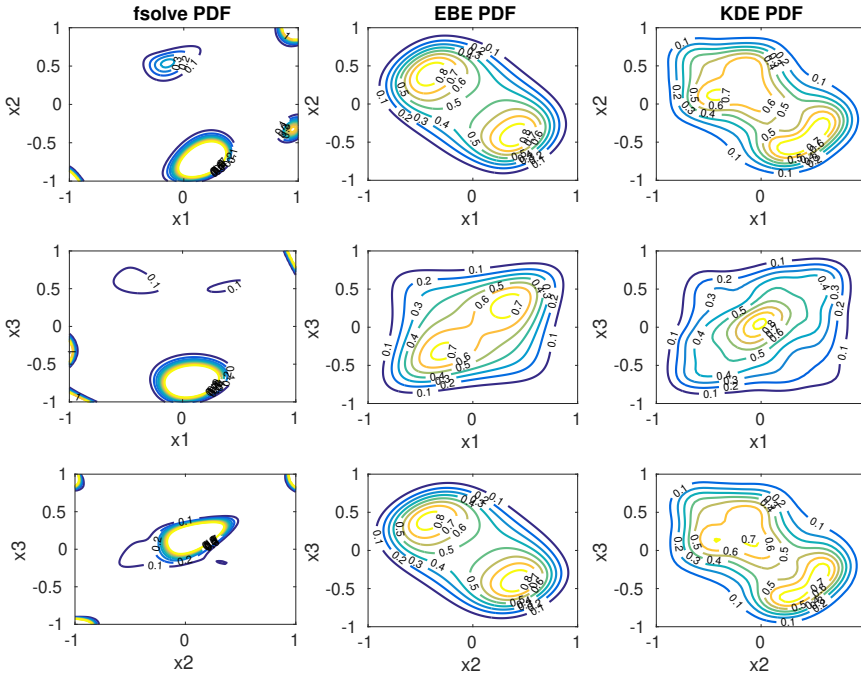
$d$	EBE method	fsolve	Newton
2	$1.098 \times 10^{-15}$	$9.779 \times 10^{-7}$	$8.128 \times 10^{-14}$
3	$4.29 \times 10^{-13}$	$3.150 \times 10^{-2}$	divergence
4	$1.19 \times 10^{-14}$	0.021	divergence
5	$2.47 \times 10^{-11}$	0.018	divergence

**Table 3.** Summary of solutions for Example 5.



**Figure 6.** The comparison of the density functions obtained by the EBE algorithm, the MATLAB fsolve function, Newton’s method, and the kernel density estimate (denoted as the measured PDF) for the two-dimensional case.

points over a domain of  $[0, 32\pi]$  as in [15]. We use initial condition  $u(x, 0) = \cos(x/(16\xi))(1 + \sin(x/16))$ , with  $\xi \sim U[0, 1]$  and integration time step 0.25. The data is generated by integrating 10 000 time steps. Based on this data set, we randomly select  $d$  components and estimate the  $d$ -dimensional joint density associated to these components. For visual comparison, we also show the results from a two-dimensional kernel density estimation method [22; 21] as a reference. Numerically, we use the MATLAB built-in function, ksdensity. Note that the BFGS algorithm [3] does not work on this data set while the classical Newton’s method only converges for the two-dimensional case. We also show the corresponding results with the MATLAB fsolve with specified Hessian function as in the previous example. The moment errors of these three schemes are reported in Table 3.

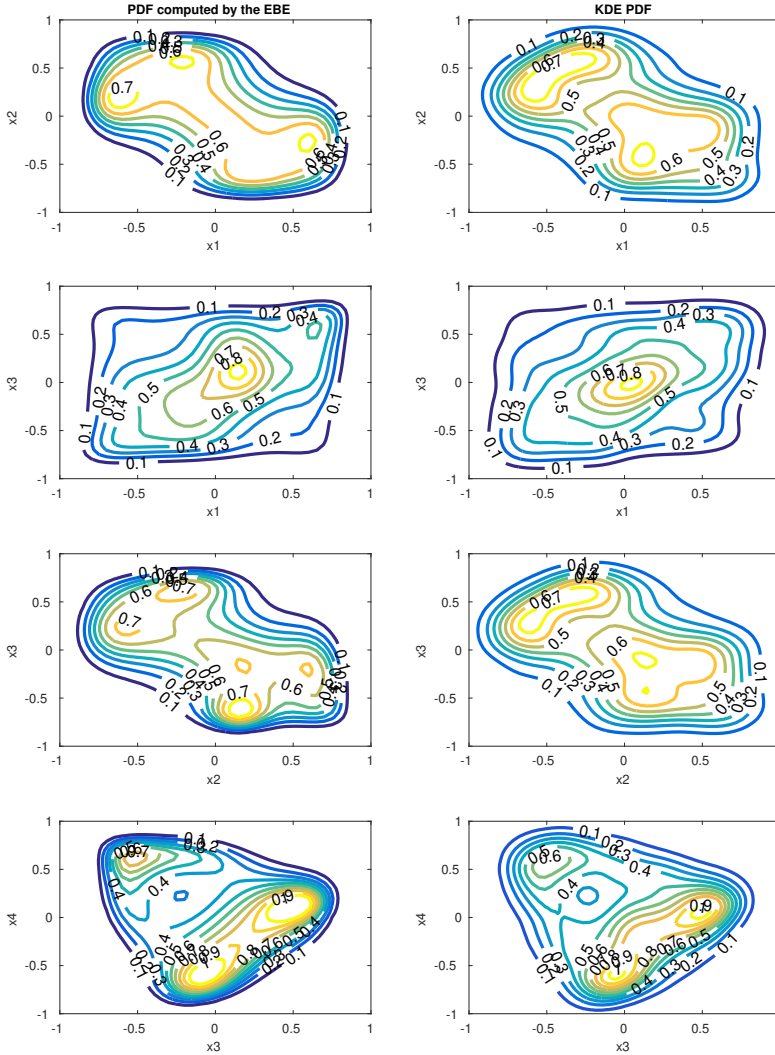


**Figure 7.** The comparison of the two-dimensional marginal density functions obtained by the MATLAB fsolve function (first column), the EBE algorithm (second column) that solves a three-dimensional problem accounting for up to order-four moment constraints, and the two-dimensional kernel density estimate (third column).

In Figure 6, we show the two-dimensional density estimated by the EBE algorithm compared to those from fsolve, the classical Newton's method, and the 2D kernel density estimate. For the two-dimensional case, the resulting densities are visually identical although the corresponding moment error of the EBE method is still the smallest compared to Newton's and the MATLAB fsolve (see Table 3). In Figure 7, we show the contour plot of the two-dimensional marginal densities obtained from solving the three-dimensional problem given four-moment constraints with the EBE method and the MATLAB fsolve. For diagnostic purposes, we also provide the corresponding contour plots of the two-dimensional kernel density estimates. Notice that the MATLAB fsolve produces a completely inaccurate estimate. The EBE method produces an estimate that qualitatively agrees to the corresponding two-dimensional KDE estimates. The slight disagreement between these estimates is expected since we only provide up to order-four moment information.

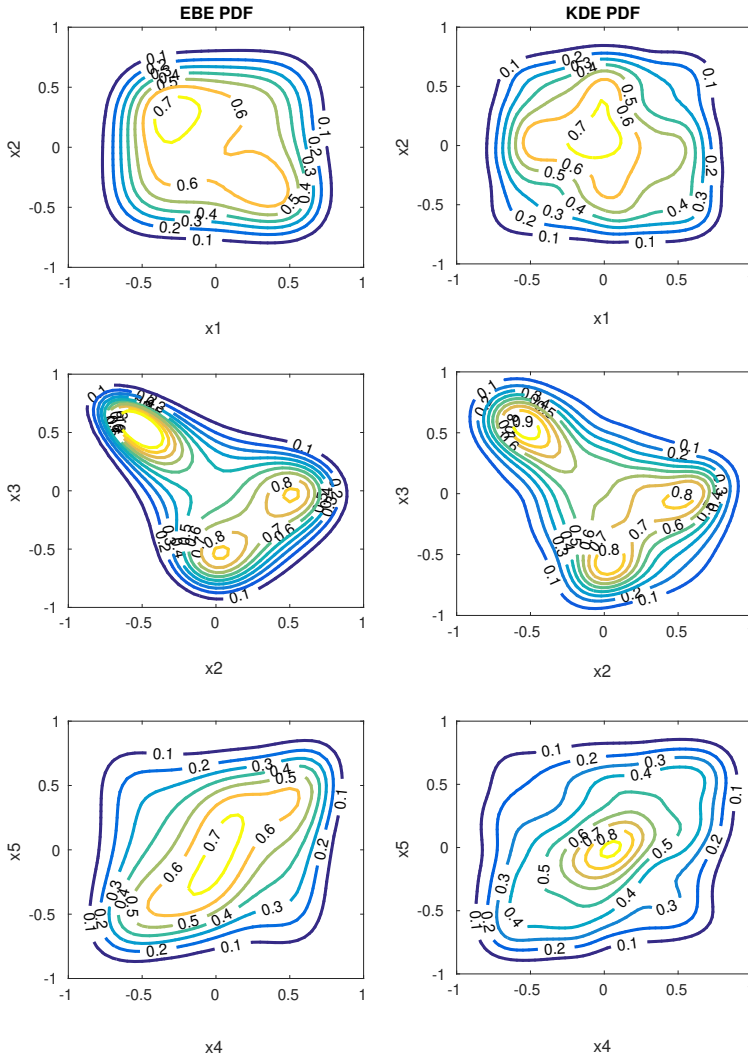
In Figure 8, we show the results for the four-dimensional problem. We do not show the estimate from the MATLAB fsolve since it is not accurate at all. Here, we include more than four-order moments. Specifically, the total number of constraints for up to order-four moments is 70 while this result is based on 87 constraints,





**Figure 8.** The comparison of the two-dimensional marginal density functions obtained by the EBE algorithm (first column) that solves a four-dimensional problem accounting for more than order-four moment constraints (see text for detail) and the two-dimensional kernel density estimate (second column).

including 17 additional higher-order moment constraints that include order-six moments,  $\mathbb{E}[x_i^6]$ ,  $i = 1, \dots, 4$ . See the movie of the density estimates for each iteration in the supplementary material [11]. Notice that the marginal densities estimated by the EBE look very similar to those estimated by the two-dimensional kernel density estimation. If more constraints are included, we found that we lose the convexity of the polynomial terms in (4). As we mentioned before, we need better criteria to preserve the convexity of the solutions.



**Figure 9.** The comparison of the two-dimensional marginal density functions obtained by the EBE algorithm (first column) that solves a five-dimensional problem accounting for the automatically selected 91 out of the prescribed 125 moments, and the two-dimensional kernel density estimate (second column).

In Figure 9, we include the result from a five-dimensional simulation. We also do not show the estimate from the MATLAB `fsolve` since it is not accurate at all. In this five-dimensional case, the EBE method automatically discards 34 equations (moment constraints). In this case, we suspect that either the maximum entropy solution that accounts for all of the constraints does not exist or the EBE method cannot find the solution. Here, the EBE method just estimates the best-fitted solution within the tolerance of  $10^{-10}$  by solving 91 out of 125 moment constraints.

## 7. Summary

In this paper, we introduced a novel equation-by-equation algorithm for solving a system of nonlinear equations arising from the moment constrained maximum entropy problem. Theoretically, we have established the local convergence and provided a sufficient condition for global convergence. Through the convergence analysis, we understood that the method, geometrically, finds the solution by searching along the surface corresponding to one component of the nonlinear equations. Numerically, we have demonstrated its accuracy and efficiency on various examples. In one of the examples, we found that the EBE algorithm produces more accurate solutions compared to the previously developed BFGS-based algorithm which does not use the Hessian information [3; 4]. In this same example, we also found that the EBE is superior compared to two schemes that use the Hessian information, including the current MATLAB built-in solver which uses the trust-region-dogleg algorithm and the classical Newton's method.

We also found that the proposed EBE algorithm is able to solve a system of 70–310 equations when the maximum entropy solution exists compared to the previously developed BFGS method which was shown to work for a system of size 44–83 equations. On the Kuramoto–Sivashinsky example, the EBE method is able to reconstruct the density of a four-dimensional problem accounting for up to order-four moments (or 70 constraints). In this case, we showed that the estimate is improved by accounting for 17 additional constraints of order-six moments. For the five-dimensional problem with moments up to order four, the EBE method reconstructs the solution within the desired precision,  $10^{-10}$ , by automatically selecting a subset of 91 constraints from the total prescribed 125 constraints induced by moments of up to order four.

While the automatic constraint selection is a desirable feature since the maximum entropy solutions within the tolerance may not be easily estimated (nor theoretically available), further study is required to fully take advantage of this feature. In particular, an important open problem is to develop a mathematical theory for ordering the constraints since the path of the solution is sensitive to the order of the constraints. Simultaneously, the ordering of the constraints needs to preserve the convexity of the polynomials in the exponential term of (4). We should stress that the EBE method is computationally not the most efficient method since it is designed to avoid singularities by tracking along the surface corresponding to one component of the nonlinear equations. Therefore, a more efficient EBE method will be one of future goals.

### Appendix: The detailed calculation of the Jacobian of the map $H_i$

In this appendix, we will give the detailed computation for the Jacobian of the map  $H_i$  in (14) evaluated at  $\mu^{(i)}$ , the solution of  $F_i(\lambda_i, \alpha_{i+1}, \dots, \alpha_n) = \mathbf{0}$ . Recall

that for  $\mathbf{H}_i = (\mathbf{H}_{i,1}, \mathbf{H}_{i,2})$  in (14),

$$\begin{aligned}\mathbf{H}_{i,1}(\boldsymbol{\lambda}_i) &= \mathbf{g}_i - \mathbf{F}_{i-1, \lambda_{i-1}}(\mathbf{g}_i, \mathbf{H}_{i,2})^{-1} \mathbf{F}_{i-1}(\mathbf{g}_i, \mathbf{H}_{i,2}), \\ \mathbf{H}_{i,2}(\boldsymbol{\lambda}_i) &= \lambda_i - \left( \frac{\partial \mathbf{F}_i}{\partial \lambda_i}(\boldsymbol{\lambda}_i) \right)^{-1} \mathbf{F}_i(\boldsymbol{\lambda}_i),\end{aligned}$$

where  $\mathbf{g}_i : \mathbb{R}^{i-1} \rightarrow \mathbb{R}^{i-1}$  is defined as in (15).

To take another derivative of  $\mathbf{H}_{i,1}$  with respect to  $\lambda_j$ , we use the fact that if  $\mathbf{F}_{i-1, \lambda_{i-1}}$  is a nonsingular matrix, then

$$\frac{\partial}{\partial \lambda_j} (\mathbf{F}_{i-1, \lambda_{i-1}})^{-1} = (\mathbf{F}_{i-1, \lambda_{i-1}})^{-1} \frac{\partial \mathbf{F}_{i-1, \lambda_{i-1}}}{\partial \lambda_j} (\mathbf{F}_{i-1, \lambda_{i-1}})^{-1},$$

and the Hessian  $\frac{\partial^2 \mathbf{F}_{i-1, \lambda_{i-1}}^*}{\partial \lambda_j^2}$  is well defined, which are Assumption 1(2)–(3). We can deduce that for  $j = 1, \dots, i$ ,

$$\begin{aligned}\frac{\partial \mathbf{H}_{i,1}}{\partial \lambda_j} &= \frac{\partial \mathbf{g}_i}{\partial \lambda_j} - (\mathbf{F}_{i-1, \lambda_{i-1}})^{-1} (\mathbf{F}_{i-1, \lambda_{i-1}})^{-1} \frac{\partial \mathbf{F}_{i-1, \lambda_{i-1}}}{\partial \lambda_j} (\mathbf{F}_{i-1, \lambda_{i-1}})^{-1} \mathbf{F}_{i-1} \\ &\quad - (\mathbf{F}_{i-1, \lambda_{i-1}})^{-1} \left( \mathbf{F}_{i-1, \lambda_{i-1}} \frac{\partial \mathbf{g}_i}{\partial \lambda_j} + \frac{\partial \mathbf{F}_{i-1}}{\partial \lambda_i} \frac{\partial \mathbf{H}_{i,2}}{\partial \lambda_j} \right),\end{aligned}\quad (21)$$

$$\frac{\partial \mathbf{H}_{i,2}}{\partial \lambda_j} = \frac{\partial \lambda_i}{\partial \lambda_j} - \frac{\partial}{\partial \lambda_j} \left( \frac{\partial \mathbf{F}_i}{\partial \lambda_i} \right)^{-1} \mathbf{F}_i - \left( \frac{\partial \mathbf{F}_i}{\partial \lambda_i} \right)^{-1} \frac{\partial \mathbf{F}_i}{\partial \lambda_j}.\quad (22)$$

Evaluating these two equations at  $\boldsymbol{\mu}^{(i)}$  and using the fact that  $\mathbf{F}_i^* := \mathbf{F}_i(\boldsymbol{\mu}^{(i)}) = \mathbf{0}$ , the second terms in the right-hand-side of (21)–(22) vanish and we have

$$\begin{aligned}\frac{\partial \mathbf{H}_{i,1}^*}{\partial \lambda_j} &= \frac{\partial \mathbf{g}_i^*}{\partial \lambda_j} - (\mathbf{F}_{i-1, \lambda_{i-1}}^*)^{-1} \left( \mathbf{F}_{i-1, \lambda_{i-1}}^* \frac{\partial \mathbf{g}_i^*}{\partial \lambda_j} + \frac{\partial \mathbf{F}_{i-1}^*}{\partial \lambda_i} \frac{\partial \mathbf{H}_{i,2}^*}{\partial \lambda_j} \right) \\ &= -(\mathbf{F}_{i-1, \lambda_{i-1}}^*)^{-1} \left( \frac{\partial \mathbf{F}_{i-1}^*}{\partial \lambda_i} \frac{\partial \mathbf{H}_{i,2}^*}{\partial \lambda_j} \right), \\ \frac{\partial \mathbf{H}_{i,2}^*}{\partial \lambda_j} &= \delta_{j,i} - \left( \frac{\partial \mathbf{F}_i^*}{\partial \lambda_i} \right)^{-1} \frac{\partial \mathbf{F}_i^*}{\partial \lambda_j}.\end{aligned}$$

where  $\delta_{j,i}$  is one only if  $j = i$  and zero otherwise.

### Acknowledgment

We thank Rafail Abramov for supplying the two-dimensional density data set for Example 4. The BFGS code that we used for comparison in Example 4 was downloaded from [2].

## References

- [1] R. V. Abramov, *An improved algorithm for the multidimensional moment-constrained maximum entropy problem*, J. Comput. Phys. **226** (2007), no. 1, 621–644. MR Zbl
- [2] ———, *The multidimensional moment-constrained maximum entropy algorithm*, 2007. Zbl
- [3] ———, *The multidimensional moment-constrained maximum entropy problem: a BFGS algorithm with constraint scaling*, J. Comput. Phys. **228** (2009), no. 1, 96–108. MR Zbl
- [4] ———, *The multidimensional maximum entropy moment problem: a review on numerical methods*, Commun. Math. Sci. **8** (2010), no. 2, 377–392. MR Zbl
- [5] R. V. Abramov, A. Majda, and R. Kleeman, *Information theory and predictability for low-frequency variability*, J. Atmospheric Sci. **62** (2005), no. 1, 65–87. MR
- [6] T. Ahrendt, *Fast computations of the exponential function*, STACS 99 (Trier) (C. Meinel and S. Tison, eds.), Lecture Notes in Comput. Sci., no. 1563, Springer, 1999, pp. 302–312. MR Zbl
- [7] D. J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wampler, *Numerically solving polynomial systems with Bertini*, Software, Environments, and Tools, no. 25, Society for Industrial and Applied Mathematics, 2013. MR Zbl
- [8] M. Frontini and A. Tagliani, *Maximum entropy in the finite Stieltjes and Hamburger moment problem*, J. Math. Phys. **35** (1994), no. 12, 6748–6756. MR Zbl
- [9] T. Gerstner and M. Griebel, *Numerical integration using sparse grids*, Numer. Algorithms **18** (1998), no. 3–4, 209–232. MR Zbl
- [10] W. Hao and J. Harlim, *Supplementary material: MATLAB software for the equation-by-equation method for solving the maximum entropy problem*, 2018.
- [11] ———, *Supplementary movie for the equation-by-equation method for solving the maximum entropy problem*, 2018.
- [12] W. Hao, J. D. Hauenstein, B. Hu, Y. Liu, A. J. Sommese, and Y.-T. Zhang, *Continuation along bifurcation branches for a tumor model with a necrotic core*, J. Sci. Comput. **53** (2012), no. 2, 395–413. MR Zbl
- [13] K. Haven, A. Majda, and R. V. Abramov, *Quantifying predictability through information theory: small sample estimation in a non-Gaussian framework*, J. Comput. Phys. **206** (2005), no. 1, 334–362. MR Zbl
- [14] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. (2) **106** (1957), 620–630. MR Zbl
- [15] A.-K. Kassam and L. N. Trefethen, *Fourth-order time-stepping for stiff PDEs*, SIAM J. Sci. Comput. **26** (2005), no. 4, 1214–1233. MR Zbl
- [16] A. Leykin, J. Verschelde, and A. Zhao, *Newton’s method with deflation for isolated singularities of polynomial systems*, Theoret. Comput. Sci. **359** (2006), no. 1–3, 111–122. MR Zbl
- [17] MathWorks, *MATLAB fsolve.m*, 2017.
- [18] J. D. McCalpin, *The statistics and sensitivity of a double-gyre model: the reduced-gravity, quasigeostrophic case*, J. Phys. Oceanogr. **25** (1995), no. 5, 806–824.
- [19] J. D. McCalpin and D. B. Haidvogel, *Phenomenology of the low-frequency variability in a reduced-gravity, quasigeostrophic double-gyre model*, J. Phys. Oceanogr. **26** (1996), no. 5, 739–752.
- [20] L. R. Mead and N. Papanicolaou, *Maximum entropy in the problem of moments*, J. Math. Phys. **25** (1984), no. 8, 2404–2417. MR

- [21] E. Parzen, *On estimation of a probability density function and mode*, Ann. Math. Statist. **33** (1962), 1065–1076. MR Zbl
- [22] M. Rosenblatt, *Remarks on some nonparametric estimates of a density function*, Ann. Math. Statist. **27** (1956), 832–837. MR Zbl
- [23] S. A. Smolyak, *Quadrature and interpolation formulas for tensor products of certain classes of functions*, Dokl. Akad. Nauk SSSR **148** (1963), no. 5, 1042–1045, In Russian; translated in Soviet Math. Dokl. **4** (1963), 240–243. MR Zbl
- [24] A. J. Sommese and C. W. Wampler, II, *The numerical solution of systems of polynomials: arising in engineering and science*, World Scientific, 2005. MR
- [25] L. N. Trefethen, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev. **50** (2008), no. 1, 67–87. MR Zbl
- [26] X. Wu, *Calculation of maximum entropy densities with application to income distribution*, J. Econometrics **115** (2003), no. 2, 347–354. MR Zbl
- [27] Z. Wu, G. N. Phillips, Jr., R. Tapia, and Y. Zhang, *A fast Newton algorithm for entropy maximization in phase determination*, SIAM Rev. **43** (2001), no. 4, 623–642. MR Zbl

Received July 4, 2017. Revised January 17, 2018.

WENRUI HAO: wxh64@psu.edu

*Department of Mathematics, The Pennsylvania State University, University Park, PA, United States*

JOHN HARLIM: jharlim@psu.edu

*Department of Mathematics, Department of Meteorology and Atmospheric Science,  
The Pennsylvania State University, University Park, PA, United States*

## SYMMETRIZED IMPORTANCE SAMPLERS FOR STOCHASTIC DIFFERENTIAL EQUATIONS

ANDREW LEACH, KEVIN K. LIN AND MATTHIAS MORZFELD

We study a class of importance sampling methods for stochastic differential equations (SDEs). A small noise analysis is performed, and the results suggest that a simple symmetrization procedure can significantly improve the performance of our importance sampling schemes when the noise is not too large. We demonstrate that this is indeed the case for a number of linear and nonlinear examples. Potential applications, e.g., data assimilation, are discussed.

### 1. Introduction

Consider a stochastic differential equation (SDE)

$$dX_t = f(X_t) dt + \sigma dB_t, \quad X_t \in \mathbb{R}^D, \quad (1-1)$$

where  $f : \mathbb{R}^D \rightarrow \mathbb{R}^D$  and  $B_t$  is  $D$ -dimensional Brownian motion. Suppose we make noisy observations of the system at times  $t = T, 2T, 3T, \dots, JT$  ( $T > 0$ , fixed), obtaining a sequence of measurements  $Y_j = m(X_{jT}) + \eta_j$ , where  $m : \mathbb{R}^D \rightarrow \mathbb{R}^d$  ( $d \leq D$ ) is the quantity being measured (the “observable”),  $\eta_j$  are independent identically distributed (IID) random variables modeling measurement errors, and  $j = 1, \dots, J$ . What is the conditional distribution of  $X_t$  for  $t \in [0, JT]$  given  $Y_1, Y_2, \dots, Y_J$ ? This problem of “nonlinear filtering” or “data assimilation” arises in many applications; see, e.g., [7; 8; 5; 27]. A variety of algorithms have been developed to address it, but efficient data assimilation, especially in high-dimensional nongaussian problems, remains a challenge [25].

This paper concerns an approach to data assimilation known as “particle filtering” (see, e.g., [8] for more details) based on *sampling* the conditional distributions. We present an asymptotic analysis of certain sampling algorithms designed to improve the efficiency of particle filtering, and based on this analysis, we propose a general way to improve their performance. The analysis relies on taking a small noise limit, but the algorithms do not require a small noise to operate (but may not be as efficient when the noise is not small). We focus on *one* step of the filtering problem;

---

MSC2010: 65C05.

Keywords: importance sampling, stochastic differential equations, small noise theory, symmetrization, data assimilation,

i.e., we set  $J = 1$  in the above, as this is sufficient to capture the computational difficulty we wish to address. For simplicity, we assume  $\eta \sim \mathcal{N}(0, rI)$ , where  $r > 0$  is a scalar and  $I$  is the  $d \times d$  identity matrix; we also assume  $\sigma > 0$  is a scalar. These assumptions can be relaxed if needed.

To take one step of particle filtering, one begins by discretizing (1-1) using, e.g., the Euler scheme, to obtain

$$X_{n+1} = X_n + \Delta t f(X_n) + \sqrt{\Delta t} \sigma \cdot \xi_n, \quad X_0 = x_0 \in \mathbb{R}^D, \quad n = 0, \dots, N-1, \quad (1-2)$$

where  $N\Delta t = T$  and the  $\xi_n$  are IID standard normal random variables. A straightforward application of Bayes' theorem tells us that the conditional distribution of interest satisfies

$$p(x_1, \dots, x_N | y) \propto \exp\left(\frac{1}{2\sigma^2\Delta t} \sum_{n=0}^{N-1} \|x_{n+1} - x_n - f(x_n)\Delta t\|^2 + \frac{\|m(x_N) - y\|^2}{2r}\right). \quad (1-3)$$

One then tries to design a Monte Carlo algorithm to generate discrete time sample paths  $(X_1, \dots, X_N)$  from (1-3), conditioned on the observation  $y$ . We refer to the distribution in (1-3) as the *target distribution*. They are the discrete time analogs of the conditional distributions introduced above, with  $J = 1$  observation.

Without the last term in the exponent in (1-3), the target distribution is just the distribution of the discretized SDE, and one can generate sample paths by carrying out the recursion in (1-2). When the last term is included, however, it is generally not feasible to sample directly from the target distribution. A solution to this problem is *importance sampling*: instead of drawing samples from the target distribution, we draw sample paths  $(Z_1, \dots, Z_N)$  from an approximation  $q$ , usually called the “proposal distribution”. Any statistics we compute based on sample paths from  $q$  will be biased. We compensate for this bias by associating a weight  $W^{(k)} > 0$  to the  $k$ -th sample path  $(Z_1^{(k)}, \dots, Z_N^{(k)})$ , with  $\sum_k W^{(k)} = 1$ , so that the *weighted* sample paths  $(Z^{(k)}, W^{(k)})$  again have the correct statistics (in a sense we make precise later).

Vanden-Eijnden and Weare [28; 29] proposed an algorithm for sampling distributions like (1-3). They showed that their algorithm is efficient in the sense that in the limit of small dynamical and observation noise, the relative variance of the weights vanishes (see [29] for precise definitions and statements). The basic idea of the sampler is to look for the most likely sample path of the target distribution (1-3) and use this information to modify the dynamics so that samples from the proposal remain close to the target distribution. In this paper, by a combination of formal asymptotic analysis and numerical examples, we show that a symmetrization procedure proposed in [17] can be applied to SDEs to improve the efficiency of importance samplers. The symmetrization and “small noise analysis” has also been discussed in the context of implicit sampling [6; 23]; see [17].



While our primary motivation here is data assimilation for SDEs, our symmetrization procedure may be effective for sequential Monte Carlo sampling of more general types of systems. As well, the class of importance sampling algorithms studied here are closely related to algorithms proposed in [12; 13; 10; 9; 11] and in [28] for sampling “rare events” in SDEs, though there are some significant differences between the two applications. We plan to explore some of these connections in future work.

*Paper organization.* The remainder of this paper is organized as follows. We state our main results in Section 2. Section 3 briefly reviews the linear map method and its symmetrization, as well as the small noise theory [17]. We explain a new sampling method, the dynamic linear map, in Section 4. We study its efficiency in the small noise regime and show how to use symmetrization to improve its efficiency in small noise problems. Several numerical examples are provided in Section 5 that illustrate our asymptotic results as well as the efficiency of our dynamic approach in multimodal problems. The continuous time limit of the dynamic linear map is discussed in Section 6, and we present conclusions in Section 7.

## 2. Problem statement and summary of results

We now formulate the problem more precisely and summarize our key findings. We consider a discretized SDE in the small noise regime

$$X_{n+1} = X_n + \Delta t \tilde{f}(X_n, \Delta t) + \sqrt{\Delta t} \sqrt{\varepsilon} \sigma \cdot \xi_n, \quad X_0 = x_0 \in \mathbb{R}^D, \quad (2-1)$$

where  $\tilde{f}(x, \Delta t) = f(x) + O(\Delta t)$  corresponds to a numerical discretization of  $\dot{x} = f(x)$  (for most of this paper, we assume the Euler discretization  $\tilde{f}(x, \Delta t) = f(x)$ ), and  $\varepsilon \ll 1$  is the “small noise parameter”. Throughout this paper we assume that the  $D$ -dimensional vector field  $\tilde{f}$  is smooth, and that the process starts at a given initial position  $x_0$  and proceeds for  $N$  time steps of size  $\Delta t$  each. The transitions are made with independent gaussian samples  $\xi_n \sim \mathcal{N}(0, I)$ . We denote the path as  $x_{1:N}$ , a sequence of positions  $x_1, \dots, x_N$ , and its likelihood in the process with the path distribution  $\rho(x_{1:N} | x_0)$ .

The observation of the state at time  $N\Delta t$  gives rise to the *likelihood*

$$\theta(x_N) := \exp\left(-\frac{1}{\varepsilon} g(x_N)\right), \quad (2-2)$$

where  $g$  is assumed to be a smooth, nonnegative function. For example, for observations  $y = m(x_N) + \eta$ ,  $\eta \sim \mathcal{N}(0, \varepsilon r I)$ , we have  $g(x_N) = (2r)^{-1} \|m(x_N) - y\|^2$ . Hereafter we will sometimes refer to  $g$  as the “log-likelihood”, in a slight abuse of standard terminology. By Bayes’ theorem, the target distribution then has the form

$$p(x_{1:N} | x_0) \propto \rho(x_{1:N} | x_0) \cdot \theta(x_N). \quad (2-3)$$

Importance sampling methods generate samples using a proposal distribution  $q$ , and attach weights

$$W^{(k)} = w(X_{1:N}^{(k)} | x_0) = p(X_{1:N}^{(k)} | x_0) / q(X_{1:N}^{(k)} | x_0) \quad (2-4)$$

to each sample, so that the weighted samples can be used to compute unbiased statistical estimates with respect to the target distribution. To measure the efficiency of the sampling methods, we evaluate the relative variance of the weights

$$Q := \frac{\text{Var}[W]}{\text{E}[W]^2}. \quad (2-5)$$

Here the expected values are computed with respect to the proposal distribution  $q$ . This relative variance  $Q$  is connected to a standard heuristic called the “effective sample size”, defined by

$$N_{\text{eff}} := \frac{N_e}{1 + Q}, \quad (2-6)$$

where  $N_e$  is the number of weighted samples (see, e.g., [4; 21; 8]). The effective sample size is meant to measure the size of an unweighted ensemble that is equivalent to the weighted ensemble of size  $N_e$ . All else being equal, the smaller the  $Q$ , the more efficient the importance sampling algorithm, and if all the samples were independent, we would have  $Q = 0$  and  $N_{\text{eff}} = N_e$ . The quantity  $Q$  is convenient because it is not tied to any specific observable; recent work [1] has also given it a more precise meaning. Other quantities that can assess effective sample sizes are discussed in [22]. We note that in practice,  $p$  and  $q$  are only known up to a constant. The algorithms we describe do not require knowing the normalization constants. Likewise,  $Q$  is invariant under rescaling of  $p$  or  $q$  by a constant.

We study two types of importance sampling methods in this paper. The first method, called the “linear map” (LM), uses a gaussian proposal distribution centered at the most likely path. The second method, called “dynamic linear map” (DLM), reapplies the linear map after each time step between  $t = 0$  and  $t = N\Delta t$  given the previous moves. Note that the linear map can be viewed as a version of implicit sampling [6; 23] applied to the path distribution of an SDE. The dynamic linear map applies this implicit sampling step repeatedly to transition densities and is also closely linked to the continuous time control method of Vanden-Eijnden and Weare [28; 29] (see also Section 6). For each method, we perform a symmetrization and exploit symmetries of the proposal distributions to increase sampling efficiency. Symmetrization was previously studied for the LM in a more general context in [17]. Here we adapt this procedure to problems involving SDE and to the dynamic linear map. Following the approach taken in [17], we show that under suitable assumptions (see Section 4), the relative variances of the various methods are as follows:

method	$Q(\varepsilon)$ scaling
linear map (LM)	$O(\varepsilon)$
symmetrized LM	$O(\varepsilon^2)$
dynamic LM (DLM)	$O(\varepsilon)$
symmetrized DLM	$O(\varepsilon^2)$

We also present examples showing that the leading coefficient of the DLM can be smaller than that of LM, suggesting that DLM may be more effective in some situations (see Section 5). We discuss the continuous time limit of LM and DLM for scalar SDE, and calculate the leading coefficient of  $Q(\varepsilon)$  in an asymptotic expansion in  $\varepsilon$ . In doing so, we show that, under additional assumptions, the sampling method discussed in [28] is recovered in the  $\Delta t \rightarrow 0$  limit of the DLM (see Section 6).

*Notes.*

- (i) The  $\varepsilon$ -expansions we will consider are formally justified as the relevant quantities; e.g., relative weight variance, are gaussian integrals.
- (ii) The insertion of the small noise parameter  $\varepsilon$  into the problem is mainly to enable asymptotic analysis. In specific problems, there is not always an identifiable small parameter, and in any case our methods do not require a small parameter to operate.

### 3. Background

We simplify notation and write  $x := x_{1:N}$ , and  $F(x) := F(x_{1:N} | x_0)$ , and consider the small noise target distribution defined in (2-3) which can be written as  $p(x) \propto \exp(-F(x)/\varepsilon)$ , where

$$F(x) = \frac{\Delta t}{2\sigma^2} \sum_{n=0}^{N-1} \left\| \frac{x_{n+1} - x_n}{\Delta t} - \tilde{f}(x_n, \Delta t) \right\|^2 + g(x_N), \tag{3-1}$$

for  $g$ , a scalar function as in (2-2). If we assume that  $F$  has a unique, nondegenerate minimum, and let

$$\varphi = \arg \min_{x \in \mathbb{R}^{D \cdot N}} F(x), \tag{3-2}$$

i.e.,  $\varphi$  is the optimal path with prescribed initial condition  $x_0$ , we can employ Laplace asymptotics to expand the target distribution around  $\varphi$ . (See, e.g., [24] for a general formulation of Laplace asymptotics.) After a change of variables

$$z = \varepsilon^{-1/2} \cdot (x - \varphi) \tag{3-3}$$

the expansion is

$$F(z) = F(\varphi) + z^T H z / 2 + \varepsilon^{1/2} C_3(z) + \varepsilon C_4(z) + O(\varepsilon^{3/2}), \tag{3-4}$$

---

Calculate  $\varphi$  and  $H$  starting from  $x_0$ .

**for**  $m = 1$  to  $M$  **do**

Sample  $X \sim \mathcal{N}(\varphi, \varepsilon H^{-1})$ .

Calculate  $W = p(X)/q(X)$ .

Return  $M$  weighted samples  $X, W$ .

---

**Algorithm 1.** Linear map.

where  $H$  is the Hessian evaluated at  $\varphi$  and  $C_k$  are the higher-order terms in the Taylor series. Here and below, we use the shorthand  $F(z) := F(\varphi + \varepsilon^{1/2}z)$ , and similarly write  $w(z)$  for  $w(\varphi + \varepsilon^{1/2}z)$ , etc. Note that while we will continue to refer to  $z := \{z_1, \dots, z_n\}$  as a “path” after the change of coordinates,  $x = \varphi + \sqrt{\varepsilon}z$  is the actual solution of (2-1).

The small noise analysis of LM, and other methods to follow will make frequent use of this expansion, as well as the “variance lemma” [17].

**Lemma** (variance lemma). *For a function  $u(z, \varepsilon)$  that can be expanded in  $\varepsilon$  at least to the terms*

$$u(z) = 1 + \varepsilon^r u_1(z) + \varepsilon^{2r} u_2(z) + O(\varepsilon^{3r}), \quad (3-5)$$

*the relative variance of  $u$  with respect to a probability density  $q$  is*

$$Q = \varepsilon^{2r} \text{Var}_q[u_1(z)] + O(\varepsilon^{3r}). \quad (3-6)$$

**3.1. Linear map.** The proposal distribution of the linear map (LM) sampling method, summarized in Algorithm 1, is gaussian and proportional to

$$q(z) \propto \exp(-z^T H z / 2). \quad (3-7)$$

The weights are the ratio of target and proposal distribution, and can be expanded as

$$w(z) = 1 - \varepsilon^{1/2} C_3(z) + O(\varepsilon). \quad (3-8)$$

Using the variance lemma we thus find that

$$Q = \varepsilon \text{Var}_q[C_3(z)] + O(\varepsilon^{3/2}), \quad (3-9)$$

i.e., the relative variance of the weights is linear in  $\varepsilon$  (see [17] for more details).

**3.2. Symmetrized linear map.** It is shown in [17] that the linear map can be “symmetrized” to improve the scaling of  $Q$  from linear to quadratic in  $\varepsilon$ . This stems from the observation that the leading-order term in the weight is an odd function with respect to the random variable  $z$ , whose probability distribution function is even. The symmetrized sampler uses a proposal distribution which reweights equally likely samples from the gaussian distribution of the linear map such that the resulting

weights have even symmetry. The odd leading-order terms in the weight expansions then cancel, which results in a quadratic scaling of  $Q$  in  $\varepsilon$ .

Specifically, the symmetrized linear map draws a sample  $z$  from the proposal distribution  $q$ . It returns  $z$  with probability  $w^+/(w^- + w^+)$ , and  $-z$  with probability  $w^-/(w^- + w^+)$ , where

$$w^+ = \frac{p(-z)}{q(z)}, \quad w^- = \frac{p(z)}{q(z)}. \quad (3-10)$$

Samples generated in this way have a nonsymmetric distribution, but even weights:

$$q_s(z) = q(z) \frac{2w^+}{w^- + w^+}, \quad w_s(z) = \frac{w^- + w^+}{2}. \quad (3-11)$$

The Taylor expansion of the symmetrized weight is

$$w_s(z) = 1 + \varepsilon \left( \frac{1}{2} C_3(z)^2 - C_4(z) \right) + O(\varepsilon^2), \quad (3-12)$$

which, together with the variance lemma shows that

$$Q_s = \varepsilon^2 \text{Var}_q \left[ \frac{1}{2} C_3(z)^2 - C_4(z) \right] + O(\varepsilon^4). \quad (3-13)$$

The symmetrization therefore improves the linear scaling of  $Q$  in  $\varepsilon$  of LM, to a quadratic scaling of  $Q$  for SLM (see [17] for more details).

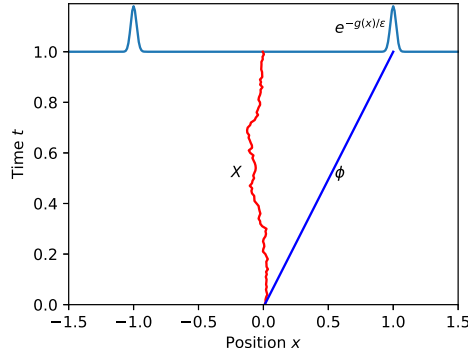
#### 4. Dynamic linear map and its symmetrization

**4.1. A multimodal example.** The linear map can be efficient when the hypotheses underlying its derivation are satisfied, i.e., when the pathspace distribution is unimodal and a gaussian approximation is appropriate. However, when there are multiple modes, LM can become inefficient. To see how this might happen, consider the simple random walk

$$X_{n+1} = X_n + \sqrt{\Delta t} \sqrt{\varepsilon} \xi_n, \quad (4-1)$$

i.e.,  $X_n = X_0 + \sqrt{\Delta t} \sqrt{\varepsilon} W_n$  where  $W_n$  is standard Wiener process. Suppose we have a bimodal likelihood function  $e^{-g(x)/\varepsilon}$  whose graph is as shown in Figure 1; this type of situation can arise when multiple states can give the same measurement, so that observations may have ambiguous interpretation. In this case, the high probability paths will be those that reach the vicinity of  $x = \pm 1$  at  $t = 1$ ; effectively, the high probability paths are sample paths of Brownian motion, conditioned to be near  $x = \pm 1$  at  $t = 1$ . The probability of this occurring by chance is exponentially small as  $\varepsilon \rightarrow 0$ , and direct sampling is unlikely to ever produce such a path.

A straightforward calculation shows that the optimal path  $\varphi$  approaches a straight line in the  $x$ - $t$  plane as  $\varepsilon \rightarrow 0$ , going to the right bump if  $X_0 > 0$  and to the left if  $X_0 < 0$  (and undefined if  $X_0 = 0$ ). With a bimodal likelihood function, the target



**Figure 1.** Brownian motion with bimodal likelihood. Here, the initial condition is  $X_0 = 0.01$ , and  $\epsilon = 0.1$ . Shown are a sample path  $X$  and the optimal path  $\varphi$  starting from  $X_0$ .

distribution  $p(x)$  is bimodal as well. If the initial condition is sufficiently to the right of  $x = 0$ , one of the two modes will dominate, and LM can be expected to be effective. As  $X_0$  moves closer to  $x = 0$ , however, the other mode will begin to make a greater contribution; at  $X_0 = 0$ , the two modes carry exactly the same weight. But LM will *always* pick the mode on the right when  $X_0 > 0$ , no matter how close  $X_0$  is to  $x = 0$ . So LM will produce essentially no sample paths going to the left, leading to a large weight variance. See Section 5 for detailed numerical results.

This is a well known problem with importance sampling algorithms. Similar issues arise in rare event simulation, and a standard solution is to dynamically recompute the optimal path. See, e.g., the discussion of Siegmund's algorithm in [2]. In our context, this leads to an algorithm we call the dynamic linear map, which is similar to the algorithms proposed in [28; 13]. We will also discuss symmetrization in this context.

**4.2. Dynamic linear map.** Roughly speaking, the dynamic linear map (DLM) consists of computing the optimal path  $\varphi$  starting from the current state  $X_n$ , taking *one* step (so that  $X_{n+1} = \varphi_{n+1}$ ) and then repeating. See Algorithm 2 for details. The DLM thus requires redoing LM *at every step*, and is therefore more expensive.<sup>1</sup> However, it can avoid some of the issues arising from multimodal target distributions. One can see this heuristically in the above example (Section 4.1): suppose we start with  $X_0$  slightly to the right of  $x = 0$ , so that the optimal path  $\varphi$  goes to the right bump. After a few steps, we may end up in a state  $X_n$  closer to the left bump. At this point, the DLM would start steering the sample path towards the left bump. Unlike LM, repeated sampling using DLM would yield sample paths that end at both the left and the right bumps (see Section 5.1).

<sup>1</sup>Suppose each cost function evaluation requires CPU time  $\propto N$ , the number of steps, and each optimization requires  $k$  function evaluations. Then all else being equal, LM has running time  $O(kN)$  and DLM  $O(kN^2)$ .

```

for  $m = 1$  to  $M$  do
  for  $n = 0$  to  $N - 1$  do
    Calculate  $\varphi$  and  $H$  starting from  $X_n$ .
    Calculate  $\Sigma_{n+1} = (H^{-1})_{1,1}/\Delta t$ .
    Sample  $X_{n+1} \sim \mathcal{N}(\varphi_{n+1}, \Delta t \varepsilon \Sigma_{n+1})$ .
    Calculate  $W_n = p(X_{n+1} | X_n)/q(X_{n+1} | X_n)$ .
  Calculate  $W = W_{N-1} \cdots W_0$ .
  Return  $M$  weighted samples  $X, W$ .
    
```

**Algorithm 2.** Dynamic linear map.

To make use of DLM, we need an expression for the associated weights. This, in turn, requires an expression for the proposal distribution  $q$  associated with DLM, which one can derive by first noting that in general, transition densities are marginals of the pathspace distribution:

$$\rho(x_{n+1} | x_n) = \int \rho(x_{n+1:N} | x_n) dx_{n+2:N}.$$

(Here we abuse notation slightly and use  $p$  and  $q$  to denote both pathspace distributions as well as their marginals.) The DLM transition density arises from making a gaussian approximation of the target distribution at each step and then taking its marginal. This leads to

$$\begin{aligned}
 q(x_{n+1} | x_n) &= \int q(x_{n+1:N} | x_n) dx_{n+2:N} \\
 &\propto \exp\left(- (x - \varphi)_{n+1}^T \Sigma_{n+1}^{-1} (x - \varphi)_{n+1} / (2\Delta t)\right). \tag{4-2}
 \end{aligned}$$

Here  $\varphi$  is the optimal path from  $x_n$  to  $x_N$  and we omit its dependence on  $x_n$  for readability of the equations; we also remind the reader that  $x = x_n, \dots, x_N$  is a path. We denote the Hessian of  $F(x)$  evaluated at the optimal path  $\varphi$  by  $H$ . We view a path from  $x_n$  to  $x_{n+k}$  as a point in  $\mathbb{R}^{kD}$ , arranged in  $k$  blocks of  $D$  entries. Accordingly, the matrix  $H$  can be viewed as an element of  $\mathbb{R}^{(N-n)D \times (N-n)D}$  and can be subdivided into  $(N-n) \times (N-n)$  blocks of dimension  $D \times D$  each. The matrix  $\Sigma_{n+1}$  in (4-2) is  $(H^{-1})_{1,1}/\Delta t$ , the first block of the inverse of the Hessian  $H$  (after rescaling).

In Algorithm 2, going from step  $n$  to  $n+1$  requires optimizing over the  $(N-n)D$  remaining variables in the path. This is done independently at every step and for every sample path. The weights for the proposal distribution of DLM can be calculated as described in Algorithm 2, or as the product of the incremental weights

$$w = \prod_{n=0}^{N-1} w_n, \quad w_n \propto \frac{p(x_{n+1} | x_n)}{q(x_{n+1} | x_n)}. \tag{4-3}$$

*Relation to Hamilton–Jacobi equation and regularity of “value functions”.* In the definitions above, it is assumed that  $q(x_{n+1} | x_n)$  is well defined for all  $(x_n, x_{n+1})$ . This is actually not always the case. To see this, consider again the example from Section 4.1. If  $x_n = 0$  at some  $n$ , there are two optimal paths pointing in opposite directions. At this point, because there is not a single optimal path,  $q(x_{n+1} | x_n)$  is undefined. This behavior is actually rather common, and not at all confined to the Brownian motion example. It is closely connected with regularity of solutions of a partial differential equation of Hamilton–Jacobi (HJ) type. As we do not make use of the theory of HJ equations in this paper, we do not go into details here. Instead, we provide a brief summary below, and refer interested readers to, e.g., [29] or [12; 13; 10; 9], for more information.

In the DLM method, the optimal path minimizes a version of the function  $F$  in (3-1), but starting with state  $x_n$  at time  $n$  rather than always at time 0. In the limit as  $\Delta t \rightarrow 0$ , the *value function*  $u(x, t)$  achieved with initial condition  $x_n = x$  at step  $n\Delta t = t$  solves an HJ equation of the form  $\partial_t u = H(x, Du)$ , with Hamiltonian  $H(x, p) = (\sigma^2/2)|p|^2 + p \cdot f(x)$ ; this is the Legendre transformation of the Freidlin–Wentzell Lagrangian  $L(x, v) = |v - f(x)|^2/(2\sigma^2)$  [14]. For the HJ equation to be well posed, one prescribes the *final condition* that  $u(x, T) = g(x)$ , where  $g$  is the likelihood in (2-2) and  $T > 0$ . The HJ equation is then solved backwards in time. The time derivative  $\dot{\phi}$  of the optimal path starting at position  $x$  and time  $t$  is given by the gradient of  $u(x, t)$  where it is differentiable. At locations  $(x, t)$  where there are multiple optimal paths, the value function  $u(x, t)$  is generally continuous but not differentiable. At such *singular points*  $x$ ,  $q(x_{n+1}, x)$  has jump discontinuities (as  $x$  varies) and is therefore undefined.

Though very much relevant to the efficacy of the type of methods discussed in this paper, the analysis of singularities of HJ equations can be highly nontrivial. As our main goal is to assess whether some version of the symmetrization procedure proposed in [17] can be extended to SDEs, we have opted to focus on the simplest possible setting, leaving more general analysis to future work. *For the remainder of the paper, we make the following standing assumption:*

$q(x_{n+1} | x_n)$  is defined everywhere, and is as smooth as needed.

The analytical results described below should therefore be interpreted as a *best-case scenario*. We also note that while the numerical algorithm is unlikely to produce an  $x_n$  *exactly* in the set of singular points in actual practice, the presence of singularities does mean that the performance of the algorithm may be worse than predicted by our analysis. We have therefore designed our numerical examples to test the extent to which the algorithms behave as predicted even when  $q(x_{n+1} | x_n)$  is not differentiable everywhere.



**4.3. Small noise analysis.** To find the scaling of the relative variance of the weights of DLM with the small noise parameter  $\varepsilon$ , we apply the same change of variables as in (3-3) to each transition density and expand the incremental weights  $w_n$  as

$$w_n = w(z_{n+1} | z_n) = 1 + \varepsilon^{1/2} \cdot w_{1,n}(z_{n+1} | z_n) + \varepsilon \cdot w_{2,n}(z_{n+1} | z_n) + O(\varepsilon^{3/2}), \quad (4-4)$$

where

$$w_{1,n}(z_{n+1} | z_n) = \frac{\int C_3(z) \exp(-z^T H z / 2) dz_{n+2:N}}{\int \exp(-z^T H z / 2) dz_{n+2:N}}, \quad (4-5)$$

$$w_{2,n}(z_{n+1} | z_n) = \frac{\int (C_3(z)^2 / 2 - C_4(z)) \exp(-z^T H z / 2) dz_{n+2:N}}{\int \exp(-z^T H z / 2) dz_{n+2:N}} - \int (C_3(z)^2 / 2 - C_4(z)) \exp(-z^T H z / 2) dz_{n+1:N}, \quad (4-6)$$

noting that (4-4) relies strongly on our standing assumption that  $q(x_{n+1} | x_n)$  is differentiable. Since the weight of a sample is the product of the incremental weights, we have

$$w(z) = 1 + \varepsilon^{1/2} \cdot w_1 + \varepsilon \cdot w_2 + O(\varepsilon^{3/2}),$$

where

$$w_1 = \sum_{n=0}^{N-1} w_{1,n}, \quad w_2 = \sum_{n=0}^{N-1} w_{2,n} + \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} w_{1,n} \cdot w_{1,m}. \quad (4-7)$$

The scaling of  $Q$  in  $\varepsilon$  now follows from the variance lemma:

$$Q^\varepsilon = \varepsilon \cdot \text{Var}_q[w_1] + O(\varepsilon^2). \quad (4-8)$$

Thus, the relative variance of DLM scales linearly in  $\varepsilon$ , the same asymptotic scaling as LM. However, we will show in numerical examples below that the dynamic approach can be more effective in practice than LM, especially when the target distribution has multiple modes.

**4.4. Symmetrization.** The leading-order term in the weight for DLM has an odd symmetry, just like the LM, and a symmetrization procedure can be applied to DLM to improve the scaling of  $Q$  in  $\varepsilon$ . The reason is that, at each time step,  $X_{n+1}$  is generated by a composition of the previous state  $X_n$  and a new gaussian sample  $\xi_n$ . While this procedure leads to a proposal distribution that is not necessarily even, the paths are constructed incrementally from gaussian samples which are even.

More specifically, the recursive composition forms a map  $h$  from the  $N \cdot D$  dimensional gaussian to the path  $X = h(\varepsilon^{1/2} \xi)$ , and for every sampled path  $X^+ = h(\varepsilon^{1/2} \xi)$ , there is a path  $X^- = h(-\varepsilon^{1/2} \xi)$  which is equally likely. Following the algorithm described in Algorithm 3, we sample  $X^+$  with probability  $W^+ / (W^+ + W^-)$ , and

---

**for**  $m = 1$  to  $M$  **do**

  Sample  $\xi \sim \mathcal{N}(0, I)$ .

  Calculate  $X^+ = h(\varepsilon^{-1/2}\xi)$  and  $X^- = h(-\varepsilon^{-1/2}\xi)$ .

  Calculate  $W^+ = p(X^+)/q(X^+)$  and  $W^- = p(X^-)/q(X^-)$ .

  Sample  $X = X^+$  with probability  $W^+/(W^+ + W^-)$  and  $X = X^-$  with probability  $W^-/(W^+ + W^-)$ .

  Calculate  $W = (W^+ + W^-)/2$ .

Return  $M$  weighted samples  $X, W$ .

---

**Algorithm 3.** Symmetrization.

$X^-$  with probability  $W^-/(W^+ + W^-)$ ; the resulting proposal is a ‘‘symmetrized’’ distribution with even weights (see (3-11)).

The symmetrized weights can be written in terms of the map as

$$w_s(h(\varepsilon^{1/2}\xi)) = \frac{w(h(\varepsilon^{1/2}\xi)) + w(h(-\varepsilon^{1/2}\xi))}{2}. \quad (4-9)$$

Recall the expansion of the weights in (4-4), and note that

$$z = \varepsilon^{-1/2}(h(\varepsilon^{1/2}\xi) - h(0)),$$

since the most likely path  $\varphi$  can be written in terms of the map as  $\varphi = h(0)$ .

If  $\varphi$  is unique (at each time step),  $h$  can be expanded around the most likely path as

$$h(\varepsilon^{1/2}\xi) = \varphi + \varepsilon^{1/2}(Dh)(0) \cdot \xi + O(\varepsilon), \quad (4-10)$$

$$h(-\varepsilon^{1/2}\xi) = \varphi - \varepsilon^{1/2}(Dh)(0) \cdot \xi + O(\varepsilon). \quad (4-11)$$

We thus have that

$$\begin{aligned} w(h(\varepsilon^{1/2}\xi)) &= 1 + \varepsilon^{1/2}w_1(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) \\ &\quad + \varepsilon w_2(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) + O(\varepsilon^{3/2}), \end{aligned} \quad (4-12)$$

$$\begin{aligned} w(h(-\varepsilon^{1/2}\xi)) &= 1 - \varepsilon^{1/2}w_1(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) \\ &\quad + \varepsilon w_2(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) + O(\varepsilon^{3/2}) \end{aligned} \quad (4-13)$$

which results in the cancellation of the leading-order term in  $\varepsilon$  of the symmetrized weight

$$w_s(h(\varepsilon^{1/2}\xi)) = 1 + \varepsilon w_2(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) + O(\varepsilon^{3/2}). \quad (4-14)$$

Applying the variance lemma completes the proof for the quadratic scaling of  $Q_s$  in  $\varepsilon$ :

$$Q_s = \varepsilon^2 \cdot \text{Var}_{q_s}[w_2] + O(\varepsilon^4). \quad (4-15)$$

## 5. Numerical examples

We now examine a number of concrete examples, both to illustrate the scaling of the proposed algorithms and to test their limitations. The source code for all examples in this section can be found on GitHub.<sup>2</sup>

**5.1. Examples with linear SDE.** We begin with the Brownian motion example from Section 4.1:

$$X_{n+1} = X_n + \sqrt{\Delta t} \sqrt{\varepsilon} \xi_n, \quad (5-1)$$

with initial condition  $X_0 = x_0$  and with likelihood  $\theta = e^{-g(X_N)/\varepsilon}$  for two different choices for  $g$ . We first consider the case of a unimodal target distribution for which the assumptions made during the small noise analysis are satisfied. We then violate the assumption of a unique optimal path to indicate limitations of DLM and our small noise analysis. For the examples below, the time step is  $\Delta t = 10^{-2}$ . The observation is collected at step  $N = 100$  (i.e.,  $T = 1$ ). Computing the optimal paths is straightforward to do analytically, and we use the analytic formulas in our implementation of the various samplers.

*Brownian motion with unimodal likelihood.* We first consider a likelihood defined by

$$g(x) = \frac{1}{24}x^4 + \frac{1}{6}x^3 + \frac{1}{2}x^2.$$

The likelihood is asymmetric in  $x$  and leads to a nongaussian and unimodal target distribution. In this example, the assumptions made in our small noise analysis are satisfied.

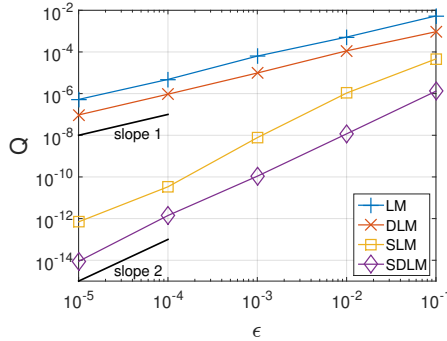
We apply LM, SLM, DLM, and SDLM to sample the target distribution over a wide range of  $\varepsilon$ , and compute the relative variance  $Q$  for each of these methods. For each  $\varepsilon$  and method (LM, SLM, DLM, and SDLM), we draw 1200 samples. The results are shown in Figure 2. As can be seen, the results show the predicted scalings for  $Q$  for a wide range of  $\varepsilon$  for all four methods: both LM and DLM are  $O(\varepsilon)$ , while SLM and SDLM are both  $O(\varepsilon^2)$ . Perhaps this is no surprise, as all assumptions that lead to the small noise theory are valid in this example. We also see that the dynamic methods (DLM and SDLM) have smaller relative variance  $Q$  at each value of  $\varepsilon$ , though they also cost more per sample.

*Brownian motion with bimodal likelihood.* Next, we examine

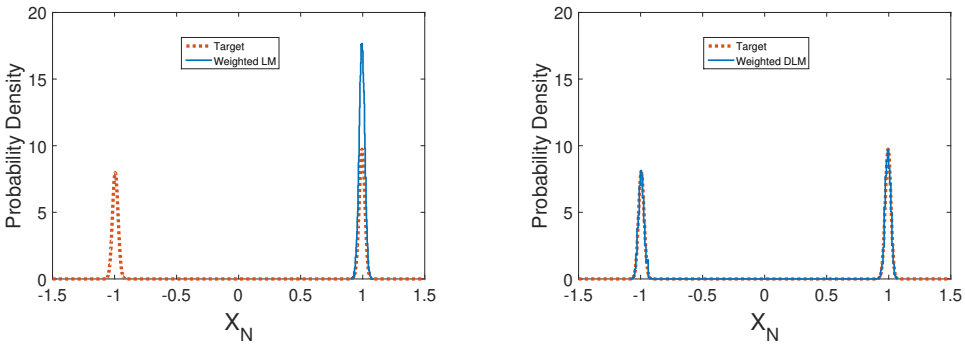
$$g(x) = 100 \cdot \left( \frac{1}{4}x^4 - \frac{1}{2}x^2 \right).$$

As explained in Section 4.1, this leads to a bimodal target distribution. We fix  $\varepsilon = 10^{-1}$ , and leave all other parameters as above. We apply LM and DLM to

<sup>2</sup>[https://github.com/AndrewLeach/SDE\\_Importance\\_Sampling](https://github.com/AndrewLeach/SDE_Importance_Sampling)



**Figure 2.** Brownian motion with asymmetric unimodal likelihood. The scaling of  $Q$  in  $\epsilon$  for LM, SLM, DLM, and SDLM are plotted.

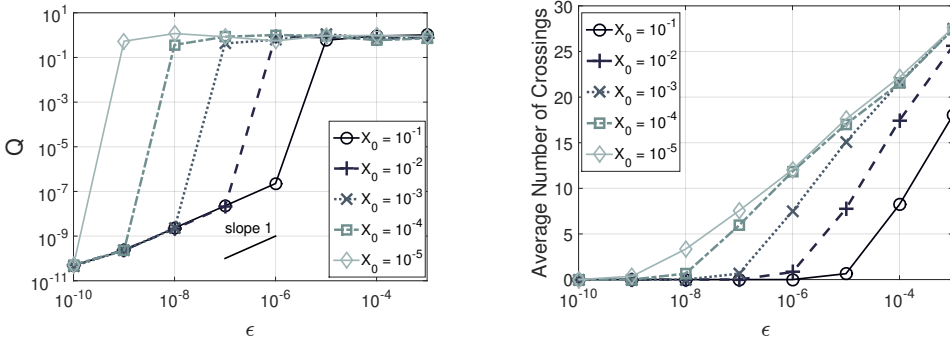


**Figure 3.** Final-time marginal distributions for Brownian motion with bimodal likelihood. Left: we plot the marginal distribution  $p(x_N | x_0)$  estimated by weighted histograms of 12000 samples generated using LM. Also shown is the target distribution. Right: we plot the same information for DLM.

compute the final-time distribution  $p(X_N | X_0)$ , using  $1.2 \times 10^4$  (weighted) samples. The results are shown in Figure 3, along with the target distribution  $\propto e^{-(g(x)+x^2/2)/\epsilon}$ .

As expected, LM essentially ignores one of the two modes, while DLM captures both modes. As explained before, even though both samplers should reproduce the target distribution in the large-sample-size limit, in practice LM produces almost no sample paths that go to the left bump. In contrast, DLM readily generates sample paths ending at both bumps, leading to a more effective sampling of the target distribution. We have experimented with increasing the sample size for LM, but even the largest sample sizes we consider did not lead to weighted samples that represent both modes.

Finally, note that empirical estimates of  $Q$  are insufficient to detect this problem: even though the true value of  $Q$  for LM should be quite large in this case, empirical estimates of  $Q$  for LM are actually quite small because none of the sample paths go to the left bump. Indeed, for Figure 3, the empirical  $Q$  for LM is  $\sim 3 \times 10^{-3}$ , while



**Figure 4.** DLM applied to the overdamped Langevin equation with bimodal likelihood. Left: the scaling of  $Q$  versus  $\epsilon$  for  $x_0$  approaching  $x = 0$ . Right: we plot the average number of  $x = 0$  crossings against  $\epsilon$ .

that of DLM is  $\sim 1$ . The example thus shows that for nongaussian and possibly multimodal distributions, DLM can be more reliable despite the same scaling of  $Q$ .

*Overdamped Langevin equation with bimodal likelihood.* The scaling arguments for DLM and its symmetrized version rely on the assumption that the most likely path  $\varphi$  is unique at every time step. We now consider an example for the DLM in which we deliberately violate this assumption. The model is

$$X_{n+1} = X_n - \Delta t \alpha \cdot X_n + \sqrt{\Delta t} \sqrt{\epsilon} \xi_n, \tag{5-2}$$

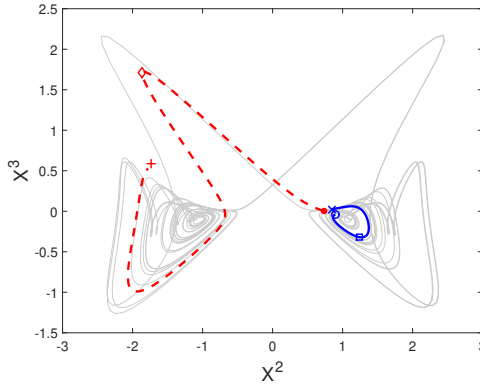
the Euler discretization of the overdamped Langevin equation  $\dot{X} = -\alpha X + \sqrt{\epsilon} \dot{B}$ . We use the log-likelihood

$$g(x) = 10 \cdot \left(\frac{1}{4}x^4 - \frac{1}{2}x^2\right).$$

As in the previous example, the optimal path goes to the right bump when  $X_0 > 0$  and to the left when  $X_0 < 0$ . At  $X_0 = 0$  there is no unique optimal path.

The linear drift makes it likely that DLM sample paths encounter the  $x = 0$  line and the small noise results may not hold in this case. To illustrate the behavior and efficiency of the methods in this situation, we perform experiments with varying values of  $\epsilon$  and  $x_0$ . Specifically, for a fixed  $\epsilon$ , we take  $N = 10^3$  time steps with DLM, starting from initial conditions ranging from  $x_0 = 10^{-1}$  to  $x_0 = 10^{-5}$ . We compute the averaging number of  $x = 0$  crossings for each experiment. Figure 4 shows the results as well as the computed values of  $Q$ .

As can be seen in Figure 4, left, the predicted asymptotic scaling of  $Q$  only emerges for small  $\epsilon$ ; the critical value of  $\epsilon$  at which the  $Q$  curve crosses over into the asymptotic regime decreases as  $x_0$  approaches 0, making crossings more likely. Comparing Figure 4, left and right, we see that the asymptotic regime corresponds to values of  $\epsilon$  small enough that the average number of crossings per sample is



**Figure 5.** The Gissinger model and its phase space geometry. Shown are trajectories of the deterministic model (light gray) projected to the  $x^2$ - $x^3$  plane. The dashed line is the most likely path with initial condition marked by “•” and measured state at time  $t = 10$  marked by “+”; this trajectory undergoes a “pole reversal” (Case (a)). The solid blue line represents the most likely path with initial condition “○” and observation “×” at  $t = 10$ , and does not exhibit a pole reversal (Case (b)). The symbols “□” and “◇” are the times at which we computed the histograms in Figure 6.

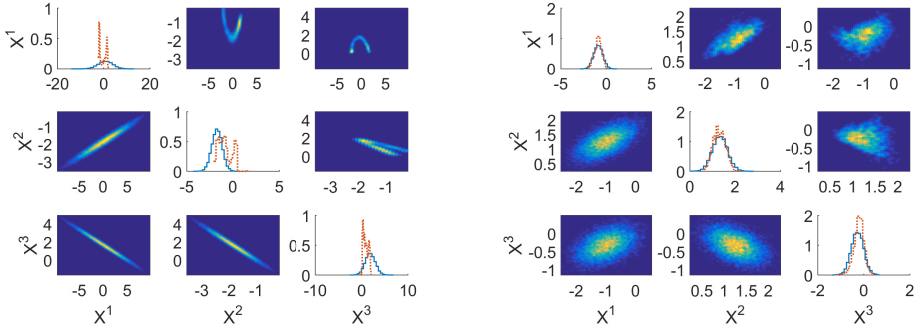
near zero. Closer examination of the data suggests that this critical  $\varepsilon$  scales roughly linearly with distance of the initial condition  $x_0$  to  $x = 0$ . The example thus suggests that the efficiency of DLM may suffer if one encounters nonunique optimal paths while constructing the proposal distribution  $q$  sequentially, but the predicted  $Q$  scaling again holds if  $\varepsilon$  is small enough.

Finally, we note that even in the preasymptotic regime, the values of  $Q$  are  $O(1)$ , meaning the effective number of samples is  $\approx N_e/2$ , which is still a significant improvement over direct sampling.

**5.2. Example with a nonlinear SDE.** Our second example is a stochastic version of an idealized geomagnetic pole reversal model due to Gissinger [16]:

$$\begin{aligned} \dot{x}^1 &= 0.119x^1 - x^2x^3 + \sqrt{\varepsilon}\dot{B}^1, \\ \dot{x}^2 &= -0.1x^2 + x^1x^3 + \sqrt{\varepsilon}\dot{B}^2, \\ \dot{x}^3 &= 0.9 - x^3 + x^1x^2 + \sqrt{\varepsilon}\dot{B}^3. \end{aligned} \tag{5-3}$$

(In this section,  $x^k$  refers to the  $k$ -th component of a vector  $x$ .) The  $\varepsilon = 0$  system of ordinary differential equations has 3 unstable fixed points:  $(0, 0, 0.9)$  and  $p_{\pm} \approx (\mp 0.96, \pm 1.05, -0.109)$ . It has a chaotic attractor on which trajectories circulate around either  $p_+$  or  $p_-$  many times before making a quick transition to the other fixed point. See Figure 5. Following [16], we refer to these transitions as “pole reversals”, since the second component  $x^2(t)$  can be thought of as a proxy for the geomagnetic dipole field, and it changes signs at these transitions.



**Figure 6.** Final-time marginal distributions for the Gissinger model. Left: Case (a). Right: Case (b). In each panel, the diagonal plots are histograms for the final-time marginal proposal distributions for of  $x^1$ ,  $x^2$ , and  $x^3$  (solid = LM and dashed = DLM). The times at which the marginals are computed are marked by “ $\diamond$ ” in Figure 5 for Case (a), and “ $\square$ ” for Case (b). Plots on the lower-triangular submatrix are two-dimensional marginal proposal distributions computed by LM, while two-dimensional marginal proposal distributions computed by DLM form the upper-triangle (see text for details).

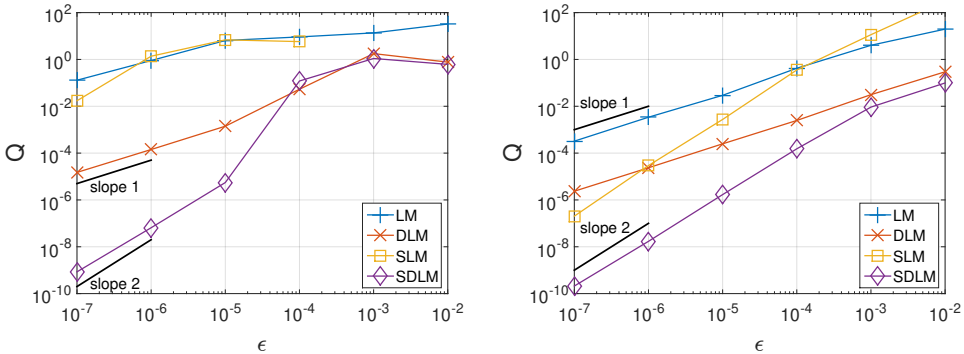
Here, we consider (5-3) with  $\varepsilon > 0$ . We start with an initial condition near  $p_+$ , and after  $N = 100$  steps make an observation with log-likelihood  $g(x) = \|x - y\|^2/2$ , where  $x = (x^1, x^2, x^3)$ . We view  $y \in \mathbb{R}^3$  as the outcome of a “measurement” made at step  $N$ .

We consider two cases:

- (a) the measured value  $y$  is near  $p_-$ , i.e., on the opposite “lobe” from the initial condition, or
- (b)  $y$  is near  $p_+$ , i.e., on the same “lobe” as the initial condition.

Figure 5 illustrates the initial conditions, data, and optimal paths for the two cases. Shown are trajectories of the deterministic model (light gray), representing the chaotic attractor. The dashed line is the most likely path with initial condition marked by “ $\bullet$ ” and with measured state at time  $t = 10$  marked by “ $+$ ”; this trajectory undergoes a “pole reversal” (Case (a)). The solid blue line represents the most likely path with initial condition “ $\circ$ ” and observation “ $\times$ ,” and does not exhibit a pole reversal (Case (b)).

To see how the two cases differ, we fix  $\varepsilon = 10^{-2}$  and apply the LM and DLM to generate 1200 sample paths in each case and plot marginals of the proposal distributions at two different times. In Case (a), we plot histograms of the marginal distributions at time  $j \Delta t$  as marked by “ $\diamond$ ” in Figure 5; in Case (b), we plot histograms of the marginal distributions at time  $j \Delta t$  as marked by “ $\square$ ”. For each method, the resulting “triangle plot” consists of histograms of the one-dimensional marginals,  $q(X_j^k | X_0)$  for  $k \in \{1, 2, 3\}$ , and the two-dimensional marginals,  $q(X_j^k, X_j^\ell | X_0)$ ,  $k \neq \ell$ , of the



**Figure 7.** Relative variance  $Q$  as a function of  $\varepsilon$  for the Gissinger model. Case (a) (left) involves a pole reversal, whereas Case (b) (right) does not.

proposal distributions. The triangle plots are shown in Figure 6. In each panel, the diagonal plots are the one-dimensional marginal distributions. The lower-triangular parts of each panel are the two-dimensional marginal distributions generated by LM, while the upper-triangular parts show marginals generated by DLM.

In Case (a), the marginal distributions of the DLM proposal are multimodal, possibly related to the underlying geometry of the strange attractor. In contrast, the LM proposal distribution misses this complexity altogether (as one might expect). Moving now to Case (b), which involves starting and end points on the same lobe connected by a shorter optimal path, the marginals are unimodal, and LM and DLM give more similar answers (though there is still significant deviation from gaussianity in the DLM proposal distribution).

Finally, we vary  $\varepsilon$  in Cases (a) and (b) and apply LM, SLM, DLM, and SDLM. For each value of  $\varepsilon$ , we estimate  $Q$  for each of the 4 methods. The results are shown in Figure 7. Not surprisingly, LM breaks down for Case (a), in which the target distribution is likely multimodal. In contrast, both DLM and SDLM exhibit the predicted scaling. For Case (b), because the target distribution is unimodal, all four methods behave as predicted by the small noise theory.

*Numerical details.* The Gissinger model requires attention to numerical implementation when we compute its statistics. We describe our numerical implementation in detail.

- (i) *Time-stepping.* The Euler scheme for the Gissinger model requires small time steps because of numerical instabilities. To improve stability, we discretize the drift part of (5-3) using a standard fourth-order Runge–Kutta (RK4) method and then adding IID  $\mathcal{N}(0, \sqrt{\varepsilon}\sqrt{\Delta t} I)$  normal random vectors at each step. This yields a model of the form (2-1), where  $\tilde{f}(x, \Delta t)$  now represents one step of the RK4 scheme. In all the examples shown above, the time step is  $\Delta t = 10^{-1}$ .



- (ii) *Estimation of  $Q$ .* In Figure 7, because of their different variances, we use 1200 sample paths to estimate  $Q$  for DLM and for SDLM, and 12000 paths for LM and for SLM.
- (iii) *Computing optimal paths.* Our methods requires computing optimal paths. For the Gissinger model, we use Newton’s method. Since explicit analytical expressions for the gradient and the Hessian are available, this is relatively straightforward to program. To reduce the (fairly significant) computational cost of computing  $\varphi$  at each time step, we “guess” a good initialization for the optimization procedure using the solution from the previous time step using the linearized dynamics. See [20] for details.

### 6. Continuous time limit of dynamic linear map

So far, we have focused on time discretizations of SDEs. A natural question is what happens to the proposed algorithms in the limit  $\Delta t \rightarrow 0$ . In this section, we sketch some analytical arguments aimed at addressing these questions for scalar SDE. Though restrictive, we believe these results yield useful insights. A more complete and rigorous analysis is left for future work, as it is expected to be more involved.

**6.1. Dynamic linear map.** For scalar SDE, the DLM can be defined through the recursion

$$X_{n+1}^{\Delta t} = \varphi_{n+1}^{\Delta t}(X_n^{\Delta t}, n) + \sqrt{\Delta t} \sqrt{\varepsilon} \sqrt{\Sigma_{n+1}^{\Delta t}(X_n^{\Delta t}, n)} \xi_n, \tag{6-1}$$

where  $\varphi_n^{\Delta t}(x_0, m)$ ,  $n \in \{m, m + 1, \dots, N\}$ , is the optimal path (3-2) with prescribed initial condition  $x_m = x_0 \in \mathbb{R}$ ,  $\Sigma_{n+1}^{\Delta t}(X_n^{\Delta t}, n)$  is the (1, 1)-th entry of the Hessian of  $F^{\Delta t}$  (see (3-1) and (4-2)), and the  $\xi_n$  are independent standard normal random variables. Keeping in mind that  $\varphi_n(x, n) = x$  for all  $n$ , the above can be written as

$$X_{n+1}^{\Delta t} = X_n^{\Delta t} + \Delta t \frac{\varphi_{n+1}^{\Delta t}(X_n^{\Delta t}, n) - \varphi_n^{\Delta t}(X_n^{\Delta t}, n)}{\Delta t} + \sqrt{\Delta t} \sqrt{\varepsilon} \sqrt{\Sigma_{n+1}^{\Delta t}(X_n^{\Delta t}, n)} \xi_n. \tag{6-2}$$

Our goal in this subsection is to sketch an argument suggesting that as  $\Delta t \rightarrow 0$ , solutions of (6-2) converge weakly [19] to those of

$$dX_t = \dot{\varphi}_t(X_t, t) dt + \sqrt{\varepsilon} \sigma \cdot dB_t \tag{6-3}$$

with  $X_0 = x_0$ . Since we consider “continuous time” and “discrete time” cases, we mark the discrete time case by a  $\Delta t$  superscript (i.e., in this section, the function in (3-1) is called  $F^{\Delta t}$ ). In (6-3), “ $\dot{\varphi}_s(x, t)$ ” denotes  $\partial_s(\varphi_s(x, t))$ , and the path

$s \mapsto \varphi_s(x_0, t)$  ( $t \leq s \leq T$ ) minimizes the *action functional* [14]

$$F(x_{t:T} \mid x_t = x_0) = \frac{1}{2\sigma^2} \int_t^T (\dot{x}_s - f(x_s))^2 ds + g(x_T), \quad \varphi_t(x_0, t) = x_0. \quad (6-4)$$

This is the continuous time analog of (3-1).

Equation (6-3) was derived in [28] as the proposal for an importance sampling algorithm. This was later used in [29] for data assimilation in the small noise regime. We assume minimizers  $\varphi$  of the action functional are twice-differentiable in the time parameter and satisfy the Euler–Lagrange equations; this can be justified via standard results from the calculus of variations (see, e.g., Section 3.1 of [15]). In what follows, we also assume that the action functional has a single global minimum for all initial positions  $x$  and initial time  $t \in [0, T]$ . This *unique optimal paths* assumption (the continuous time analog of the unimodality of  $p(x)$ ) implies that  $\dot{\varphi}_t(x, t)$  is defined everywhere. Without unique optimal paths, any analysis will require more care; see, e.g., [28] and references therein for a discussion of these and related issues. The assumption is natural for linear systems with unimodal likelihood functions  $e^{-g/\varepsilon}$ , and may hold (approximately) in nonlinear systems when  $T$  is small.

We now sketch our argument. We begin by recalling that a numerical approximation of an SDE *converges weakly with weak order  $k$*  if, for all test functions  $\psi \in C^{k+1}$  with at most polynomial growth,

$$|\mathbb{E}(\psi(X_N^{\Delta t}) \mid X_0) - \mathbb{E}(\psi(X_T) \mid X_0)| = O(\Delta t^k) \quad (6-5)$$

as  $\Delta t \rightarrow 0$ . By standard results in the numerical analysis of SDEs, weak convergence is implied by “weak consistency” plus some mild polynomial growth conditions; see, e.g., Section 14.5 in [19] for details.

In this context, consistency means that the factors  $(\varphi_{n+1}^{\Delta t}(x, n) - \varphi_n^{\Delta t}(x, n))/\Delta t$  and  $\Sigma_{n+1}^{\Delta t}(x, n)$  in (6-2) approximate the corresponding factors in (6-3) ( $\dot{\varphi}_t(x, n\Delta t)$  and  $\sigma^2$ , respectively). These we now prove:

**Proposition.** *Under the unique optimal path assumption, we have*

$$\frac{\varphi_{n+1}^{\Delta t}(x, n) - \varphi_n^{\Delta t}(x, n)}{\Delta t} = \dot{\varphi}_{n\Delta t}(x, n\Delta t) + O(\Delta t) \quad \text{for all } n = 1, \dots, N \text{ and } x \in \mathbb{R}, \quad (6-6)$$

$$\Sigma_{n+1}^{\Delta t}(x, n) = \sigma^2 + O(\Delta t). \quad (6-7)$$

*Proof of (6-6).* We begin by proving that  $\varphi$  and  $\varphi^{\Delta t}$  satisfy the first variational equations for  $F$  and  $F^{\Delta t}$ , respectively (see (6-4) and (3-1)). Without loss of generality, set  $t = 0$  and  $n = 0$ , and write  $\varphi(s) := \varphi_s(x_0, 0)$  for a given  $x_0$ . Then

the first variational equation of  $F$  is the boundary value problem

$$-\ddot{\varphi}(s) + f'(\varphi(s))f(\varphi(s)) = 0, \tag{6-8}$$

$$\varphi(0) - x(0) = 0, \tag{6-9}$$

$$\dot{\varphi}(T) - f(\varphi(T)) + \sigma g'(\varphi(T)) = 0, \tag{6-10}$$

and the first variational equation for  $F^{\Delta t}$  is

$$\begin{aligned} -\frac{\varphi_{k-1}^{\Delta t} - 2\varphi_k^{\Delta t} + \varphi_{k+1}^{\Delta t}}{\Delta t^2} + f'(\varphi_k^{\Delta t})f(\varphi_k^{\Delta t}) \\ + \frac{f(\varphi_k^{\Delta t}) - f(\varphi_{k-1}^{\Delta t})}{\Delta t} - f'(\varphi_k^{\Delta t})\frac{\varphi_{k+1}^{\Delta t} - \varphi_k^{\Delta t}}{\Delta t} = 0, \end{aligned} \tag{6-11}$$

$$\varphi_0^{\Delta t} - x_0 = 0, \tag{6-12}$$

$$\frac{\varphi_N^{\Delta t} - \varphi_{N-1}^{\Delta t}}{\Delta t} - f(\varphi_{N-1}^{\Delta t}) + \sigma g'(\varphi_N^{\Delta t}) = 0. \tag{6-13}$$

By the unique optimal path assumption, (6-8) is well posed. Equation (6-8) is equivalent to the system

$$-\dot{v} + f'(\varphi)f(\varphi) = 0, \quad \dot{\varphi} = v \tag{6-14}$$

with boundary conditions  $\varphi(0) = 0$  and  $v(T) - f(\varphi(T)) + \sigma g'(\varphi(T)) = 0$ , and (6-11) is equivalent to the first-order-accurate finite difference approximation

$$\begin{aligned} -\frac{v_k - v_{k-1}}{\Delta t} + f'(\varphi_k^{\Delta t})f(\varphi_k^{\Delta t}) + \frac{f(\varphi_k^{\Delta t}) - f(\varphi_{k-1}^{\Delta t})}{\Delta t} - f'(\varphi_k^{\Delta t})v_k = 0, \\ v_k = \frac{\varphi_{k+1}^{\Delta t} - \varphi_k^{\Delta t}}{\Delta t}. \end{aligned}$$

Convergence results for numerical approximations of two-point boundary value problems tell us that for first-order-accurate finite difference schemes, pointwise errors are uniformly bounded by  $C\Delta t$  for some  $C > 0$  (see, e.g., [18] and references therein). In particular, we have  $(\varphi_{n+1}^{\Delta t} - \varphi_n^{\Delta t})/\Delta t = v_n = \dot{\varphi}(n\Delta t) + O(\Delta t)$  for each  $n$ , as claimed.  $\square$

*Proof of (6-7).* To prove (6-7), we consider the second variational equations of  $F$  and  $F^{\Delta t}$ . For  $F$ , we obtain a Sturm–Liouville boundary value problem

$$\begin{aligned} (Lu)(s) &= 0, \\ u(0) &= 0, \end{aligned} \tag{6-15}$$

$$u'(T) + (-f'(\varphi(s)) + \sigma g''(\varphi(T)))u(T) = 0,$$

where the operator  $L$  is defined by

$$Lu = -u''(s) + (f'(\varphi(s))^2 + f''(\varphi(s))f(\varphi(s)))u(s),$$

$\varphi$  is the solution to the first variational equation, and  $u$  is a test function. The second variational equation for  $F^{\Delta t}$  is

$$\begin{aligned} (H/\Delta t)u^{\Delta t} &= 0, \\ u_0^{\Delta t} &= 0, \\ \frac{u_N^{\Delta t} - u_{N-1}^{\Delta t}}{\Delta t} - f'(\varphi_{N-1}^{\Delta t})u_{N-1}^{\Delta t} + \sigma g''(\varphi_N^{\Delta t})u_N^{\Delta t} &= 0, \end{aligned}$$

where  $H$  is the Hessian of  $F^{\Delta t}$ , and

$$\begin{aligned} (H/\Delta t)u^{\Delta t} &= -\frac{u_{k-1}^{\Delta t} - 2u_k^{\Delta t} + u_{k+1}^{\Delta t}}{\Delta t^2} + (f'(\varphi_k^{\Delta t})^2 + f(\varphi_k^{\Delta t}) \cdot f''(\varphi_k^{\Delta t}))u_k^{\Delta t} \\ &+ \frac{f'(\varphi_k^{\Delta t})u_k^{\Delta t} - f'(\varphi_{k-1}^{\Delta t})u_{k-1}^{\Delta t}}{\Delta t} - \frac{u_{k+1}^{\Delta t} - u_k^{\Delta t}}{\Delta t} \cdot f'(\varphi_k^{\Delta t}) - \frac{\varphi_{k+1}^{\Delta t} - \varphi_k^{\Delta t}}{\Delta t} \cdot f''(\varphi_k^{\Delta t})u_k^{\Delta t}. \end{aligned}$$

Note that the discrete equations can also be obtained by applying a first-order discretization scheme to the continuous equations.

The differential operator  $L$  has an associated Green's function

$$K(t, s) = \frac{1}{y_1'(0)y_2(0)} \begin{cases} y_1(t)y_2(s), & 0 < t < s, \\ y_2(t)y_1(s), & 0 < s \leq t, \end{cases}$$

where  $y_1$  is a solution that satisfies the left Dirichlet boundary condition, while the solution  $y_2$  satisfies the mixed boundary condition on the right. The analog of the Green's function for the discretized problem is  $H^{-1}$ . Specifically, the first element of the first row of  $H^{-1}$  is a second-order approximation of the Green's function  $K(\Delta t, \Delta t)$ :

$$(H^{-1})_{1,1} = K(\Delta t, \Delta t) + O(\Delta t^2). \tag{6-16}$$

A Taylor expansion of  $K$  at the origin gives

$$K(\Delta t, \Delta t) = \sigma^2 \Delta t + \Delta t^2 \frac{y_2'(0)}{y_2(0)} + O(\Delta t^3). \tag{6-17}$$

Combined, we thus have

$$(H^{-1})_{1,1} = \sigma^2 \Delta t + O(\Delta t^2). \tag{6-18}$$

Since  $\Sigma_{n+1}^{\Delta t}(x, n) = (H^{-1})_{1,1}/\Delta t$ , this shows that  $\Sigma_{n+1}^{\Delta t}(x, n) = \sigma^2 + O(\Delta t)$ .  $\square$

**6.2. Small noise analysis for the continuous time limit of DLM.** We investigate how the efficiency of the dynamic linear map, as measured by the quantity  $Q$  (see (2-5)), is affected by taking the  $\Delta t \rightarrow 0$  limit, and apply the theory presented in [26] to show that  $Q$  scales linearly in the small noise parameter  $\varepsilon$  even as  $\Delta t \rightarrow 0$ .

First, we note that the weights of the continuous limit of the DLM follow from the Cameron–Martin–Girsanov theorem [14]

$$w(X) \propto \exp\left(-\frac{1}{\sqrt{\varepsilon}} \int_0^T v(X_s, s) \cdot dB_s - \frac{1}{2\varepsilon} \int_0^T v(X_s, s)^2 ds - \frac{1}{\varepsilon} g(X_T)\right) \quad (6-19)$$

where  $v(x, t) = \sigma^{-1} \cdot (\varphi'_t(x, t) - f(x))$ . The relative variance of the weights can be written as

$$Q = e^{-(V(0, x_0) - 2G(0, x_0))/\varepsilon} - 1, \quad (6-20)$$

where

$$G(x, t) = -\varepsilon \log(\mathbb{E}_q[w \mid x_t = x]),$$

$$V(x, t) = -\varepsilon \log(\mathbb{E}_q[(w)^2 \mid x_t = x]).$$

In [26], it was shown that  $V$  can be expanded in powers of  $\varepsilon$  when the minimizer  $\varphi$  of (6-4) is unique for all  $(x, t)$  in the domain. A calculation shows that  $G$  can also be expanded in powers of  $\varepsilon$ , with similar coefficients. In summary, we have

$$G(x, t) = G_0(x, t) + \varepsilon \cdot G_1(x, t) + \varepsilon^2 \cdot G_2(x, t) + O(\varepsilon^3), \quad (6-21)$$

$$V(x, t) = V_0(x, t) + \varepsilon \cdot V_1(x, t) + \varepsilon^2 \cdot V_2(x, t) + O(\varepsilon^3), \quad (6-22)$$

where the coefficients  $G_i, V_i, i = 0, 1, 2$ , satisfy the following system of PDEs:

$$\begin{aligned} \partial_t G_0 + f \partial_x G_0 - \frac{\sigma^2}{2} (\partial_x G_0)^2 &= 0, & G_0(x, T) &= g(x), \\ \partial_t V_0 + (f + \sigma^2 \partial_x G_0) \cdot \partial_x V_0 - \frac{\sigma^2}{2} (\partial_x V_0)^2 - \sigma^2 (\partial_x G_0)^2 &= 0, & V_0(x, T) &= 2g(x), \\ \partial_t G_1 + f \cdot \partial_x G_1 + \frac{\sigma^2}{2} \partial_{xx} G_0 - \sigma^2 \partial_x G_0 \cdot \partial_x G_1 &= 0, & G_1(x, T) &= 0, \\ \partial_t V_1 + (f + \sigma^2 \partial_x G_0) \cdot \partial_x V_1 + \frac{\sigma^2}{2} \partial_{xx} V_0 - \sigma^2 \partial_x V_0 \cdot \partial_x V_1 &= 0, & V_1(x, T) &= 0, \\ \partial_t G_2 + f \cdot \partial_x G_2 + \frac{\sigma^2}{2} \partial_{xx} G_1 - \sigma^2 \partial_x G_0 \cdot \partial_x G_2 \\ &\quad - \frac{\sigma^2}{2} (\partial_x G_1)^2 &= 0, & G_2(x, T) &= 0, \\ \partial_t V_2 + (f + \sigma^2 \partial_x G_0) \cdot \partial_x V_2 + \frac{\sigma^2}{2} \partial_{xx} V_1 - \sigma^2 \partial_x V_0 \cdot \partial_x V_2 \\ &\quad - \frac{\sigma^2}{2} (\partial_x V_1)^2 &= 0, & V_2(x, T) &= 0. \end{aligned}$$

(These equations are similar in structure to those of the WKB approximation [3], with the leading-order term given by a nonlinear PDE of Hamilton–Jacobi type and a hierarchy of linear transport equations for the higher-order terms.) One can check

that  $V_0 = 2G_0$  and  $V_1 = 2G_1$ , but  $V_2 \neq 2G_2$ . Combining the expansions (6-21) and (6-22) we thus have

$$V(x_0, 0) - 2G(x_0, 0) = \varepsilon^2 K_2 + O(\varepsilon^3), \quad (6-23)$$

where  $K_2 = V_2 - 2G_2$  satisfies

$$\partial_t K_2 + f \cdot \partial_x K_2 - \sigma^2 \partial_x G_0 \cdot \partial_x K_2 - \sigma^2 (\partial_x G_1)^2 = 0, \quad K_2(x, T) = 0. \quad (6-24)$$

Using (6-23) in the expression of the relative variance  $Q$  in (6-20), and expanding in  $\varepsilon$  results in

$$Q = \varepsilon \cdot K_2(x_0, 0) + O(\varepsilon^2). \quad (6-25)$$

Thus, the performance criterion  $Q$  for this continuous time method scales linearly with  $\varepsilon$ .

## 7. Concluding discussion

In this paper, we study a class of importance samplers for SDEs designed for data assimilation tasks in the small (observation and dynamic) noise regime. We have extended a small noise analysis for implicit samplers [17] to importance sampling for SDEs. We have also shown that a symmetrization procedure, originally proposed in [17], can be applied effectively to obtain higher-order samplers for SDEs. Moreover, we have shown that a dynamic version of the importance sampler retains the same asymptotic performance but is more robust in problems with multimodal distributions.

Our work also points to a number of directions for future research:

- (i) *Multimodal distributions.* Our analysis is limited to unimodal target distributions, but multimodal distributions do occur in practice. We believe an analysis for such problems (which necessarily means dealing with  $q(x_{n+1} | x_n)$  with jump discontinuities), possibly on concrete examples, would yield useful insights into the performance of DLM in more general situations than the ones examined here. One use for such an analysis is to compare DLM with other data assimilation methods, e.g., the ensemble Kalman filter, which may require less computation in nearly gaussian problems.
- (ii) *Continuous time limits.* In discrete time, the dimension of the sampling problem we consider is equal to the dimension of a discretized path of an SDE and, thus, equal to the product of the state dimension and the number of time steps of the path. Our continuous time limit of the DLM for scalar SDE indicates that a large dimension due to a small time step is unproblematic, but our results do not indicate how the efficiency of DLM degrades when the dimension of the SDE is large.

- (iii) *Symmetrization in continuous time.* Our results with symmetrized methods in discrete time are encouraging, but we currently do not have theoretical results on symmetrization in continuous time.
- (iv) *Long time scales.* As mentioned in Section 1, the methods discussed in this paper bear a close resemblance to methods proposed in [28] and [13] for rare event simulation. However, in this paper we have assumed a fixed final time  $T$ , whereas for many (if not most) rare event problems of interest, the relevant time scale tends to  $\infty$  as  $\varepsilon \rightarrow 0$  (e.g.,  $T = O(1/\varepsilon)$ ), and our methods are not expected to perform well on such long time scales. It would be of theoretical and practical interest to extend the ideas described here to the setting of rare event simulation, particularly the idea of symmetrization.
- (v) *Problems that do not come from SDEs.* Also mentioned in Section 1 is the possibility of extending the methods proposed here, in particular symmetrization, to more general sequential Monte Carlo sampling problems.

### Acknowledgments

Lin and Leach were supported in part by NSF grant DMS-1418775. Morzfeld was supported by NSF grant DMS-1619630, the Office of Naval Research (grant number N00173-17-2-C003), and the Alfred P. Sloan Foundation. The authors thank Professors Jonathan Goodman, Jonathan Weare, and Kostas Spiliopoulos for many helpful conversations and some of the references.

### References

- [1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart, *Importance sampling: intrinsic dimension and computational cost*, *Statist. Sci.* **32** (2017), no. 3, 405–431. MR
- [2] S. Asmussen and P. W. Glynn, *Stochastic simulation: algorithms and analysis*, *Stochastic Modelling and Applied Probability*, no. 57, Springer, 2007. MR Zbl
- [3] C. M. Bender and S. A. Orszag, *Advanced mathematical methods for scientists and engineers, I: Asymptotic methods and perturbation theory*, Springer, 1999. MR Zbl
- [4] N. Bergman, *Recursive Bayesian estimation: navigation and tracking applications*, Ph.D. thesis, Linköping University, 1999.
- [5] M. Bocquet, C. A. Pires, and L. Wu, *Beyond Gaussian statistical modeling in geophysical data assimilation*, *Mon. Weather Rev.* **138** (2010), no. 8, 2997–3023.
- [6] A. Chorin, M. Morzfeld, and X. Tu, *Implicit particle filters for data assimilation*, *Commun. Appl. Math. Comput. Sci.* **5** (2010), no. 2, 221–240. MR Zbl
- [7] A. J. Chorin and O. H. Hald, *Stochastic tools in mathematics and science*, 3rd ed., *Texts in Applied Mathematics*, no. 58, Springer, 2013. MR Zbl
- [8] A. Doucet, N. de Freitas, and N. Gordon (eds.), *Sequential Monte Carlo methods in practice*, Springer, 2001. MR

- [9] P. Dupuis, K. Spiliopoulos, and H. Wang, *Rare event simulation for rough energy landscapes*, Proceedings of the 2011 Winter Simulation Conference (S. Jain, R. Creasey, and J. Himmelspach, eds.), IEEE, 2011, pp. 504–515.
- [10] ———, *Importance sampling for multiscale diffusions*, Multiscale Model. Simul. **10** (2012), no. 1, 1–27. MR Zbl
- [11] P. Dupuis, K. Spiliopoulos, and X. Zhou, *Escaping from an attractor: importance sampling and rest points, I*, Ann. Appl. Probab. **25** (2015), no. 5, 2909–2958. MR Zbl
- [12] P. Dupuis and H. Wang, *Importance sampling, large deviations, and differential games*, Stoch. Stoch. Rep. **76** (2004), no. 6, 481–508. MR Zbl
- [13] ———, *Subsolutions of an Isaacs equation and efficient schemes for importance sampling*, Math. Oper. Res. **32** (2007), no. 3, 723–757. MR Zbl
- [14] M. I. Freidlin and A. D. Wentzell, *Random perturbations of dynamical systems*, Grundlehren der Mathematischen Wissenschaften, no. 260, Springer, 1984. MR Zbl
- [15] M. Giaquinta and S. Hildebrandt, *Calculus of variations, I: The Lagrangian formalism*, Grundlehren der Mathematischen Wissenschaften, no. 310, Springer, 1996. MR
- [16] C. Gissinger, *A new deterministic model for chaotic reversals*, Eur. Phys. J. B **85** (2012), no. 4, 137.
- [17] J. Goodman, K. K. Lin, and M. Morzfeld, *Small-noise analysis and symmetrization of implicit Monte Carlo samplers*, Comm. Pure Appl. Math. **69** (2016), no. 10, 1924–1951. MR Zbl
- [18] H. B. Keller, *Numerical methods for two-point boundary value problems*, Dover, 1992. MR
- [19] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, Applications of Mathematics (New York), no. 23, Springer, 1992. MR Zbl
- [20] A. Leach, *Monte Carlo methods for stochastic differential equations and their applications*, Ph.D. thesis, University of Arizona, 2017. MR
- [21] J. S. Liu and R. Chen, *Sequential Monte Carlo methods for dynamic systems*, J. Amer. Statist. Assoc. **93** (1998), no. 443, 1032–1044. MR Zbl
- [22] L. Martino, V. Elvira, and F. Louzada, *Effective sample size for importance sampling based on discrepancy measures*, Signal Process. **131** (2017), 386–401.
- [23] M. Morzfeld, X. Tu, E. Atkins, and A. J. Chorin, *A random map implementation of implicit filters*, J. Comput. Phys. **231** (2012), no. 4, 2049–2066. MR
- [24] E. I. Ostrovskii, *Exact asymptotics of Laplace integrals for nonsmooth functions*, Math. Notes **73** (2003), no. 6, 838–842. MR Zbl
- [25] C. Snyder, T. Bengtsson, and M. Morzfeld, *Performance bounds for particle filters using the optimal proposal*, Mon. Weather Rev. **143** (2015), no. 11, 4750–4761.
- [26] K. Spiliopoulos, *Nonasymptotic performance analysis of importance sampling schemes for small noise diffusions*, J. Appl. Probab. **52** (2015), no. 3, 797–810. MR Zbl
- [27] P. J. van Leeuwen, *Particle filtering in geophysical systems*, Mon. Weather Rev. **137** (2009), no. 12, 4089–4144.
- [28] E. Vanden-Eijnden and J. Weare, *Rare event simulation of small noise diffusions*, Comm. Pure Appl. Math. **65** (2012), no. 12, 1770–1803. MR Zbl
- [29] ———, *Data assimilation in the low noise regime with application to the Kuroshio*, Mon. Weather Rev. **141** (2013), no. 6, 1822–1841.

Received July 7, 2017. Revised March 7, 2018.



ANDREW LEACH: [imaleach@gmail.com](mailto:imaleach@gmail.com)

*Program in Applied Mathematics, University of Arizona, Tucson, AZ, United States*

*Current address: Google, Mountain View, CA, United States*

KEVIN K. LIN: [klin@math.arizona.edu](mailto:klin@math.arizona.edu)

*Program in Applied Mathematics, University of Arizona, Tucson, AZ, United States*

and

*Department of Mathematics, University of Arizona, Tucson, AZ, United States*

MATTHIAS MORZFELD: [mmo@math.arizona.edu](mailto:mmo@math.arizona.edu)

*Program in Applied Mathematics, University of Arizona, Tucson, AZ, United States*

and

*Department of Mathematics, University of Arizona, Tucson, AZ, United States*



## EFFICIENT HIGH-ORDER DISCONTINUOUS GALERKIN COMPUTATIONS OF LOW MACH NUMBER FLOWS

JONAS ZEIFANG, KLAUS KAISER, ANDREA BECK,  
JOCHEN SCHÜTZ AND CLAUS-DIETER MUNZ

We consider the efficient approximation of low Mach number flows by a high-order scheme, coupling a discontinuous Galerkin (DG) discretization in space with an implicit/explicit (IMEX) discretization in time. The splitting into linear implicit and nonlinear explicit parts relies heavily on the incompressible solution. The method has been originally developed for a singularly perturbed ODE and applied to the isentropic Euler equations. Here, we improve, extend, and investigate the so-called RS-IMEX splitting method. The resulting scheme can cope with a broader range of Mach numbers without running into roundoff errors, it is extended to realistic physical boundary conditions, and it is shown to be highly efficient in comparison to more standard solution techniques.

### 1. Introduction

Computing solutions to singularly perturbed problems can be cumbersome and expensive due to their multiscale nature. However, they do often occur in physical reality. A typical example is the transition from the compressible to the incompressible Navier–Stokes equations that constitutes a singularly perturbed limit [29; 41; 49]. Another example is multiphase flows, in which small liquid droplets are suspended in a gaseous phase. In such problems, the Mach number  $\varepsilon$  — the local ratio of flow speed to the speed of sound — can vary by orders of magnitude. In particular, some parts are very close to the incompressible regime, meaning that the Mach number is close to zero, while in other parts,  $\varepsilon$  is of the order of one. Flows of this nature are sometimes called *all-speed* flows.

Besides some issues such as the high-order treatment of the ubiquitous shocks or the treatment of turbulence, the efficient discretization of the  $\varepsilon = \mathcal{O}(1)$  case is by now rather well understood; see, e.g., [47] and the references therein. In this work, we therefore focus on the efficient discretization of the  $\varepsilon \ll 1$  case as a milestone towards *all-speed* schemes.

---

MSC2010: 35L65, 65N30.

Keywords: discontinuous Galerkin, IMEX-Runge–Kutta, low Mach number, splitting, asymptotic preserving.

The first idea that comes to mind is to treat the low Mach case as incompressible. In many situations however, compressible effects matter even in the vicinity of the incompressible regime. An example is given by the computation of transcritical droplets in a surrounding (supercritical) gas phase, where a strong coupling of thermodynamics and hydrodynamics in the droplets occurs. This phase state is also called the “compressible liquid” state and of current research interest [27; 26; 44]. Other situations occur in meteorological flows, where density gradients have to be considered but acoustic waves do not have to be resolved [14], and situations with strong temperature gradients but low velocities, e.g., natural convection [35]. As a consequence, we therefore stick to solution procedures for the compressible equations. The incompressible equations will nonetheless serve as a building block in our discretization.

Computing solutions to the low Mach equations ( $\varepsilon \ll 1$ ) using classical discretization paradigms which mostly rely on *explicit* time stepping methods leads to the unwanted encounter of having to choose an extremely small time step size ( $\Delta t \lesssim \varepsilon \Delta x$ ) to obtain a stable algorithm. Furthermore, due to the excessive amount of numerical viscosity that is added to the approximate solution, the explicit method may yield an incorrect solution [38]. One remedy is to use an implicit-explicit (IMEX) time stepping method [2; 28; 10] that relies on a splitting of the flux functions into a stiff part, which accounts for the singularity in the problem and is treated implicitly, and a nonstiff part, which only has a mild dependency on  $\varepsilon$ , and not on  $\varepsilon^{-1}$ . A number of splittings have been developed over the past few years; see, e.g., [30; 9; 13; 21]. All of those splittings have the disadvantage that it is very difficult to adapt them to other physical situations at hand, because they are developed for a specific set of equations.

To circumvent this problem, Kaiser et al. have introduced a general splitting and have applied it to the isentropic Euler equations in [25] that is based on the incompressible limit solution (called the *reference solution*); the splitting was hence termed *RS-IMEX*. The RS-IMEX idea is conceptually similar to the one introduced in [42] where the underlying problem is a singularly perturbed ODE. Related ideas have already been published earlier in [16; 9; 19]. One of the advantages of the splitting is that its idea, at least from a conceptual point of view, is independent from the underlying singularly perturbed problem and thereby not specific to a fixed system of equations. Furthermore, the implicit part is always linear in the solution variable, which usually reduces the time-to-solution tremendously, as the resulting algebraic equations are then also linear. Those linear equations are usually solved through a Krylov-type iteration method, which means that only matrix-vector products are needed, where the Jacobian is being frequently approximated via finite differences. For a nonlinear operator at low  $\varepsilon$ , this can pose severe problems for the approximation quality. However, as the implicit part of the RS-IMEX is linear,

the finite differences are not approximations but exact. In [24], the splitting has been used within a high-order IMEX discontinuous Galerkin (DG) solver and it has been shown that the algorithm preserves the asymptotics of the problem, which means that for  $\varepsilon \rightarrow 0$ , the discrete solution converges towards a discretization of the incompressible Euler equations. The latter is an important property as it means that no spurious effects stemming from the discretization and polluting the solution are introduced for small Mach numbers.

The purpose of this work is to improve, extend, and investigate the method introduced in [24] towards engineering problems.

- We improve the scheme for very small Mach numbers to suffer less from roundoff errors [33]. This is achieved through a reformulation of the method in a perturbed variable, partly following the lines of [43]. Thereby, we alleviate the dependency of the method on  $\varepsilon^{-1}$  to a great extent, which is the core source of roundoff problems.
- We extend the scheme to cope with more realistic physical settings by adding appropriate boundary conditions ([24] works with periodic ones) and considering three dimensions.
- We investigate the scheme with respect to runtime in the framework of a modern parallel architecture solver and compare it against other methods. For solving the algebraic system, we take advantage of the linearity of RS-IMEX in the solution process. We demonstrate the advantages of this novel method with respect to runtime and accuracy as a function of  $\varepsilon$  for nontrivial test cases.

The paper is structured as follows. In Section 2 we introduce the underlying isentropic Euler equations. Section 3 introduces the RS-IMEX splitting for those equations; subsequently, in Section 4, the discontinuous Galerkin discretization including the IMEX time discretization is detailed. In Section 5, we validate the method through a manufactured solution. Furthermore, we explain in detail how to circumvent a problem with machine accuracy due to low Mach numbers  $\varepsilon$ . In Section 6 we present more involved numerical examples and discussions concerning accuracy and efficiency in the low Mach number case. Finally, in Section 7 we offer conclusion and outlook.

## 2. Equations

In this work we consider the isentropic Euler equations on a domain  $\Omega \subset \mathbb{R}^d$ . Nondimensionalized, those equations are given by

$$\partial_t \mathbf{w} + \nabla_x \cdot \mathbf{F}(\mathbf{w}) = 0 \quad \text{with } \mathbf{w} := \begin{pmatrix} \rho \\ \rho \mathbf{u} \end{pmatrix} \text{ and } \mathbf{F}(\mathbf{w}) := \begin{pmatrix} \rho \mathbf{u} \\ \rho \mathbf{u} \otimes \mathbf{u} + (1/\varepsilon^2) p(\rho) \cdot \mathbf{Id} \end{pmatrix} \quad (1)$$

with  $\mathbf{u}$  and  $\rho$  denoting velocity and density, respectively. The subscripts  $t$  and  $x$  denote temporal and spatial derivatives, respectively. The reference Mach number, defined as

$$\varepsilon := \frac{u^*}{\sqrt{p(\rho^*)/\rho^*}}$$

with  $u^*$  and  $\rho^*$  reference values used in the nondimensionalization process, is a measure for the compressibility of the system. The pressure  $p$  is defined by the equation of state

$$p(\rho) = \kappa \rho^\gamma, \tag{2}$$

with  $\gamma \geq 1$  the isentropic coefficient and  $\kappa > 0$  being a constant.

The eigenvalues of  $\frac{\partial}{\partial \mathbf{w}} \mathbf{F}(\mathbf{w}) \cdot \mathbf{n}$  are, for  $\Omega \subset \mathbb{R}^3$  and with  $c := \sqrt{\gamma p/\rho}$  being the speed of sound, given by

$$\lambda_{1,2} = \mathbf{u} \cdot \mathbf{n}, \quad \lambda_{3,4} = \mathbf{u} \cdot \mathbf{n} \pm \frac{c}{\varepsilon}. \tag{3}$$

It is obvious that for  $\varepsilon \ll 1$  (the low Mach case) those waves have extremely different speeds, i.e., wave speeds are in  $\mathcal{O}(1)$  and  $\mathcal{O}(\varepsilon^{-1})$ . In the limit  $\varepsilon \rightarrow 0$  two wave speeds tend to infinity, meaning that the associated hyperbolic equation degenerates. This means that some information travels infinitely fast, and some at finite speed. Besides that, one can show that under certain conditions [29], the compressible equations (1) transform towards the incompressible Euler equations, which are given by

$$\begin{aligned} \partial_t \begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix} \mathbf{w}_{(0)} + \nabla_x \cdot \mathbf{G}(\mathbf{w}_{(0)}, p_{(2)}) &= 0 \quad \text{and} \quad \rho_{(0)} \equiv \text{const} \\ \text{with} \quad \mathbf{w}_{(0)} &:= \begin{pmatrix} \rho_{(0)} \\ (\rho \mathbf{u})_{(0)} \end{pmatrix} \\ \text{and} \quad \mathbf{G}(\mathbf{w}_{(0)}, p_{(2)}) &:= \begin{pmatrix} (\rho \mathbf{u})_{(0)} \\ (\rho \mathbf{u})_{(0)} \otimes (\rho \mathbf{u})_{(0)} / \rho_{(0)} + p_{(2)} \cdot \mathbf{Id} \end{pmatrix}. \end{aligned} \tag{4}$$

The relation between the compressible and the incompressible equations can be understood best if we assume that every quantity of (1) can be represented by an asymptotic expansion, e.g.,

$$\rho = \rho_{(0)} + \varepsilon \rho_{(1)} + \varepsilon^2 \rho_{(2)} + \mathcal{O}(\varepsilon^3),$$

and compute the formal limit  $\varepsilon \rightarrow 0$ . The incompressible equations (4) are then obtained under the assumption of *well prepared* initial and boundary conditions; see, e.g., [21] for a detailed computation and [29] for a more formal proof.

**Definition** (well prepared initial and boundary conditions). We call initial conditions *well prepared* if they possess an asymptotic expansion in positive powers of  $\varepsilon$ ,

and

$$\rho(t=0) = \text{const} + \mathcal{O}(\varepsilon^2) \quad \text{and} \quad \nabla_x \cdot (\rho \mathbf{u})(t=0) = \mathcal{O}(\varepsilon).$$

We call boundary conditions *well prepared* if they ensure

$$\int_{\partial\Omega} (\rho \mathbf{u})|_{\partial\Omega} \cdot \mathbf{n} = 0 \quad \text{and} \quad \rho|_{\partial\Omega} = \text{const} + \mathcal{O}(\varepsilon^2).$$

For *well prepared* initial conditions the corresponding incompressible state can be calculated by setting  $\varepsilon = 0$ . Consequently, incompressible density is the constant value  $\rho_{(0)}$  and the velocity field is  $\mathbf{u}_{(0)}$ . To initialize  $p_{(2)}$  we compute

$$p = p(\rho_{(0)}) + p'(\rho_{(0)})\varepsilon^2 \rho_{(2)} + \mathcal{O}(\varepsilon^4) = p_{(0)} + \varepsilon^2 p_{(2)} + \mathcal{O}(\varepsilon^4).$$

After reformulation one obtains

$$p_{(2)} = p'(\rho_{(0)})\rho_{(2)} = \kappa \gamma (\rho_{(0)})^{\gamma-1} \rho_{(2)} = \frac{1}{\varepsilon^2} \kappa \gamma (\rho_{(0)})^{\gamma-1} (\rho - \rho_{(0)}). \quad (5)$$

### 3. RS-IMEX

In this section, we explain the basic ideas of the RS-IMEX splitting of the isentropic Euler equations for nearly incompressible flows; more details of the final algorithm are given in Section 4. Previously in Section 2, we have seen that the isentropic Euler equations (1) in the low Mach regime transform to the incompressible Euler equations (4) if  $\varepsilon \rightarrow 0$ . A proper numerical method should be designed in such a way that it can imitate this behavior; i.e., for  $\varepsilon \rightarrow 0$  the method should formally transform into a discretization of the incompressible equations. This is the so-called *asymptotic preserving* property; see, e.g., [23].

One way to handle this type of equation is to split the flux function  $F$  into two parts and treat one part  $\widehat{F}$  with an explicit and the other part  $\widetilde{F}$  with an implicit method:

$$\partial_t \mathbf{w} + \nabla_x \cdot (\widetilde{F}(\mathbf{w}) + \widehat{F}(\mathbf{w})) = 0. \quad (6)$$

This technique results in IMEX time integration schemes; see, e.g., [2; 28] and Section 4. For stability, efficiency, and accuracy it is important to find a suitable splitting. One splitting developed in the past years is the so-called RS-IMEX, where RS stands for *reference solution*. This splitting fulfills the asymptotic preserving property in the setting of low- and high-order discretizations for the isentropic Euler equations [25; 24]. Furthermore, it gave promising results for different types of equations in the sense of stability [24; 50], efficiency [25], and accuracy [25; 24].

**Definition** (RS-IMEX splitting). The RS-IMEX splitting is defined by

$$\widetilde{F}(\mathbf{w}) = F(\mathbf{w}_{(0)}) + F'(\mathbf{w}_{(0)})(\mathbf{w} - \mathbf{w}_{(0)}) \quad \text{and} \quad \widehat{F} = F(\mathbf{w}) - \widetilde{F}(\mathbf{w})$$

for a given reference solution  $\mathbf{w}_{(0)}$  and  $\mathbf{F}'(\mathbf{w}_{(0)})$  the Jacobian of the flux:

$$\mathbf{F}'(\mathbf{w}_{(0)}) = \frac{\partial \mathbf{F}(\mathbf{w}_{(0)})}{\partial \mathbf{w}_{(0)}}.$$

In general one could choose an arbitrary reference solution, but in the following we use the limit  $\mathbf{w}_{(0)} = \lim_{\varepsilon \rightarrow 0} \mathbf{w}$ , which corresponds to the solution of the incompressible equation. Applying the splitting to the isentropic Euler equations gives the implicit and explicit parts

$$\begin{aligned} \tilde{\mathbf{F}} &= \begin{pmatrix} \rho \mathbf{u} \\ -\rho \mathbf{u}_{(0)} \otimes \mathbf{u}_{(0)} + \rho \mathbf{u} \otimes \mathbf{u}_{(0)} + \rho \mathbf{u}_{(0)} \otimes \mathbf{u} + \frac{(p(\rho_{(0)}) + p'(\rho_{(0)})(\rho - \rho_{(0)}))}{\varepsilon^2} \cdot \mathbf{Id} \end{pmatrix}, \\ \hat{\mathbf{F}} &= \begin{pmatrix} 0 \\ \rho(\mathbf{u} - \mathbf{u}_{(0)}) \otimes (\mathbf{u} - \mathbf{u}_{(0)}) + \frac{p(\rho) - p(\rho_{(0)}) - p'(\rho_{(0)})(\rho - \rho_{(0)})}{\varepsilon^2} \cdot \mathbf{Id} \end{pmatrix}. \end{aligned}$$

Considering the  $\mathcal{O}(\varepsilon^{-2})$  terms, one obtains one motivation for the chosen reference solution. Since  $\rho - \rho_{(0)} = \mathcal{O}(\varepsilon)$  one obtains

$$p(\rho) - p(\rho_{(0)}) - p'(\rho_{(0)})(\rho - \rho_{(0)}) = \mathcal{O}(\varepsilon^2).$$

So upon inserting the exact solution, there are no stiff terms remaining in the explicit part. Of course this is only a rationale that stands behind the scheme. Computing the eigenvalues of the nonstiff part explicitly, one obtains

$$\hat{\lambda}_{1,2} = (\mathbf{u} - \mathbf{u}_{(0)}) \cdot \mathbf{n}, \quad \hat{\lambda}_3 = 0, \quad \text{and} \quad \hat{\lambda}_4 = 2(\mathbf{u} - \mathbf{u}_{(0)}) \cdot \mathbf{n},$$

and indeed, these eigenvalues are  $\mathcal{O}(1)$ . Even more, upon inserting the exact solution, they would be in  $\mathcal{O}(\varepsilon)$  for this given choice of the reference solution. Fast waves are solely solved with the implicit method, which is a core requirement for unconditional stability with respect to  $\varepsilon$ .

**Remark.** A similar technique has been used for the stiff collision operator of kinetic equations in [16] and for the pressure gradient of shallow-water equations in [9; 20].

### 4. Discretization

**4.1. High-order discontinuous Galerkin IMEX framework.** Discontinuous Galerkin (DG) schemes have recently gained considerable interest as baseline schemes for multiscale problems; see, e.g., [5; 3; 11; 47] and the references therein. They can be interpreted as a hybrid finite element–finite volume formulation, where an elementwise Galerkin variational formulation is coupled weakly to its neighbors through a numerical flux term. Each inner-element solution is approximated by a polynomial function of given order  $\mathcal{N}$ . Penalization of discontinuities and the



locality of the basis make DG suitable for hyperbolic problems. In addition, the compact operator with small memory and communication footprint leads to excellent parallel scaling properties and the element-based approximation enables unstructured meshing of complex domains.

To obtain a DG discretization, we assume that the domain is separated into a finite number of independent cells. Then we seek a piecewise smooth function  $\mathbf{w}_h$ ; i.e., it is a polynomial of maximal degree  $\mathcal{N}$  on every cell, which fulfills the weak discontinuous Galerkin formulation, given by

$$\frac{\partial}{\partial t} \int_E \mathbf{w}_h \phi(\mathbf{x}) \, d\mathbf{x} + \oint_{\partial E} \mathbf{F}_n^* \phi(\mathbf{x}) \, ds - \int_E \mathbf{F}(\mathbf{w}_h) \cdot \nabla_x \phi(\mathbf{x}) \, d\mathbf{x} = 0, \quad (7)$$

on every cell  $E$  for every polynomial test function  $\phi(\mathbf{x})$  of maximal degree  $\mathcal{N}$ . Note that  $\mathbf{F}_n^*$  denotes the surface normal numerical flux function, given by  $\mathbf{F}_n^* := \mathbf{F}_n^*(\mathbf{w}_h^+, \mathbf{w}_h^-)$  and superscripts  $\pm$  denote the values at the grid cell interface from the neighbor and the local grid cell  $E$ , respectively.

The current investigations are based on a particularly efficient variant of the general DG formulation (7), namely the discontinuous Galerkin spectral element method (DGSEM) proposed by [32]. In this formulation, the solution  $\mathbf{w}$  is approximated as a tensor product of one-dimensional Lagrange interpolating polynomials of degree  $\mathcal{N}$ . The  $\mathcal{N} + 1$  Legendre–Gauss quadrature points  $\{\xi_i\}_{i=0}^{\mathcal{N}}$  are chosen as interpolation nodes. This collocation of interpolation and integration nodes significantly reduces the number of operations per degree of freedom. In particular, the tensor product structure of the solution ansatz transfers to the operator itself, avoiding element–global volume operations. Instead, the multidimensional operator is constructed of consecutive one-dimensional operations. One disadvantage is that this reduces the flexibility of DG with respect to meshing, as only quadrilateral meshes can be used in order not to destroy the tensor product structure.

Details on the implementation and efficiency of the solver are given by Hindenlang et al. [22]. Extension of the framework to include multiphase flow based on a sharp interface approach, large eddy simulation methods, and shock capturing strategies are given by [15; 17; 7; 45]. The full FLEXI framework, including pre- and postprocessing tools, is available as open source software.<sup>1</sup>

For the extension of this solver to an implicit-explicit time discretization, we consider again a splitting as in (6). (Note that, with  $\tilde{\mathbf{F}}(\mathbf{w}) = \mathbf{F}(\mathbf{w})$  and  $\hat{\mathbf{F}}(\mathbf{w}) = 0$ , also a fully implicit scheme falls into this framework.) IMEX schemes are defined by their Butcher tableaux featuring the coefficients  $\tilde{A}$ ,  $\hat{A}$ ,  $\tilde{c}$ , and  $\hat{c}$ . In semidiscrete form the implicit-explicit Runge–Kutta time discretization for the  $i$ -th stage and

<sup>1</sup><https://www.flexi-project.org>, GNU General Public License v3.0.

the  $n$ -th time step can be written as

$$\mathbf{w}^{n,i} - \mathbf{w}^n + \Delta t \left( \sum_{j=1}^i \tilde{A}_{i,j} \nabla_x \cdot \tilde{\mathbf{F}}(\mathbf{w}^{n,j}, t^n + \tilde{c}_j \Delta t) + \sum_{j=1}^{i-1} \hat{A}_{i,j} \nabla_x \cdot \hat{\mathbf{F}}(\mathbf{w}^{n,j}, t^n + \hat{c}_j \Delta t) \right) = 0. \quad (8)$$

A reformulation of (8) yields

$$\begin{aligned} & \mathbf{w}^{n,i} + \Delta t \tilde{A}_{i,i} \nabla_x \cdot \tilde{\mathbf{F}}(\mathbf{w}^{n,i}, t^n + \tilde{c}_i \Delta t) \\ &= \mathbf{w}^n - \Delta t \sum_{j=1}^{i-1} [\tilde{A}_{i,j} \nabla_x \cdot \tilde{\mathbf{F}}(\mathbf{w}^{n,j}, t^n + \tilde{c}_j \Delta t) + \hat{A}_{i,j} \nabla_x \cdot \hat{\mathbf{F}}(\mathbf{w}^{n,j}, t^n + \hat{c}_j \Delta t)], \end{aligned}$$

where the right-hand side is either known from previous stages or can be computed explicitly. In the following, this equation is abbreviated by

$$(\mathbf{Id} - \Delta t \tilde{A}_{i,i} \tilde{\mathbf{R}}) \mathbf{w}^{n,i} = \mathbf{b},$$

with  $\tilde{\mathbf{R}}$  denoting the spatial operator with the implicitly treated fluxes. To solve this potentially nonlinear (for a fully implicit scheme it is; for the RS-IMEX it is linear!) system, a standard root finding algorithm such as Newton's method can be applied. Therefore, the IMEX-Runge–Kutta scheme for the  $k$ -th Newton's iteration reads

$$\begin{aligned} & \mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \Delta \mathbf{w}, \\ & \Delta \mathbf{w} - \Delta t \tilde{A}_{i,i} \frac{\partial \tilde{\mathbf{R}}(\mathbf{w}^{(k)})}{\partial \mathbf{w}} \Delta \mathbf{w} = -\mathbf{w}^{(k)} + \Delta t \tilde{A}_{i,i} \tilde{\mathbf{R}}(\mathbf{w}^{(k)}) + \mathbf{b}. \end{aligned} \quad (9)$$

For ease of presentation, we have omitted the superscript  $n, i$ .

Equation (9) is a linear system for every Newton iteration  $k$ , which can be solved with a standard linear solving algorithm. To minimize computational costs for calculating and storing the Jacobian, the matrix-free GMRES linear solving algorithm by Saad and Schultz [40] is applied. In [18] it has been shown that a matrix-free approach can be superior to a matrix-based approach for a high-order three-dimensional DG scheme. The matrix-vector product including the Jacobian in (9) is approximated via a finite difference

$$\frac{\partial \tilde{\mathbf{R}}(\mathbf{w}^{(k)})}{\partial \mathbf{w}} \Delta \mathbf{w} \approx \frac{\tilde{\mathbf{R}}(\mathbf{w}^{(k)} + \Delta_{\text{FD}} \Delta \mathbf{w}) - \tilde{\mathbf{R}}(\mathbf{w}^{(k)})}{\Delta_{\text{FD}}} \quad (10)$$

for a small  $\Delta_{\text{FD}}$  which can be calculated according to Qin et al. [37] and Knoll and Keynes [31] as

$$\Delta_{\text{FD}} = \frac{\sqrt{\text{eps}}}{\|\Delta \mathbf{w}\|_2},$$

with  $\text{eps}$  being the machine accuracy or the maximum achievable accuracy. As the implicit flux of the RS-IMEX splitting is linear, this finite difference can be simplified, but special care has to be taken of Dirichlet-type boundary conditions. Hence, for the RS-IMEX splitting, the matrix-vector product can be simplified to

$$\frac{\partial \tilde{\mathbf{R}}(\mathbf{w}^{(k)})}{\partial \mathbf{w}} \Delta \mathbf{w} = \tilde{\mathbf{R}}(\Delta \mathbf{w}) - \tilde{\mathbf{R}}(\mathbf{0}). \quad (11)$$

In our implementation, we use a standard block-Jacobian preconditioner due to the small building and storing costs of the preconditioner. This turned out to be beneficial for a DG setup with a very large number of processors [8]. In [8] it is shown that more sophisticated preconditioners like full SGS and multilevel preconditioners are not superior to the standard block-Jacobian preconditioner regarding computational time for a parallel DG setup.

**4.2. Incompressible solver.** The RS-IMEX splitting, defined on page 247, requires the corresponding incompressible state. Therefore, an incompressible solver in the discontinuous Galerkin framework is needed. We start with the incompressible Euler equations as given in (4) in three dimensions and reformulate them as

$$\partial_t \begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix} \mathbf{U} + \nabla_x \cdot \tilde{\mathbf{G}}(\mathbf{U}) = 0,$$

for the state vector  $\mathbf{U} = (p_{(2)}, u_{(0),1}, u_{(0),2}, u_{(0),3})^T$  and with the flux

$$\tilde{\mathbf{G}}(\mathbf{U}) := \begin{pmatrix} \mathbf{u}_{(0)} \\ \mathbf{u}_{(0)} \otimes \mathbf{u}_{(0)} + (p_{(2)}/\rho_{(0)}) \cdot \mathbf{Id} \end{pmatrix}.$$

Note as a reminder that  $p_{(2)}$  denotes the hydrodynamic pressure,  $\rho_{(0)}$  is a constant positive value, and  $\mathbf{u}_{(0)} = (u_{(0),1}, u_{(0),2}, u_{(0),3})^T$  denotes the three-dimensional velocity vector. As the divergence-free condition for the velocity field is not a time evolution equation for the hydrodynamic pressure, a numerical flux function is required to couple the velocity and pressure field. A flux which satisfies this condition for solving incompressible flows with a discontinuous Galerkin scheme has been proposed by Bassi et al. [4]. In order to obtain a flux at the interfaces, artificial compressibility is added for the solution of the Riemann problem. An iterative Godunov-type Riemann solver is used in [4] to obtain the interface fluxes. Following [6] this artificial compressibility approach allows an equal-order discretization for the pressure and velocity. Moreover it is shown to be a consistent discretization of the incompressible Euler equations as the added compressibility is zero for vanishing jumps at the cell interfaces. A further advantage of this approach is that it offers the possibility to use the same high-order numerical methods for space and time discretization as for the compressible method. As the accuracy of the incompressible reference solution for the RS-IMEX splitting is not crucial, we

use a cheaper Lax–Friedrichs-type Riemann solver motivated by the asymptotic analysis, which reads

$$\tilde{\mathbf{G}}^* = \frac{1}{2} \left( \tilde{\mathbf{G}}(\mathbf{U}^+) + \tilde{\mathbf{G}}(\mathbf{U}^-) + \mathbf{Diag} \left( \frac{\rho^{(0)^{1-\gamma}}}{\kappa^\gamma}, 1, 1, 1 \right) (\mathbf{U}^+ - \mathbf{U}^-) \right),$$

with  $\kappa$  and  $\gamma$  from the equation of state (2).

## 5. Validation and reformulation

In this section, we validate the code and give first impressions of its performance. Furthermore, we indicate how to avoid problems with machine accuracy when  $\varepsilon$  is very small. As an underlying example, we use the two-dimensional smooth traveling vortex presented in [24]. For both compressible and incompressible isentropic Euler equations, the solution is a mere transport of an initial vortex in the  $x_1$  direction with speed 0.5. The compressible solution reads

$$\rho(\mathbf{x}, t) = \rho \left( \begin{pmatrix} x_1 - 0.5t \\ x_2 \end{pmatrix}, 0 \right), \quad \mathbf{u}(\mathbf{x}, t) = \mathbf{u} \left( \begin{pmatrix} x_1 - 0.5t \\ x_2 \end{pmatrix}, 0 \right),$$

for the initial conditions

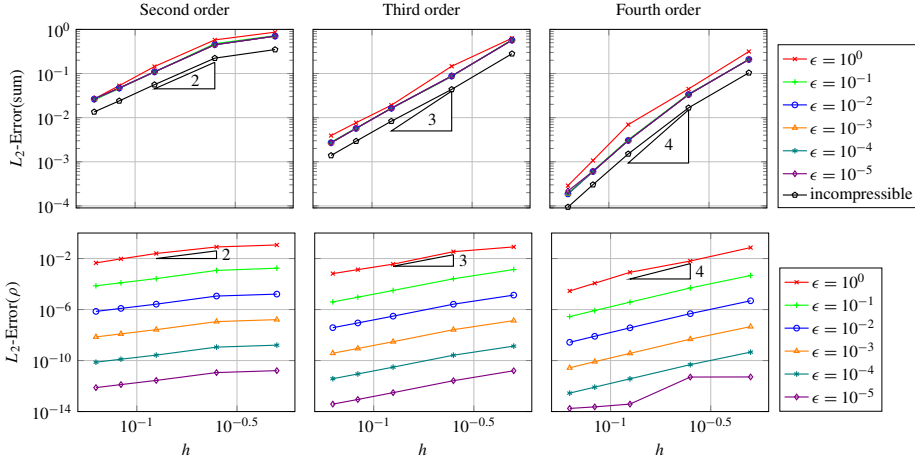
$$\begin{aligned} \rho(\mathbf{x}, t = 0) &= 2 + (500\varepsilon)^2 \cdot \begin{cases} 0.5e^{2/\Delta r} \Delta r - \text{Ei}(2/\Delta r) & \text{for } r < 0.5, \\ 0 & \text{otherwise,} \end{cases} \\ \mathbf{u}(\mathbf{x}, t = 0) &= \begin{pmatrix} 0.5 \\ 0 \end{pmatrix} + 500 \begin{pmatrix} -x_2 + 0.5 \\ x_1 - 0.5 \end{pmatrix} \cdot \begin{cases} e^{1/\Delta r} & \text{for } r < 0.5, \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (12)$$

with  $r := \sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2}$ ,  $\Delta r := r^2 - 0.25$ , and the equation of state  $p(\rho) = \kappa\rho^\gamma$  with  $\kappa = 0.5$  and  $\gamma = 2$ . Ei denotes the exponential integral function

$$\text{Ei}(x) := \int_{-\infty}^x \frac{e^t}{t} dt.$$

In our implementation we use the algorithm by Press et al. [36] for the exponential integral function. Boundary conditions of the domain  $\Omega = [0, 1]^2$  are chosen to be periodic. The initialization of the incompressible pressure is obtained via (5).

**5.1. Validation.** Here, we present numerical results validating the solver. As time integrators, we use the IMEX Runge–Kutta schemes IMEX-ARS-222 and IMEX-ARS-443 by [2] as second- and third-order schemes and IMEX-ARK-4A2 from [34] as a fourth-order scheme. All schemes are given with their Butcher tableaux in the Appendix; see Tables 2, 3, and 4. In the numerical results, an appropriate polynomial degree is chosen so that the overall order is the order of the time integration scheme.



**Figure 1.**  $h$ -convergence of second-, third-, and fourth-order incompressible and RS-IMEX schemes for traveling vortex in overall  $L_2$ -error (top) and  $L_2$ -error in density (bottom) for different Mach numbers.

Figure 1 shows the convergence of the overall  $L_2$ -error including the errors in momentum and density (top) for the incompressible solver and the RS-IMEX splitting. Overall, the  $L_2$  error is computed by

$$\|\mathbf{w}_h - \mathbf{w}\|_{L_2(\Omega)}^2 := \int_{\Omega} \|\mathbf{w}_h(\mathbf{x}) - \mathbf{w}(\mathbf{x})\|_2^2 dx$$

where  $\mathbf{w}_h$  is the computed numerical approximation and  $\mathbf{w}$  the exact solution at the final time instance. Note that for the incompressible equation the error is computed in  $p_{(2)}$  and  $\mathbf{u}_{(0)}$  and for the compressible equation the error is computed in  $\rho$  and  $\rho\mathbf{u}$ . Both the incompressible solver itself and the RS-IMEX splitting which uses the incompressible solver show the correct order of convergence. Only the third-order case shows an order that is slightly too low, but this is inherent to the test case and has already been observed in [24] for under-resolved explicit calculations. Note that the overall  $L_2$ -errors for Mach numbers ranging from  $\epsilon = 10^{-1}$  to  $\epsilon = 10^{-5}$  nearly coincide. Additionally, Figure 1 shows the convergence in density for the RS-IMEX splitting (bottom). Here, the correct order is obtained from second to fourth order and in contrary to the overall  $L_2$ -error, the  $L_2$ -error in density scales with  $\epsilon^2$ . This is due to the structure of the test case and the asymptotic preserving property of the method: the density can be expressed as  $\rho = \text{const} + \mathcal{O}(\epsilon^2)$ , which is a disturbance in  $\epsilon^2$  added to a constant—this can be reproduced exactly by the DG scheme due to the AP property. Momentum can be expressed as  $\rho\mathbf{u} = \mathcal{O}(1)$ , and therefore, the error does not scale with  $\epsilon$ .

**5.2. Efficiency.** In this subsection, we evaluate the efficiency of the RS-IMEX splitting in the low Mach number limit. A desirable method has the following properties.

- It is computationally cheaper than a fully implicit scheme. We have hope that this will be the case due to the linearity of the implicit flux  $\tilde{F}$ .
- The scheme should — for small Mach numbers — be more efficient than a fully explicit scheme. This can also be expected, because the RS-IMEX scheme should be stable under a time step restriction that depends solely on  $\Delta x$ , and not on  $\varepsilon$ . An explicit scheme will always have a time step restriction of form  $\Delta t \lesssim \varepsilon \Delta x$  due to the CFL condition.

For relatively large Mach numbers, we expect the RS-IMEX splitting scheme to be computationally more expensive as additional equations have to be solved. The task of this section is to identify the “sweet spot” between an explicit scheme and the RS-IMEX scheme.

We do not use the standard Lax–Friedrichs Riemann solver for the explicit and fully implicit solver as it is known to give wrong results in the low Mach number limit. The standard Lax–Friedrichs Riemann solver is defined as

$$\mathbf{F}_{\text{LF}}^* = \frac{1}{2}(\mathbf{F}(\mathbf{w}^+) + \mathbf{F}(\mathbf{w}^-) + \lambda_{\max}(\mathbf{w}^+ - \mathbf{w}^-)),$$

with

$$\lambda_{\max} = \max(|\mathbf{u}^+ \cdot \mathbf{n}|, |\mathbf{u}^- \cdot \mathbf{n}|) + \frac{\max(c^+, c^-)}{\varepsilon}.$$

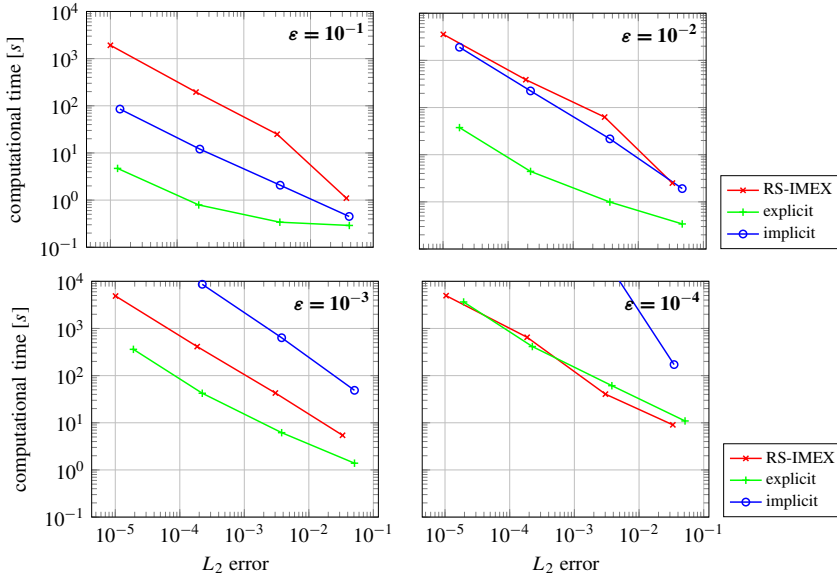
Inspired by the low Mach number fix for the Roe Riemann solver by Rieper [39], we utilize the low Mach Lax–Friedrichs Riemann solver

$$\mathbf{F}_{\text{LF LMFix}}^* = \frac{1}{2}(\mathbf{F}(\mathbf{w}^+) + \mathbf{F}(\mathbf{w}^-) + \lambda_{\max} \mathbf{Diag}(1, \varepsilon, \varepsilon, \varepsilon)(\mathbf{w}^+ - \mathbf{w}^-)). \quad (13)$$

Note that the idea of a different scaling of density and momentum jump with respect to the Mach number has also been applied for the numerical flux of the implicit part of the RS-IMEX splitting. We show later in Section 6 that a modification multiplying the whole jump in the Riemann solver with  $\varepsilon$  is not sufficient.

We compare the computational effort for a fully implicit, a fully explicit, and the RS-IMEX scheme in Figure 2. The results have been obtained on sixteen cores with a temporal and spatial order of four. As the time integration scheme we used the IMEX-ARK-4A2 [34] for RS-IMEX, the implicit part of the same scheme for the implicit method, and a five-stage Runge–Kutta scheme [12] (see Table 6) for the explicit part. For all computations we start with the same grid and perform several refinements.

CFL numbers were chosen as  $\text{CFL} = 0.9$  for the explicit scheme,  $\text{CFL} = 150$  for the implicit scheme, and  $\text{CFL} = 0.5$  for the RS-IMEX scheme. For the fully



**Figure 2.** Comparison of computational time for two-dimensional traveling vortex (fourth order in space and time) with respect to overall  $L_2$  error.

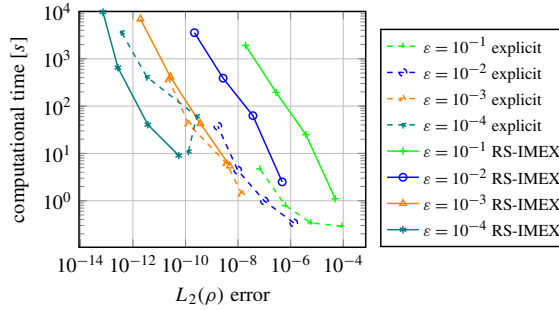
implicit and explicit scheme, the CFL condition is calculated using the eigenvalues of the unsplit system (3), whereas the RS-IMEX splitting only uses the convective eigenvalue ( $\lambda_{1,2}$  in (3)). Our computations showed that for the implicit scheme  $\text{CFL} = 150$  is a good compromise between required time steps and required iterations per time step. (Note that the performance of a *linear* solver depends heavily on  $\Delta t$ ,  $\Delta x$ , and  $\varepsilon$ .)

First of all, we can conclude from Figure 2 that RS-IMEX computes a smaller error on the same grid compared to the other methods (the  $i$ -th dot of each graph corresponds to the same grid).

It can be seen from Figure 2 that the computational time of the explicit and the implicit scheme scales somehow inversely to the Mach number. Since the equation system gets more and more stiff for  $\varepsilon \ll 1$ , the computational cost of the implicit method grows faster than the explicit ones. For the RS-IMEX only a slight increase in computational time is noticeable for a decreasing Mach number.

If the efficiency is defined as the quotient of error and computational effort, the efficiency of the explicit and implicit scheme decreases stronger than for the RS-IMEX splitting with decreasing  $\varepsilon$  due to the aforementioned scaling.

The implicit method shows an extreme growth in computational cost and therefore for  $\varepsilon < 10^{-2}$  the efficiency of the implicit method becomes worse than the efficiency of the RS-IMEX method. The explicit method reaches this sweet spot for a much smaller value of  $\varepsilon$ , i.e., for  $\varepsilon \leq 10^{-4}$ , since the computational cost of the explicit



**Figure 3.** Comparison of computational time for 2d traveling vortex (4th order in space and time) with respect to  $L_2$  error in density.

method is much smaller. Note that fully implicit calculations with  $\varepsilon = 10^{-4}$  are very expensive as machine accuracy issues caused by the finite difference (10) lead to an extremely strong increase in computational time due to slow convergence.

We computed the same tests for a lower spatial order (second order in space and fourth order in time) and a higher spatial order (eighth order in space and fourth order in time) and obtained similar results with an earlier (low-order case) and later (higher-order case) break-even point. This behavior can be explained by the worsening of an implicit high-order scheme due to increasing storage requirements.

More improvements concerning efficiency are obtained if the error in density is considered, displayed in Figure 3. Again, the  $i$ -th symbol of each line corresponds to the same mesh. Therefore, it is visible that for low Mach numbers one obtains significantly lower errors with the RS-IMEX scheme than with the fully explicit scheme with the same mesh. The graph shows that the RS-IMEX splitting is more efficient than the explicit scheme for Mach numbers  $\varepsilon < 10^{-3}$ . The steepening of the  $\varepsilon = 10^{-4}$  RS-IMEX line is due to round-off errors, which occur due to machine precision. We take a closer look on this problem in the next subsection.

**5.3. Solving in the perturbation.** It has to be noted that for very small Mach numbers, the equation becomes extremely stiff and therefore limited machine accuracy can be a problem. Indeed, in [24] the authors observed problems with the accuracy for the RS-IMEX discretization for small values of  $\varepsilon$  which cannot be explained by order reduction [10]. Similar problems have been seen in Figure 3. This observation serves as a motivation to rewrite the method similarly to the proceeding in [43]. The key trick is to rewrite the solution  $\mathbf{w}$  as

$$\mathbf{w} = \underbrace{\mathbf{w}_{(0)}}_{\text{reference solution}} + \varepsilon \underbrace{(\mathbf{w}_{(1)} + \varepsilon \mathbf{w}_{(2)} + \mathcal{O}(\varepsilon^2))}_{\text{perturbation}} =: \mathbf{w}_{(0)} + \varepsilon \delta \mathbf{w}$$



and to observe that  $\mathbf{w}_{(0)}$  is already part of the algorithm and therefore known, so one only has to solve in the perturbation  $\delta\mathbf{w}$  which fulfills the equation

$$\partial_t \mathbf{w}_{(0)} + \varepsilon \partial_t \delta\mathbf{w} + \nabla_x \cdot (\tilde{\mathbf{F}}(\mathbf{w}_{(0)} + \varepsilon \delta\mathbf{w}) + \hat{\mathbf{F}}(\mathbf{w}_{(0)} + \varepsilon \delta\mathbf{w})) = 0.$$

In the setting of the isentropic Euler equations,  $\partial_t \mathbf{w}_{(0)}$  can be identified by the corresponding incompressible equations. Therefore, we can replace it by the flux function  $\mathbf{G}(\mathbf{w}_{(0)}, p_{(2)})$  given in (4). This results in

$$\partial_t \delta\mathbf{w} + \frac{1}{\varepsilon} \nabla_x \cdot (\tilde{\mathbf{F}}(\mathbf{w}_{(0)} + \varepsilon \delta\mathbf{w}) - \mathbf{G}(\mathbf{w}_{(0)}, p_{(2)}) + \hat{\mathbf{F}}(\mathbf{w}_{(0)} + \varepsilon \delta\mathbf{w})) = 0,$$

where  $\mathbf{G}$  is added to the stiff part of the equation, i.e., handled with an implicit method, but does not change the implicit matrix, since the values are given. Computing the eigenvalues of the explicit part and using  $\delta(\rho\mathbf{u}) = \rho_{(0)}\delta\mathbf{u} + \mathbf{u}\delta\rho + \varepsilon\delta\rho\delta\mathbf{u}$  yields

$$\hat{\lambda}_{1,2} = \varepsilon(\delta\mathbf{u} \cdot \mathbf{n}), \quad \hat{\lambda}_3 = 0, \quad \text{and} \quad \hat{\lambda}_4 = 2\varepsilon(\delta\mathbf{u} \cdot \mathbf{n}).$$

Consequently, the explicit part has eigenvalues in  $\mathbb{O}(\varepsilon)$  and the resulting method is supposed to show similar stability properties with an improved accuracy because many of the  $\mathbb{O}(\varepsilon^{-1})$  terms drop out.

However, not all the terms cancel directly. One remaining term in the explicit flux is

$$\frac{1}{\varepsilon^2} (p(\rho_{(0)} + \varepsilon\delta\rho) - p(\rho_{(0)}) - p'(\rho_{(0)})\varepsilon\delta\rho).$$

Using a Taylor expansion for  $p$  gives

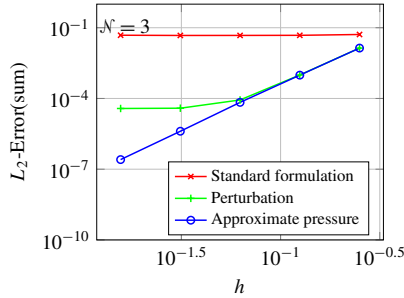
$$p(\rho_{(0)} + \varepsilon\delta\rho) = p(\rho_{(0)}) + \varepsilon p'(\rho_{(0)})\delta\rho + \varepsilon^2 p''(\rho_{(0)})\delta\rho^2 + \mathbb{O}(\varepsilon^3\delta\rho^3),$$

and therefore, the terms read

$$\frac{1}{\varepsilon^2} (p(\rho_{(0)} + \varepsilon\delta\rho) - p(\rho_{(0)}) - p'(\rho_{(0)})\varepsilon\delta\rho) = p''(\rho_{(0)})\delta\rho^2 + \mathbb{O}(\varepsilon\delta\rho^3) \approx p''(\rho_{(0)})\delta\rho^2.$$

We can therefore substitute the expression on the left-hand side by the one on the right-hand side; we call this proceeding *approximate pressure*. Note that—in general—this introduces an additional error in  $\mathbb{O}(\varepsilon\delta\rho^3)$  to the equation, but in our setting  $\delta\rho = \mathbb{O}(\varepsilon)$  and therefore the error would be in  $\mathbb{O}(\varepsilon^4)$ . For the low Mach case, this can safely be assumed to be negligibly small. Note furthermore that for  $\gamma = 2$  this does *not* introduce an additional error.

In Figure 4 results are presented for a very small  $\varepsilon$ . Spatial and temporal accuracy is set to fourth order, i.e., we are using  $\mathcal{N} = 3$  and the IMEX-ARK-4A2 scheme. We show errors for the “straightforward” RS-IMEX discretization, for solving in the perturbation only and for solving in the perturbation with an approximated pressure. Note that for the high-order vortex example the approximated pressure is an exact reformulation since  $\gamma = 2$ . Figure 4 shows that due to the reformulation the



**Figure 4.** Convergence behavior for a fourth-order RS-IMEX discretization for the traveling vortex example for a very low Mach number of  $\varepsilon = 10^{-6}$ .

problems caused by machine accuracy are tremendously reduced. All computations have been done with an exact reference solution to neglect influences due to an inaccurate incompressible solver.

## 6. Numerical results

In this section, we present numerical results for test cases which are more physically motivated than the one considered in the previous subsection. We start with a two-dimensional flow over a cylinder, and subsequently, we investigate the three-dimensional inviscid Taylor–Green vortex.

**6.1. Flow over a cylinder.** This test case demonstrates the ability to use different boundary conditions in our implementation and illustrates the importance of the *asymptotic preserving* property. We compute the two-dimensional, inviscid flow over a cylinder at low Mach numbers. We apply Euler wall boundary conditions on the surface of the cylinder and Dirichlet-type boundary conditions at all other boundaries of the domain. For Euler wall boundaries we can directly prescribe the flux in normal direction at the boundaries as the normal velocity is zero:

$$\begin{aligned}\widetilde{\mathbf{F}}_n &= \frac{(p(\rho_{(0)}) + p'(\rho_{(0)})(\rho - \rho_{(0)}))}{\varepsilon^2} \begin{pmatrix} 0 \\ n_1 \\ n_2 \end{pmatrix}, \\ \widehat{\mathbf{F}}_n &= \frac{p(\rho) - p(\rho_{(0)}) - p'(\rho_{(0)})(\rho - \rho_{(0)})}{\varepsilon^2} \begin{pmatrix} 0 \\ n_1 \\ n_2 \end{pmatrix}, \\ \mathbf{G}_n &= p_{(2)} \begin{pmatrix} 0 \\ n_1 \\ n_2 \end{pmatrix},\end{aligned}$$

whereas pressure and density are prescribed from the inner side. Dirichlet-type boundary conditions are imposed weakly, meaning that a state is prescribed on the boundaries and is used as one state required for the Riemann solver. We use a uniform two-dimensional state  $\mathbf{w}_\infty = (\rho_\infty, u_{1,\infty}, u_{2,\infty})^T = (1.0, 1.0, 0)^T$  (in nondimensional quantities) as initialization and for the Dirichlet boundaries. For the incompressible solver the state  $\mathbf{w}_\infty$  is transformed to  $\mathbf{u}_{\infty,\text{incomp}} = (u_{1,\infty}, u_{2,\infty})^T = (1.0, 0)^T$ ,  $p_{(2),\infty} = 0$ , and  $\rho_{(0),\infty} = \rho_\infty$ . Again, the equation of state  $p(\rho) = \kappa\rho^\gamma$  with  $\kappa = 0.5$  and  $\gamma = 2$  has been utilized. In the low Mach number limit, the exact solution is given by a potential flow field [1]. One measure of solution quality is the pressure coefficient  $C_p$ . It can be computed in two ways: once via the equation of state

$$C_p^{\text{EOS}} = \frac{1}{\varepsilon^2} \frac{p - p_\infty}{\frac{1}{2}\rho_\infty \|\mathbf{u}_\infty\|_2^2} = \frac{1}{\varepsilon^2} \frac{\kappa(\rho^\gamma - \rho_\infty^\gamma)}{\frac{1}{2}\rho_\infty \|\mathbf{u}_\infty\|_2^2}, \quad (14)$$

and once via Bernoulli's hypothesis for an incompressible, inviscid flow [1]

$$C_p^{\text{Bernoulli}} = \frac{1}{\varepsilon^2} \frac{p - p_\infty}{\frac{1}{2}\rho_\infty \|\mathbf{u}_\infty\|_2^2} = 1 - \frac{\rho \|\mathbf{u}\|_2^2}{\rho_\infty \|\mathbf{u}_\infty\|_2^2}. \quad (15)$$

For an incompressible, inviscid flow the result of (14) should coincide with the results of (15), and therefore should satisfy

$$C_p^{\text{EOS}} = C_p^{\text{Bernoulli}} = 1 - 4 \sin^2(\theta),$$

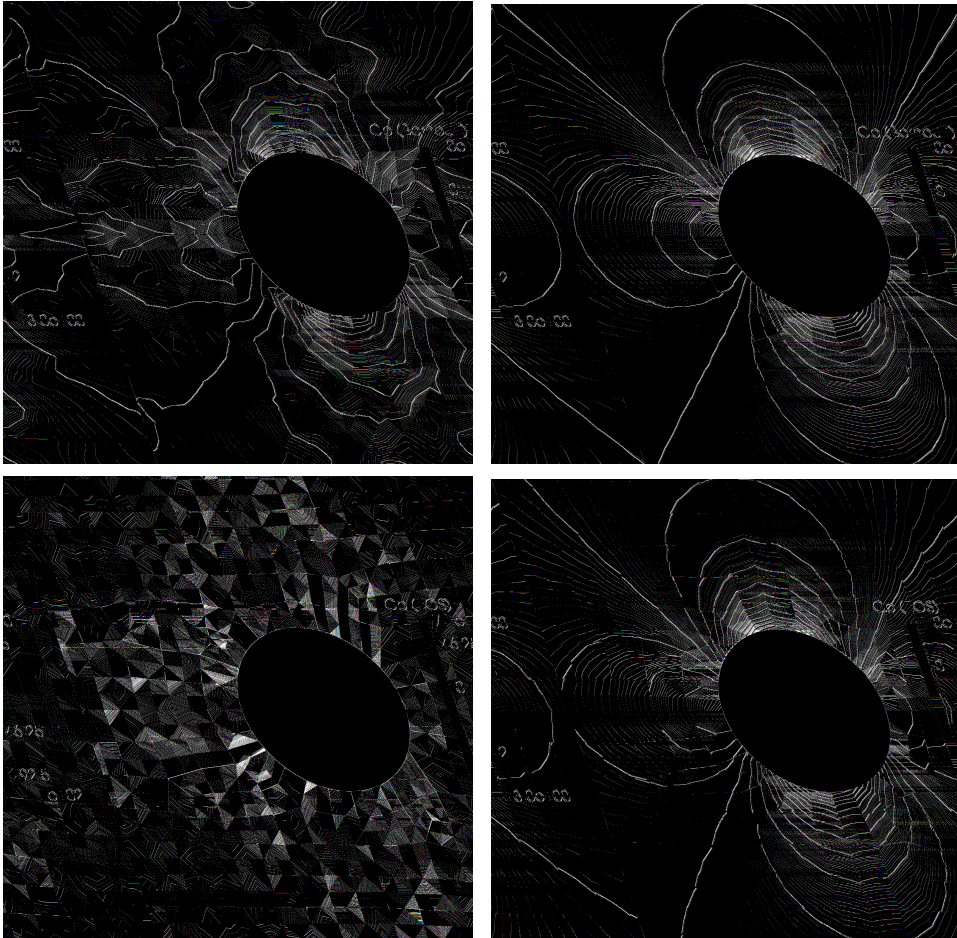
with  $\theta$  being the angular coordinate of the cylinder's polar coordinates ranging from 0 to  $2\pi$  [1]. Hence, the maximum of the pressure coefficient is  $C_p = 1$  at the stagnation points and the minimum  $C_p = -3$  is reached at the positions with maximum velocity on the top and bottom.

Rieper [38] showed that an explicit scheme with a standard HLL-type Riemann solver reproduces the wrong pressure distribution in the low Mach number limit, as it adds too much numerical viscosity. Therefore, the explicit scheme converges to creeping flow where the dynamic pressure is several orders of magnitudes too high. In contrast, an *asymptotic preserving* scheme would reproduce the potential flow correctly. This is given since we can show for a method which is asymptotic preserving that also on the discrete level

$$\rho_h = \rho_{(0)} + \mathcal{O}(\varepsilon^2)$$

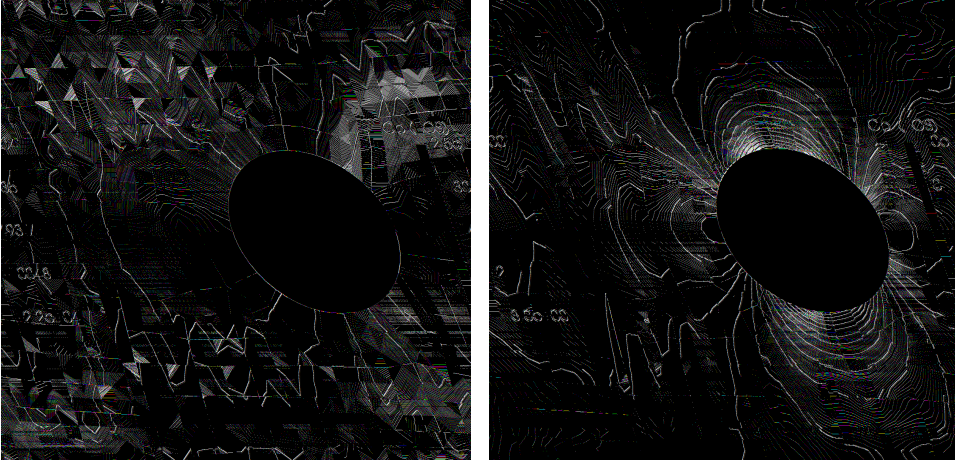
holds. Note that in this case  $\rho_{(0)} = \rho_\infty$ . Therefore, using a Taylor expansion in (14) we obtain

$$C_p^{\text{EOS}} = \frac{1}{\varepsilon^2} \frac{\kappa \mathcal{O}(\varepsilon^2)}{\frac{1}{2}\rho_\infty \|\mathbf{u}_\infty\|_2^2} = \mathcal{O}(1).$$



**Figure 5.** Isolines and colors of pressure coefficient  $C_p$  calculated via Bernoulli's hypothesis (upper) and via the equation of state (lower) for third-order explicit standard Lax–Friedrichs scheme (left) and RS-IMEX splitting (right) at  $\varepsilon = 10^{-5}$ .

If the method is not asymptotic preserving, the difference in the pressure might be in  $\mathcal{O}(\varepsilon)$  or worse, and therefore, the pressure coefficient  $C_p^{\text{EOS}}$  becomes  $\mathcal{O}(\varepsilon^{-1})$  or worse. This only affects the pressure coefficient computed via the equation of state, which is therefore an important measure of asymptotic quality of the method. Figure 5 shows the results of a calculation with 1646 elements and a polynomial degree of  $\mathcal{N} = 2$  using the standard explicit Lax–Friedrichs Riemann solver on the left and the RS-IMEX splitting on the right. The Mach number is set to  $\varepsilon = 10^{-5}$ . If the pressure coefficient is evaluated via (15), meaning it is mainly influenced by the velocity distribution (upper row in Figure 5), both schemes are able to predict potential flow. A different behavior is observed if the dynamic pressure is evaluated

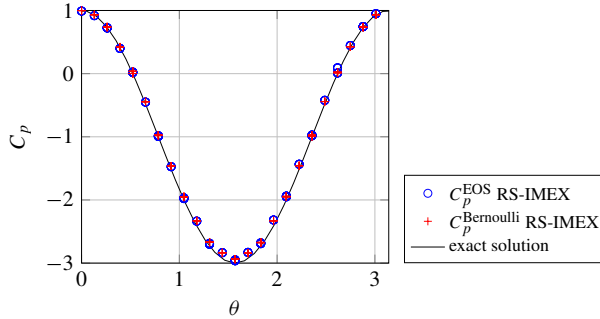


**Figure 6.** Isolines and colors of pressure coefficient  $C_p$  calculated via the equation of state for explicit third-order schemes with insufficient modification of the Lax–Friedrichs scheme (left) and with low Mach Lax–Friedrichs scheme (right) at  $\varepsilon = 10^{-5}$ .

via the equation of state (lower row). Whereas the explicit scheme with a standard Lax–Friedrichs solver does not show the correct flow pattern and has a pressure coefficient several orders or magnitude too high, the RS-IMEX scheme is able to reproduce the potential flow. This illustrates the *asymptotic preserving* property of the scheme. We use the low Mach fix proposed in (13) to show its similar asymptotic behavior compared to the asymptotic preserving RS-IMEX scheme. Figure 6 illustrates that the simple multiplication of the jump with  $\varepsilon$  is not sufficient (left) as it shows the flow pattern of a creeping flow. However, the explicit scheme with the low Mach Lax–Friedrichs Riemann solver (right) is able to predict potential flow. A further validation of the RS-IMEX splitting can be seen in Figure 7 where the  $C_p$  distribution on the upper surface of the cylinder evaluated with the equation of state and with Bernoulli’s hypothesis is compared with the solution for potential flow.

**6.2. Taylor–Green vortex.** The Taylor–Green vortex introduced in [46] is originally a three-dimensional, incompressible viscous test case to study the transition to turbulence and its decay. For nonviscous equation systems such as the isentropic Euler equations it can be used to investigate the amount of dissipation added by a numerical scheme. The standard incompressible initial conditions are given by

$$\begin{aligned} \rho_{(0)} &= 1, \\ \mathbf{u}_{(0)}(\mathbf{x}, t = 0) &= V_0 \begin{pmatrix} \cos(x_1) \cos(x_2) \cos(x_3) \\ -\cos(x_1) \sin(x_2) \cos(x_3) \\ 0 \end{pmatrix}, \\ p_{(2)}(\mathbf{x}, t = 0) &= \frac{\rho_{(0)} V_0^2}{16} (\cos(2x_1) + \cos(2x_2)) (\cos(2x_3) + 2), \end{aligned}$$



**Figure 7.** Distribution of  $C_p^{\text{EOS}}$  and  $C_p^{\text{Bernoulli}}$  for third-order RS-IMEX splitting at  $\varepsilon = 10^{-5}$  calculated with the equation of state or Bernoulli's equation in comparison with potential flow.

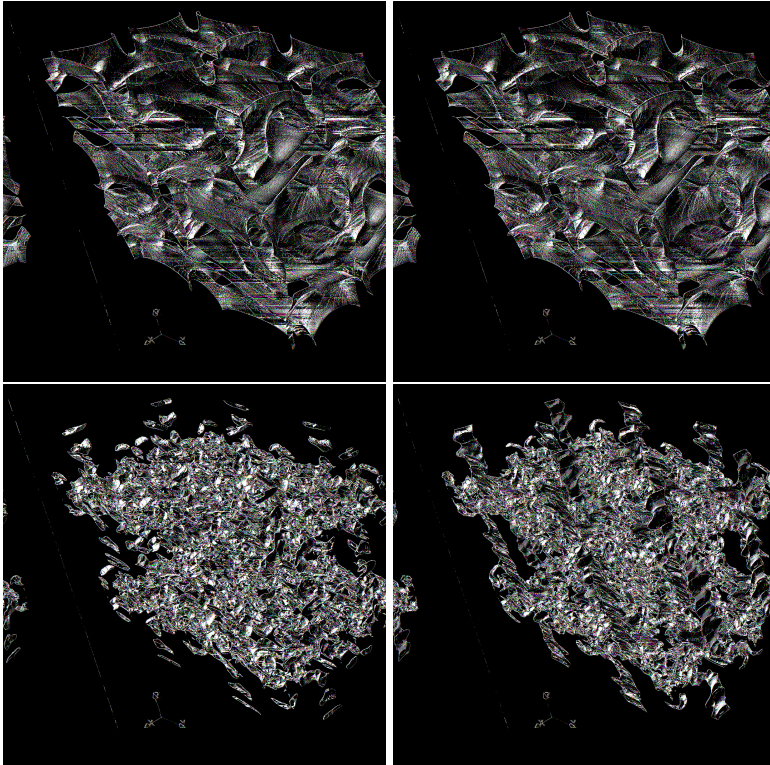
where  $V_0$  denotes a constant initial velocity which is chosen to be  $V_0 = 1$ ;  $x_1$ ,  $x_2$ , and  $x_3$  denote the spatial coordinates of the periodic box  $\Omega = [0, 2\pi]^3$ . We adapt the initialization for the compressible isentropic Euler equations according to (5) to obtain a consistent initial dataset for the incompressible initialization

$$\rho(\mathbf{x}, t = 0) = \rho_{(0)} + \varepsilon^2 \frac{V_0^2 \rho_{(0)}^{2-\gamma}}{16\gamma\kappa} (\cos(2x_1) + \cos(2x_2))(\cos(2x_3) + 2),$$

$$\mathbf{u}(\mathbf{x}, t = 0) = V_0 \begin{pmatrix} \cos(x_1) \cos(x_2) \cos(x_3) \\ -\cos(x_1) \sin(x_2) \cos(x_3) \\ 0 \end{pmatrix},$$

with  $p = \kappa\rho^\gamma$ ,  $\kappa = 0.5$ , and  $\gamma = 2$ . All calculations were conducted on a regular grid with  $16^3$  elements and a polynomial degree of  $\mathcal{N} = 3$ . For the temporal discretization, the third-order IMEX-ARS-443 scheme by Ascher et al. [2] is used. Again, a fully implicit method is obtained if only the implicit Butcher tableau is considered. The explicit calculations were made with a standard three-stage third-order Runge–Kutta scheme [48] (see Table 5). We consider the isosurfaces of the velocity field to compare the results of the RS-IMEX splitting with the explicit scheme in a qualitative manner. Figure 8 exemplarily shows the velocity field at a Mach number of  $\varepsilon = 10^{-4}$  for two different times  $t$ . In the top row, the solutions of both the explicit and the RS-IMEX scheme are identical. For consistent schemes, this is to be expected, since at this early (pretransition) state, the chosen discretization is sufficient to completely resolve the occurring scales. This notion is also supported in Figure 9, where the kinetic energy, defined as

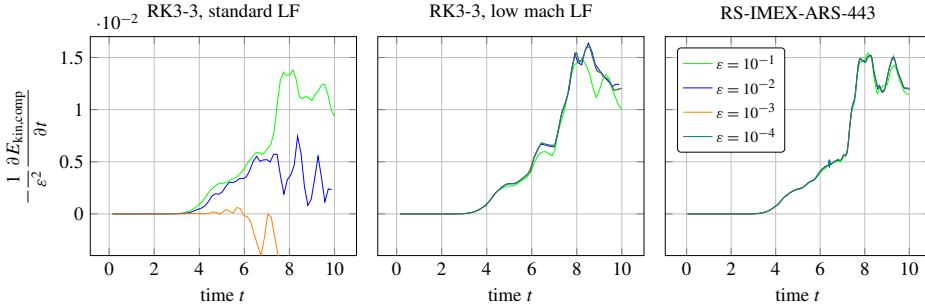
$$E_{\text{kin,comp}} = \frac{\varepsilon^2}{2} \int_{\Omega} \rho \|\mathbf{u}\|_2^2 d\Omega,$$



**Figure 8.** Isosurfaces of velocity magnitude at a physical time of  $t = 3$  (top) and  $t = 7$  (bottom) for explicit (left) and RS-IMEX scheme (right) at  $\varepsilon = 10^{-4}$ .

is preserved at  $t = 3$  for both schemes. The kinetic energy can be used as a benchmark of numerical dissipation properties of a scheme for inviscid flows. In the bottom row of Figure 8, the solutions for  $t = 7$  are shown. Here, clear qualitative differences exist and the kinetic energy is no longer conserved, which can be attributed to the different numerical dissipation mechanisms at work in both schemes. Calculations with other Mach numbers showed analogous results and can be seen as a further validation of the RS-IMEX scheme.

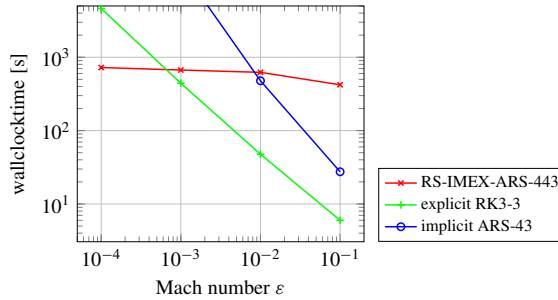
Comparisons of the dissipation rate with the compressible kinetic energy as a measure of quality, displayed in Figure 9, confirm that the explicit scheme with low-Mach Riemann solver and the RS-IMEX method behave similarly in this setting. Differences are due to the slightly different numerical dissipation added by the Riemann solvers. It is visible that a non-asymptotic preserving scheme as the explicit scheme with standard Lax–Friedrichs Riemann solver shows a Mach number dependent behavior which is not desirable. Concluding, we see that the RS-IMEX splitting is able to reproduce a complex three-dimensional physical behavior such as the Taylor–Green vortex.



**Figure 9.** Scaled change rate of compressible kinetic Energy for TGV with explicit scheme with standard Lax–Friedrichs Riemann solver (left), explicit scheme with low Mach Lax–Friedrichs Riemann solver (middle), and RS-IMEX (right) (all fourth-order in space).

$\Delta t_{\text{init}}$	explicit	implicit	RS-IMEX
$\varepsilon = 10^{-1}$	$1.47 \cdot 10^{-3}$	$2.68 \cdot 10^{-1}$	$3.87 \cdot 10^{-2}$
$\varepsilon = 10^{-2}$	$1.53 \cdot 10^{-4}$	$2.79 \cdot 10^{-2}$	$3.87 \cdot 10^{-2}$
$\varepsilon = 10^{-3}$	$1.53 \cdot 10^{-5}$	$2.80 \cdot 10^{-3}$	$3.87 \cdot 10^{-2}$
$\varepsilon = 10^{-4}$	$1.53 \cdot 10^{-6}$		$3.87 \cdot 10^{-2}$

**Table 1.** Initial time steps of calculations with explicit, implicit, and RS-IMEX scheme for the TGV at different Mach numbers and  $\Delta x = \pi/8$ .



**Figure 10.** Comparison of computational time with 528 cores of RS-IMEX splitting and explicit and implicit schemes for TGV with  $16^3$  spatial elements and fourth-order in space.

Focusing on the question of efficiency, the required time for calculations with different discretization methods for several Mach numbers is displayed in Figure 10. The corresponding time steps summarized in Table 1 are given by a constant CFL number for each scheme. Computational effort increases with decreasing Mach number for the explicit scheme as the time step decreases accordingly. A strong increase of computational effort for the fully implicit scheme is noticeable as the



stiffness of the equation system increases. For “high” Mach numbers, more computational time is needed for the RS-IMEX scheme as an additional partial differential equation has to be approximated. But only a slight increase in computational effort for decreasing Mach number is observed as the stiffness is hidden in the linear system instead of the nonlinear system as for a fully implicit discretization. This constitutes obviously a huge advantage of the RS-IMEX splitting compared to a fully implicit scheme. Whereas the stiffness of the fully implicit scheme is increased in the nonlinear system, the Jacobian-vector product in (9) has to be approximated via the finite difference (10). The approximation of the Jacobian-vector product with the finite difference gets worse for an increasing stiffness of the equation system, and therefore, computational time strongly increases for the fully implicit scheme. Using the RS-IMEX splitting the Jacobian-vector product can be calculated exactly with (11). Hence, an increasing stiffness only slightly increases the computational effort. Consequently, large savings concerning computational costs can be obtained by using the RS-IMEX splitting for very low Mach numbers  $\varepsilon < 10^{-3}$  compared to the explicit scheme and  $\varepsilon < 10^{-2}$  compared to the implicit scheme.

## 7. Conclusion and outlook

The efficient and accurate numerical solution of physical phenomena that belong to the class of singularly perturbed problems is still an area of active research. These problems can be seen as a special case of multiscale problems, in which large differences in scale with regards to the average state occur in a spatially confined region of the solution. This becomes especially challenging when high accuracy in the limit is sought, i.e., the discretization should obey the underlying asymptotic properties of the equation.

In this work, we have taken steps towards the development of an efficient high-order DG scheme for all-speed flows at an engineering scale. Starting from the novel operator splitting technique RS-IMEX for the isentropic Euler equations proposed in [25], we have reformulated the discrete equations to significantly extend the Mach number range of the scheme without the occurrence of machine accuracy problems and demonstrated its capability to prevent a stall in convergence.

The RS-IMEX splitting has been implemented in an existing high-order DGSEM framework. The incompressible reference solution is solved by an artificial compressibility-type scheme, which couples the velocity and pressure field through a numerical flux function and thereby introduces a hyperbolic equation for the pressure. Numerical results have shown the efficiency of the method also in the context of realistic three-dimensional applications.

Since the RS-IMEX is conceptually independent from the underlying equations, its naive application to other systems is straightforward. However, it is not a priori

0	0	0	0	0	0	0	0
$\gamma$	0	$\gamma$	0	$\gamma$	$\gamma$	0	0
1	0	$1-\gamma$	$\gamma$	1	$\delta$	$1-\delta$	0
	0	$1-\gamma$	$\gamma$		$\delta$	$1-\delta$	0

**Table 2.** Second-order scheme IMEX-ARS-222 [2] with  $\gamma = (2 - \sqrt{2})/2 \approx 0.293$  and  $\delta = 1 - 1/(2\gamma) \approx -0.707$ .

0	0	0	0	0	0	0	0	0	0	0
1/2	0	1/2	0	0	0	1/2	1/2	0	0	0
2/3	0	1/6	1/2	0	0	2/3	11/18	1/18	0	0
1/2	0	-1/2	1/2	1/2	0	1/2	5/6	-5/6	1/2	0
1	0	3/2	-3/2	1/2	1/2	1	1/4	7/4	3/4	-7/4
	0	3/2	-3/2	1/2	1/2		1/4	7/4	3/4	-7/4

**Table 3.** Third-order scheme IMEX-ARS-443 [2].

clear whether this splitting guarantees hyperbolicity of the explicit part. Current research efforts are underway to answer this question and to explore the possibilities of extending the splitting to the full Euler equations. Furthermore, the application of the splitting to multiphase flows is of current interest.

### Acknowledgments

The authors would like to thank Sebastian Noelle and Nico Kraiss for the fruitful discussions.

Jonas Zeifang has been supported by the German Research Foundation (DFG) through the International Research Training Group GRK 2160: Droplet Interaction Technologies (DROPIT). The computations with the FLEXI framework have been conducted on the Cray XC40 at the High Performance Computing Center Stuttgart under the *hpcdg* project.

Klaus Kaiser has been partially supported by the German Research Foundation (DFG) through project NO 361/6-1; his study was supported by the Special Research Fund (BOF) of Hasselt University.

### Appendix

For the purpose of completeness, we list the Runge–Kutta schemes we have used throughout this paper; see Tables 2, 3, 4, 5, and 6. The left tableaux of the IMEX-Runge–Kutta schemes denote the Butcher tableaux of the part treated implicitly ( $\widetilde{(\cdot)}$ ); the right Butcher tableaux correspond to the explicit part ( $\widehat{(\cdot)}$ ).

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1/3	-1/6	1/2	0	0	0	0	0	0	1/3	1/3	0	0	0	0	0
1/3	1/6	-1/3	1/2	0	0	0	0	0	1/3	1/6	1/6	0	0	0	0
1/2	3/8	-3/8	0	1/2	0	0	0	0	1/2	1/8	0	3/8	0	0	0
1/2	1/8	0	3/8	-1/2	1/2	0	0	0	1/2	1/8	0	3/8	0	0	0
1	-1/2	0	3	-3	1	1/2	0	0	1	1/2	0	-3/2	0	2	0
1	1/6	0	0	0	2/3	-1/2	2/3	0	1	1/6	0	0	0	2/3	1/6
	1/6	0	0	0	2/3	-1/2	2/3	0		1/6	0	0	0	2/3	1/6

**Table 4.** Fourth-order scheme IMEX-ARK-4A2 [34].

$i$	$A_i$	$B_i$	$c_i$
1	0	1/3	0
2	-5/9	15/16	1/3
3	-153/128	8/15	3/4

**Table 5.** Third-order low-storage explicit Runge–Kutta scheme [48].

$i$	$A_i$	$B_i$	$c_i$
1	0	$\frac{1432997174477}{9575080441755}$	0
2	$-\frac{567301805773}{1357537059087}$	$\frac{5161836677717}{13612068292357}$	$\frac{1432997174477}{9575080441755}$
3	$-\frac{2404267990393}{2016746695238}$	$\frac{1720146321549}{2090206949498}$	$\frac{2526269341429}{6820363962896}$
4	$-\frac{3550918686646}{2091501179385}$	$\frac{3134564353537}{4481467310338}$	$\frac{2006345519317}{3224310063776}$
5	$-\frac{1275806237668}{842570457699}$	$\frac{2277821191437}{14882151754819}$	$\frac{2802321613138}{2924317926251}$

**Table 6.** Fourth-order low-storage explicit Runge–Kutta scheme [12]

The explicit schemes are given in the  $2N$ -storage form [12] for the coefficients  $A_i$ ,  $B_i$ , and  $c_i$ .

### References

- [1] J. D. Anderson, Jr., *Fundamentals of aerodynamics*, 3rd ed., McGraw-Hill, 2001.
- [2] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge–Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math. **25** (1997), no. 2–3, 151–167. MR Zbl

- [3] F. Bassi, L. Botti, A. Colombo, D. A. Di Pietro, and P. Tesini, *On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations*, J. Comput. Phys. **231** (2012), no. 1, 45–65. MR Zbl
- [4] F. Bassi, A. Crivellini, D. A. Di Pietro, and S. Rebay, *An artificial compressibility flux for the discontinuous Galerkin solution of the incompressible Navier–Stokes equations*, J. Comput. Phys. **218** (2006), no. 2, 794–815. MR Zbl
- [5] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations*, J. Comput. Phys. **131** (1997), no. 2, 267–279. MR Zbl
- [6] F. Bassi, A. Crivellini, D. A. Di Pietro, and S. Rebay, *An implicit high-order discontinuous Galerkin method for steady and unsteady incompressible flows*, Comput. & Fluids **36** (2007), no. 10, 1529–1546. MR Zbl
- [7] A. D. Beck, T. Bolemann, D. Flad, H. Frank, G. J. Gassner, F. Hindenlang, and C.-D. Munz, *High-order discontinuous Galerkin spectral element methods for transitional and turbulent flow simulations*, Internat. J. Numer. Methods Fluids **76** (2014), no. 8, 522–548. MR
- [8] P. Birken, G. Gassner, M. Haas, and C.-D. Munz, *Preconditioning for modal discontinuous Galerkin methods for unsteady 3D Navier–Stokes equations*, J. Comput. Phys. **240** (2013), 20–35. MR Zbl
- [9] G. Bispen, K. R. Arun, M. Lukáčová-Medvid’ová, and S. Noelle, *IMEX large time step finite volume methods for low Froude number shallow water flows*, Commun. Comput. Phys. **16** (2014), no. 2, 307–347. MR Zbl
- [10] S. Boscarino, *Error analysis of IMEX Runge–Kutta methods derived from differential-algebraic systems*, SIAM J. Numer. Anal. **45** (2007), no. 4, 1600–1621. MR Zbl
- [11] A. Buffa, T. J. R. Hughes, and G. Sangalli, *Analysis of a multiscale discontinuous Galerkin method for convection-diffusion problems*, SIAM J. Numer. Anal. **44** (2006), no. 4, 1420–1440. MR Zbl
- [12] M. H. Carpenter and C. A. Kennedy, *Fourth-order 2N-storage Runge–Kutta schemes*, technical memorandum 109112, National Aeronautics and Space Administration, 1994.
- [13] P. Degond and M. Tang, *All speed scheme for the low Mach number limit of the isentropic Euler equations*, Commun. Comput. Phys. **10** (2011), no. 1, 1–31. MR Zbl
- [14] D. R. Durran, *A physically motivated approach for filtering acoustic waves from the equations governing compressible stratified flow*, J. Fluid Mech. **601** (2008), 365–379. MR Zbl
- [15] S. Fechter and C.-D. Munz, *A discontinuous Galerkin-based sharp-interface method to simulate three-dimensional compressible two-phase flow*, Internat. J. Numer. Methods Fluids **78** (2015), no. 7, 413–435. MR
- [16] F. Filbet and S. Jin, *A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources*, J. Comput. Phys. **229** (2010), no. 20, 7625–7648. MR Zbl
- [17] D. Flad, A. Beck, and C.-D. Munz, *Simulation of underresolved turbulent flows by adaptive filtering using the high order discontinuous Galerkin spectral element method*, J. Comput. Phys. **313** (2016), 1–12. MR Zbl
- [18] M. Franciolini, A. Crivellini, and A. Nigro, *On the efficiency of a matrix-free linearly implicit time integration strategy for high-order discontinuous Galerkin solutions of incompressible turbulent flows*, Comput. & Fluids **159** (2017), 276–294. MR
- [19] F. X. Giraldo and M. Restelli, *High-order semi-implicit time-integrators for a triangular discontinuous Galerkin oceanic shallow water model*, Internat. J. Numer. Methods Fluids **63** (2010), no. 9, 1077–1102. MR Zbl

- [20] F. X. Giraldo, M. Restelli, and M. Läuter, *Semi-implicit formulations of the Navier–Stokes equations: application to nonhydrostatic atmospheric modeling*, SIAM J. Sci. Comput. **32** (2010), no. 6, 3394–3425. MR Zbl
- [21] J. Haack, S. Jin, and J.-G. Liu, *An all-speed asymptotic-preserving method for the isentropic Euler and Navier–Stokes equations*, Commun. Comput. Phys. **12** (2012), no. 4, 955–980. MR Zbl
- [22] F. Hindenlang, G. J. Gassner, C. Altmann, A. Beck, M. Staudenmaier, and C.-D. Munz, *Explicit discontinuous Galerkin methods for unsteady problems*, Comput. & Fluids **61** (2012), 86–93. MR Zbl
- [23] S. Jin, *Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review*, Riv. Math. Univ. Parma (N.S.) **3** (2012), no. 2, 177–216. MR Zbl
- [24] K. Kaiser and J. Schütz, *A high-order method for weakly compressible flows*, Commun. Comput. Phys. **22** (2017), no. 4, 1150–1174. MR
- [25] K. Kaiser, J. Schütz, R. Schöbel, and S. Noelle, *A new stable splitting for the isentropic Euler equations*, J. Sci. Comput. **70** (2017), no. 3, 1390–1407. MR Zbl
- [26] S. Kawai, *Direct numerical simulation of transcritical turbulent boundary layers at supercritical pressures with strong real fluid effects*, 54th AIAA Aerospace Sciences Meeting, American Institute of Aeronautics and Astronautics, 2016, 2016-1934.
- [27] S. Kawai, H. Terashima, and H. Negishi, *A robust and accurate numerical method for transcritical turbulent flows at supercritical pressure with an arbitrary equation of state*, J. Comput. Phys. **300** (2015), 116–135. MR Zbl
- [28] C. A. Kennedy and M. H. Carpenter, *Additive Runge–Kutta schemes for convection-diffusion-reaction equations*, Appl. Numer. Math. **44** (2003), no. 1–2, 139–181. MR Zbl
- [29] S. Klainerman and A. Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math. **34** (1981), no. 4, 481–524. MR Zbl
- [30] R. Klein, *Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics, I: One-dimensional flow*, J. Comput. Phys. **121** (1995), no. 2, 213–237. MR Zbl
- [31] D. A. Knoll and D. E. Keyes, *Jacobian-free Newton–Krylov methods: a survey of approaches and applications*, J. Comput. Phys. **193** (2004), no. 2, 357–397. MR Zbl
- [32] D. A. Kopriva, *Implementing spectral methods for partial differential equations: algorithms for scientists and engineers*, Springer, 2009. MR Zbl
- [33] S.-H. Lee, *Cancellation problem of preconditioning method at low Mach numbers*, J. Comput. Phys. **225** (2007), no. 2, 1199–1210. Zbl
- [34] H. Liu and J. Zou, *Some new additive Runge–Kutta methods and their applications*, J. Comput. Appl. Math. **190** (2006), no. 1–2, 74–98. MR Zbl
- [35] H. Paillere, C. Viozat, A. Kumbaro, and I. Toumi, *Comparison of low Mach number models for natural convection problems*, Heat Mass Transfer **36** (2000), no. 6, 567–573.
- [36] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in Fortran 77: the art of scientific computing*, 2nd ed., Cambridge University, 1996.
- [37] N. Qin, D. K. Ludlow, and S. T. Shaw, *A matrix-free preconditioned Newton/GMRES method for unsteady Navier–Stokes solution*, Int. J. Numer. Meth. Fl. **33** (2000), no. 2, 223–248. Zbl
- [38] F. Rieper, *On the dissipation mechanism of upwind-schemes in the low Mach number regime: a comparison between Roe and HLL*, J. Comput. Phys. **229** (2010), no. 2, 221–232. MR Zbl

- [39] F. Rieber, *A low-Mach number fix for Roe's approximate Riemann solver*, J. Comput. Phys. **230** (2011), no. 13, 5263–5287. MR Zbl
- [40] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput. **7** (1986), no. 3, 856–869. MR Zbl
- [41] S. Schochet, *The mathematical theory of low Mach number flows*, M2AN Math. Model. Numer. Anal. **39** (2005), no. 3, 441–458. MR Zbl
- [42] J. Schütz and K. Kaiser, *A new stable splitting for singularly perturbed ODEs*, Appl. Numer. Math. **107** (2016), 18–33. MR Zbl
- [43] J. Sesterhenn, B. Müller, and H. Thomann, *On the cancellation problem in calculating compressible low Mach number flows*, J. Comput. Phys. **151** (1999), no. 2, 597–615. MR Zbl
- [44] J. R. Simões Moreira and J. E. Shepherd, *Evaporation waves in superheated dodecane*, J. Fluid Mech. **382** (1999), 63–86.
- [45] M. Sonntag and C.-D. Munz, *Efficient parallelization of a shock capturing for discontinuous Galerkin methods using finite volume sub-cells*, J. Sci. Comput. **70** (2017), no. 3, 1262–1289. MR Zbl
- [46] G. I. Taylor and A. E. Green, *Mechanism of the production of small eddies from large ones*, P. Roy. Soc. Lond. A Mat. **158** (1937), no. 895, 499–521. JFM
- [47] Z. J. Wang, K. Fidkowski, R. Abgrall, and et al., *High-order CFD methods: current status and perspective*, Internat. J. Numer. Methods Fluids **72** (2013), no. 8, 811–845. MR
- [48] J. H. Williamson, *Low-storage Runge–Kutta schemes*, J. Comput. Phys. **35** (1980), no. 1, 48–56. MR Zbl
- [49] W.-A. Yong, *A note on the zero Mach number limit of compressible Euler equations*, Proc. Amer. Math. Soc. **133** (2005), no. 10, 3079–3085. MR Zbl
- [50] H. Zakerzadeh and S. Noelle, *A note on the stability of implicit-explicit flux-splittings for stiff systems of hyperbolic conservation laws*, Commun. Math. Sci. **16** (2018), no. 1, 1–15. Zbl

Received May 24, 2017. Revised March 28, 2018.

JONAS ZEIFANG: [zeifang@iag.uni-stuttgart.de](mailto:zeifang@iag.uni-stuttgart.de)

*Institut für Aerodynamik und Gasdynamik, Universität Stuttgart, Stuttgart, Germany*

KLAUS KAISER: [kaiser@igpm.rwth-aachen.de](mailto:kaiser@igpm.rwth-aachen.de)

*Institut für Geometrie und Praktische Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany*

and

*Faculteit Wetenschappen, Universiteit Hasselt, Diepenbeek, Belgium*

ANDREA BECK: [beck@iag.uni-stuttgart.de](mailto:beck@iag.uni-stuttgart.de)

*Institut für Aerodynamik und Gasdynamik, Universität Stuttgart, Stuttgart, Germany*

JOCHEN SCHÜTZ: [jochen.schuetz@uhasselt.be](mailto:jochen.schuetz@uhasselt.be)

*Faculteit Wetenschappen, Universiteit Hasselt, Diepenbeek, Belgium*

CLAUS-DIETER MUNZ: [munz@iag.uni-stuttgart.de](mailto:munz@iag.uni-stuttgart.de)

*Institut für Aerodynamik und Gasdynamik, Universität Stuttgart, Stuttgart, Germany*

# A NUMERICAL STUDY OF THE RELATIVISTIC BURGERS AND EULER EQUATIONS ON A SCHWARZSCHILD BLACK HOLE EXTERIOR

PHILIPPE G. LEFLOCH AND SHUYANG XIANG

We study the dynamical behavior of compressible fluids evolving on the outer domain of communication of a Schwarzschild background. For both the relativistic Burgers equation and the relativistic Euler system, assuming spherical symmetry we introduce numerical methods that take the Schwarzschild geometry and, specifically, the steady state solutions into account. The schemes we propose preserve the family of steady state solutions and enable us to study the nonlinear stability of fluid equilibria and the behavior of solutions near the black hole horizon. We state and numerically demonstrate several properties about the late-time behavior of perturbed steady states.

1. Introduction	271
2. Overview of the theory of the relativistic Burgers model	275
3. A finite volume scheme for the relativistic Burgers model	277
4. Numerical experiments for the relativistic Burgers model, I	280
5. A random choice scheme for the relativistic Burgers model	282
6. Numerical experiments for the relativistic Burgers model, II	284
7. Overview of the theory of the relativistic Euler model	289
8. A finite volume method for the relativistic Euler model	292
9. Numerical experiments for the relativistic Euler model	296
References	299

## 1. Introduction

We are interested in compressible fluid flows on a Schwarzschild black hole background. Motivated by earlier works on relativistic fluid problems posed on curved spacetimes by LeFloch et al. [1; 4; 8; 18; 2; 19; 20] and on numerical methods by Glimm et al. [12; 13] and Russo et al. [28; 29; 30], who argue that steady state solutions should be included in the design of the scheme, as well as relying on the further analytical advances by LeFloch and Xiang [22], we design several

*MSC2010:* 35L60, 35L65, 65M08, 76L05, 76M12.

*Keywords:* relativistic fluid, Schwarzschild black hole, steady state solution, generalized Riemann problem, random choice method, finite volume scheme.

numerical schemes for the approximation of shock wave solutions to the relativistic Burgers equation and to the compressible Euler system. We assume that the flows under consideration are spherically symmetric, and we design schemes that are asymptotic-preserving and allow us to investigate the late-time behavior of solutions. An important challenge we address here is taking the curved geometry into account at the level of the discretization and handling the behavior of solutions near the horizon of the black hole.

The *relativistic Burgers equation on a Schwarzschild background* reads as (see [23] for further details)

$$\partial_t \left( \frac{v}{(1 - 2M/r)^2} \right) + \partial_r \left( \frac{v^2 - 1}{2(1 - 2M/r)} \right) = 0, \quad r > 2M, \quad (1-1)$$

where we have normalized the light speed to unit and the unknown is the function  $v = v(t, r) \in [-1, 1]$ . This equation can also be put in the nonconservative form

$$\partial_t v + \partial_r \left( \left( 1 - \frac{2M}{r} \right) \frac{v^2 - 1}{2} \right) = \frac{2M}{r^2} (v^2 - 1), \quad r > 2M. \quad (1-2)$$

Here  $M > 0$  denotes the mass of the black hole and, clearly, we recover the standard Burgers equations when the mass vanishes.

For the relativistic Burgers model, we design here a finite volume method as well as a random choice method which both preserve steady state solutions. Then, we use these schemes and provide some support as well as some generalization to our theoretical results (briefly reviewed in Theorems 2.1–2.3 below). We treat the following issues:

- the global-in-time existence theory for the generalized Riemann problem and
- the late-time behavior of a steady state (and possibly discontinuous) solution under some initial perturbation.

In addition, our numerical study leads us to the following observations about general initial data.

**Claim 1.1** (relativistic Burgers model). *Given any compactly perturbed steady shock taken as initial data, the corresponding solution to the relativistic Burgers model (1-1) converges (asymptotically in time) to a steady shock.*

**Claim 1.2** (relativistic Burgers model). *Given initial data  $v_0 = v_0(r) \in [-1, 1]$  defined on  $[2M, +\infty)$  and prescribed at some time  $t_0$ , the corresponding solution  $v = v(t, r)$  to the relativistic Burgers model (1-1) enjoys the following properties:*

- *If  $v_0(2M) = 1$ , then there exists a finite time  $t_1 > t_0$  such that, for all  $t > t_1$ , the solution  $v = v(t, r)$  is a single shock connecting the left-hand state 1 to the right-hand escape velocity profile  $-\sqrt{2M/r}$ .*



- If  $v_0(2M) < 1$  and  $\lim_{r \rightarrow +\infty} v_0(r) > 0$ , then there exists a finite time  $t_1 > t_0$  such that, for all  $t > t_1$ , the solution globally coincides with the escape velocity profile  $v(t, r) = -\sqrt{2M/r}$ .
- If  $v_0(2M) < 1$  and  $\lim_{r \rightarrow +\infty} v_0(r) \leq 0$ , then there exists a finite time  $t_1 > t_0$  such that, for all  $t > t_1$ , the solution coincides with

$$v(t, r) = -\sqrt{1 - (1 - (v_0^\infty)^2) \left(1 - \frac{2M}{r}\right)}, \quad \lim_{r \rightarrow +\infty} v_0(r) =: v_0^\infty \leq 0.$$

When the pressure is not assumed to vanish, we consider isothermal fluid flows with pressure law  $p = k^2 \rho$  where  $k \in (0, 1)$  represents the (constant) sound speed. Such an assumption guarantees the hyperbolicity and genuine nonlinearity of the Euler system which, on a Schwarzschild background, reads

$$\begin{aligned} \partial_t \left( r^2 \frac{1+k^2v^2}{1-v^2} \rho \right) + \partial_r \left( r(r-2M) \frac{1+k^2}{1-v^2} \rho v \right) &= 0, \\ \partial_t \left( r(r-2M) \frac{1+k^2}{1-v^2} \rho v \right) + \partial_r \left( (r-2M)^2 \frac{v^2+k^2}{1-v^2} \rho \right) & \tag{1-3} \\ &= 3M \left( 1 - \frac{2M}{r} \right) \frac{v^2+k^2}{1-v^2} \rho - M \frac{r-2M}{r} \frac{1+k^2v^2}{1-v^2} \rho + 2 \frac{(r-2M)^2}{r} k^2 \rho, \end{aligned}$$

where the light speed is normalized to unit. By formally letting  $k \rightarrow 0$ , we recover the pressureless Euler system, from which in turn we derive the relativistic Burgers equation above. On the other hand, by letting the black hole mass  $M \rightarrow 0$ , we recover the relativistic Euler system.

We will also write the Euler equations in the alternative form

$$\begin{aligned} \partial_t \left( \frac{1+k^2v^2}{1-v^2} \rho \right) + \partial_r \left( (1-2M/r) \frac{1+k^2}{1-v^2} \rho v \right) &= -\frac{2}{r} (1-2M/r) \frac{1+k^2}{1-v^2} \rho v, \\ \partial_t \left( \frac{1+k^2}{1-v^2} \rho v \right) + \partial_r \left( (1-2M/r) \frac{v^2+k^2}{1-v^2} \rho \right) & \tag{1-4} \\ &= \frac{-2r+5M}{r^2} \frac{v^2+k^2}{1-v^2} \rho - \frac{M}{r^2} \frac{1+k^2v^2}{1-v^2} \rho + 2 \frac{r-2M}{r^2} k^2 \rho. \end{aligned}$$

We are going to design a finite volume method, with second-order accuracy, that preserves the family of steady state solutions to the Euler equations on a Schwarzschild background. Our numerical study suggests a global-in-time existence theory for the generalized Riemann problem, whose explicit form is not yet known theoretically. In particular, we will be able to exhibit solutions containing up to three steady state components, connected by a 1-wave and a 2-wave.

**Claim 1.3** (relativistic Euler model). *Let  $(\rho_*, v_*) = (\rho_*, v_*)(r)$ ,  $r > 2M$ , be a smooth steady state solution to the relativistic Euler equations on a Schwarzschild background (1-3), and consider the initial data  $(\rho_0, v_0) = (\rho_0, v_0)(r) = (\rho_*, v_*)(r) + (\delta_\rho, \delta_v)(r)$  prescribed at some time  $t_0$ , where the perturbation  $(\delta_\rho, \delta_v) = (\delta_\rho, \delta_v)(r)$  has compact support. Then, for sufficiently large times the corresponding solution  $(\rho, v) = (\rho, v)(t, r)$  to (1-3) coincides with the given steady state solution; in other words, for some time  $t_1 > t_0$ , one has  $(\rho, v)(t, r) = (\rho_*, v_*)(r)$  for all  $t > t_1$ .*

Using steady shocks (discussed in Section 7), we also have:

**Claim 1.4** (relativistic Euler model). *Let  $(\rho_*, v_*) = (\rho_*, v_*)(r)$ ,  $r > 2M$ , be a steady shock, and let  $(\rho_0, v_0) = (\rho_*, v_*)(r) + (\delta_\rho, \delta_v)(r)$  where  $(\delta_\rho, \delta_v) = (\delta_\rho, \delta_v)(r)$  is a compactly supported perturbation. Then there exists a finite time  $t > t_0$  such that the solution is a steady shock for all later times.*

Our numerical random choice scheme is motivated by the methodology in Glimm, Marshall, and Plohr [13] for quasi-one-dimensional gas flows. We rely on static solutions and on the generalized Riemann problem, which we studied in [22] for the relativistic models under consideration here. The numerical analysis of hyperbolic problems posed on curved spacetimes was initiated in [1; 8; 18; 19; 20] using the finite volume methodology. For further background we also refer to [3; 5; 6; 7; 27; 10; 11; 14; 15; 16; 17; 21; 24; 25; 26; 31].

This paper is organized as follows. In Section 2, we briefly overview our theoretical results for the relativistic Burgers model. We include a full description of the family of steady state solutions, as well as some outline of the existence theory for the initial data problem and the nonlinear stability of piecewise steady solutions (see Figure 1). In Section 3, we introduce a finite volume method for the relativistic Burgers model (1-1), which is second-order accurate. In Section 4, we apply our scheme in order to study the generalized Riemann problem and to elucidate the late-time behavior of perturbations of steady solutions.

Building on our theoretical results, in Section 5 we implement a generalized Glimm scheme for the relativistic Burgers model (1-1). Our numerical method is based on an explicit generalized Riemann solver, and therefore, our method preserves all steady state solutions. Numerical experiments are presented in Section 6, in which we are able to validate and expand the theoretical results in Section 2. Our method avoids introducing numerical diffusion and provides an efficient approach for computing shock wave solutions. Furthermore, we apply both methods to the study of the initial problem for the relativistic Burgers equation when the initial velocity is rather arbitrary and we validate our Claims 1.1 and 1.2 and, along the way, clarify the behavior of the fluid flow near the black hole horizon.

Next, in Section 7, we turn our attention to the relativistic Euler model on a Schwarzschild background. We begin by reviewing some theoretical results,

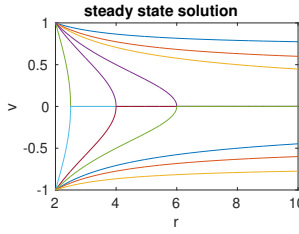


Figure 1. Burgers: steady state solutions.

including the existence theory for steady state solutions, the construction of a solver for the generalized Riemann problem, and the existence theory for the initial value problem. In Section 8 we design a finite volume method for the relativistic Euler model which is second-order accurate. With the proposed algorithm, in Section 9, we are able to tackle the generalized Riemann problem (whose solution is not known in a closed form) and we study the nonlinear stability of steady state solutions when the perturbation has compact support. This allows us to demonstrate numerically the validity of Claims 1.3 and 1.4 above.

### 2. Overview of the theory of the relativistic Burgers model

An important class of solutions to the relativistic Burgers model (1-1) is provided by the *steady state solutions*, that is, solutions depending on the space variable  $r$  only:

$$\partial_r \left( \frac{v^2 - 1}{2(1 - 2M/r)} \right) = 0. \tag{2-1}$$

The steady state solutions are given by

$$v(r) = \pm \sqrt{1 - K^2(1 - 2M/r)}, \tag{2-2}$$

in which  $K > 0$  is an arbitrary constant and, clearly, the sign of a steady state cannot change. Clearly, each such solution is smooth in  $r$  and admits a finite limit  $\lim_{r \rightarrow 2M} v(r) = \pm 1$  at the black hole horizon. Moreover, we have:

- When  $0 < K < 1$ , then the limit at space infinity is  $\lim_{r \rightarrow +\infty} v(r) = \pm \sqrt{1 - K^2}$ .
- When  $K = 1$ , the solution is the critical steady state solution  $v_*^\pm = \pm \sqrt{2M/r}$ , which vanishes at infinity and also coincides with the *escape velocity profile*.
- When  $K > 1$ , the steady state solution is defined only on a bounded interval and stops being defined as the radius  $r^\natural = 2MK^2/(1 - K^2)$ .

In addition to the smooth steady state solutions, we can also define the class of steady shock solutions to the relativistic Burgers equation

$$v(r) = \begin{cases} \sqrt{1 - K^2(1 - 2M/r)}, & 2M < r < r_0, \\ -\sqrt{1 - K^2(1 - 2M/r)}, & r > r_0, \end{cases} \tag{2-3}$$

where  $K > 0$  is a constant and  $r_0 > 2M$  is a given radius. The relevant solutions to the relativistic Burgers equation  $v = v(t, r)$  have a range bounded by the light speed, that is,  $v \in [-1, 1]$  for all  $t > 0$  and  $r > 2M$ . An initial problem of particular importance is given by the generalized Riemann problem, associated with initial data made of two steady states separated by a jump discontinuity located at some given radius.

**Theorem 2.1** (the generalized Riemann problem for the relativistic Burgers model). *There exists a unique solution to the generalized Riemann problem defined for all  $t > 0$  realized either by a shock wave or a rarefaction wave. Moreover, the wave location*

- *tends to the black hole horizon if it initially converges towards the black hole,*
- *tends to the space infinity if it initially converges away from the black hole, and*
- *does not change if it is initially steady.*

In connection with the general existence theory for (1-1), we introduce the auxiliary variable  $z := \text{sgn}(v)\sqrt{(v^2 - 1)/(1 - 2M/r) + 1}$ . It is obvious that  $z$  is a constant if  $v$  is a steady state solution. With this notation, we have the following result from [23].

**Theorem 2.2** (existence theory for the relativistic Burgers model). *Consider the relativistic Burgers equation (1-1) posed on the outer domain of a Schwarzschild black hole with mass  $M$ . Then, for any initial velocity  $v_0 = v_0(r) \in (-1, 1)$  such that  $v_0 = v_0(r)$  has locally bounded total variation, there exists a corresponding weak solution to (1-1)  $z = z(t, r)$  with locally finite total variation in space.*

We are going to design several numerical methods to study these solutions. In particular, we are interested in the behavior of solutions when the initial data  $v_0 = v_0(r)$  is a piecewise smooth and steady state solution, to which we will add a compactly supported perturbation; i.e., we consider

$$v_0(r) = \begin{cases} v_L(r), & 2M < r < r_L, \\ \text{arbitrary values,} & r_L < r < r_R, \\ v_R(r), & r > r_R, \end{cases} \tag{2-4}$$

where  $v_L = v_L(r)$  and  $v_R = v_R(r)$  are two steady state solutions given by (2-2) and  $r_L, r_R$  are two fixed points.

**Theorem 2.3** (time-asymptotic properties for the relativistic Burgers model). *Consider the asymptotic behavior of a relativistic Burgers solution  $v = v(t, r)$  on a Schwarzschild background (1-1) whose initial data are composed of steady state solutions  $v_L, v_R$  with a compactly supported perturbation.*

- *If  $v_L > v_R$ , then the solution  $v = v(t, r)$  converges asymptotically to a shock curve generated by a left-hand state  $v_L$  and a right-hand state  $v_R$ .*

- If  $v_L < v_R$ , then a generalized  $N$ -wave  $N = N(t, r)$  can be defined such that inside a rarefaction fan one has  $|v(t, r) - N(t, r)| = O(t^{-1})$  while in a region supporting the evolution of the initial data one has  $|v(t, r) - N(t, r)| = O(t^{-1/2})$ . Otherwise, one has  $v(t, r) = N(t, r)$ .
- If  $v_L = v_R$ , then  $\|v(t, r) - v_R(t, r)\|_{L^1(2M, +\infty)} = O(t^{-1/2})$ .

### 3. A finite volume scheme for the relativistic Burgers model

*The first-order formulation.* In this section, we propose a finite volume method for the relativistic Burgers equation (1-2) which takes the Schwarzschild geometry into consideration. In order to construct our approximations, we rely on the Riemann solver for the standard Burgers equation:

$$\partial_t v + \partial_x \frac{v^2}{2} = 0, \tag{3-1}$$

that is, an initial data problem with  $v(t, r) = v_0(r)$  where  $v_0 = v_0(r)$  is given as a piecewise constant function

$$v_0 = \begin{cases} v_L, & r < r_0, \\ v_R, & r > r_0, \end{cases}$$

for some fixed  $r_0$  and two constants  $v_L, v_R$ . The solution to the standard Riemann problem reads

$$v(t, r) = \begin{cases} v_L, & r < s_L t + r_0, \\ (r - r_0)/t, & s_L t + r_0 < r < s_R t + r_0, \\ v_R, & r > s_R t + r_0, \end{cases} \tag{3-2}$$

with

$$s_L = \begin{cases} v_L, & v_L < v_R, \\ (v_L + v_R)/2, & v_L > v_R, \end{cases} \quad s_R = \begin{cases} v_R, & v_L < v_R, \\ (v_L + v_R)/2, & v_L > v_R. \end{cases} \tag{3-3}$$

Denote by  $\Delta t, \Delta r$  the mesh lengths in time and in space, respectively, with ratio denoted by  $\Lambda = \Delta t / \Delta r$ . We also set  $t_n = n \Delta t$  and  $r_j = 2M + j \Delta r$ . Introduce also the mesh point  $(t_n, r_j)$ ,  $n \geq 0$  and  $j \geq 0$ , and the rectangle  $R_{nj} = \{t_n \leq t < t_{n+1}, r_{j-1/2} \leq r < r_{j+1/2}\}$ .

Integrate (1-2) from  $r_{j-1/2}$  to  $r_{j+1/2}$  in space and from  $t_n$  to  $t_{n+1}$  in time:

$$\int_{r_{j-1/2}}^{r_{j+1/2}} (v(t_{n+1}, r) - v(t_n, r)) dr + \int_{t_n}^{t_{n+1}} \left( (1 - 2M/r_{j+1/2}) \left( \frac{v(t, r_{j+1/2})^2 - 1}{2} \right) - (1 - 2M/r_{j-1/2}) \left( \frac{v^2(t, r_{j-1/2}) - 1}{2} \right) \right) dt - \int_{r_{j-1/2}}^{r_{j+1/2}} \int_{t_n}^{t_{n+1}} \frac{2M}{r^2} (v^2 - 1) dt dr = 0.$$

Denote by

$$V_j^n \simeq \frac{1}{\Delta r} \int_{r_{j-1/2}}^{r_{j+1/2}} v(t_n, r) dr$$

the approximate average of the solution in the space interval  $(r_{j-1/2}, r_{j+1/2})$ , and let us write a finite volume scheme for the relativistic Burgers equation on a Schwarzschild background in the form

$$V_j^{n+1} = V_j^n - \frac{\Delta t}{\Delta r} (F_{j+1/2} - F_{j-1/2}) - \Delta t \frac{2M}{r_j^2} (V_j^{n2} - 1), \tag{3-4}$$

where  $F_{j+1/2} = \mathfrak{F}(r_{j+1/2}, V_j^n, V_{j-1}^n)$  with

$$\mathfrak{F}(r, V_L, V_R) = \left(1 - \frac{2M}{r}\right) \frac{1}{2} (q(V_L, V_R)^2 - 1) \tag{3-5}$$

with  $q(\cdot, \cdot)$  the standard solution to the Riemann problem centered at  $r$  given by (3-2). The CFL condition

$$\Lambda \max\left(1 - \frac{2M}{r}\right) v \leq 1$$

(the maximum being taken over all relevant values) guarantees that the solution to the Riemann problem does not leave the rectangle  $R_{n,j}$  within one time step.

We now consider the boundary condition of our finite volume scheme. Let  $J$  be the number of the space mesh points, and we introduce ghost cells at the space boundaries:  $R_{n,0} = \{t_n \leq t < t_{n+1}, r_{-1/2} \leq r < r_{1/2}\}$  and  $R_{n,J} = \{t_n \leq t < t_{n+1}, r_{J-1/2} \leq r < r_{J+1/2}\}$ . We solve the Riemann problem at the boundary of the interval  $[r_1, r_2]$  with initial conditions

$$V_0(r) = \begin{cases} 1, & r < r_0, \\ V_0^n, & r > r_0, \end{cases} \quad V_J(r) = \begin{cases} V_J^n, & r < r_J, \\ -1, & r > r_J. \end{cases}$$

*A consistency property.*

**Claim 3.1.** *The finite volume method for the relativistic Burgers model introduced in (3-4) satisfies the following properties:*

- *The scheme suitably preserves the steady state solutions to the Euler equations (7-1).*
- *The scheme is consistent; that is, if  $v = v(t, r)$  is an exact solution to the relativistic Burgers model given by the ordinary differential equation (2-1), then for every fixed point  $r > 2M$*

$$\mathfrak{F}(r_R, V_L, V_R) - \mathfrak{F}(r_L, V_L, V_R) = \frac{2M}{r^2} (v^2 - 1)(r_R - r_L) + O(r_R - r_L)^2 \tag{3-6}$$

*holds as  $V_L, V_R \rightarrow v$  and  $r_L, r_R \rightarrow r$ .*

*Proof.* We write

$$\begin{aligned} F_{j+1/2} - F_{j-1/2} &= (1 - 2M/r_{j+1/2}) \frac{q^2(V_j^n, V_{j+1}^n) - 1}{2} - (1 - 2M/r_{j-1/2}) \frac{q^2(V_{j-1}^n, V_j^n) - 1}{2} \\ &= \int_{j-1/2}^{j+1/2} \frac{2M}{r^2} (v^2 - 1) dr = \frac{2M}{r_j^2} (V_j^{n2} - 1), \end{aligned}$$

and therefore,  $V_j^n = V_j^{n+1}$  holds. Next, recall that  $\mathcal{F}(r, V_L, V_R) = (1 - 2M/r) \times (q^2(r, V_L, V_R) - 1)/2$  is the numerical flux of the scheme determined by the standard Riemann solution. A Taylor expansion gives

$$\begin{aligned} 1 - \frac{2M}{r'} &= 1 - \frac{2M}{r} + \frac{2M}{r^2} (r - r') + O(r - r')^2, \\ \frac{q^2(r', V_L, V_R) - 1}{2} &= \frac{v^2 - 1}{2} + v \partial_r v (r - r') + O(r - r')^2. \end{aligned}$$

Hence, we have

$$\begin{aligned} \mathcal{F}(r_R, V_L, V_R) - \mathcal{F}(r_L, V_L, V_R) &= \frac{2M}{r^2} \frac{v^2 - 1}{2} + \left(1 - \frac{2M}{r}\right) v \partial_r v (r_R - r_L) + O(r_R - r_L)^2 \\ &= \partial_r \left( (1 - 2M/r) \frac{v^2 - 1}{2} \right) + O(r_R - r_L)^2 \\ &= \frac{2M}{r^2} (v^2 - 1) (r_R - r_L) + O(r_R - r_L)^2. \quad \square \end{aligned}$$

*A second-order accurate formulation.* We now extend the method to second-order accuracy. We follow the MUSCL methodology in order to achieve second-order accuracy in the space variable. Hence, the solution is now discretized as a piecewise linear function, and we define the min-mod expression

$$\Delta_j^n V = \begin{cases} \min(2|\Delta_{j-1/2} V^n|, 2|\Delta_{j+1/2} V^n|, |\Delta_j V^n|) \\ \quad \text{if } \text{sgn } \Delta_{j-1/2} V^n = \text{sgn } \Delta_{j+1/2} V^n = \text{sgn } \Delta_j V^n, \\ 0 \quad \text{otherwise,} \end{cases} \quad (3-7)$$

where

$$\Delta_j V^n = \frac{1}{2} (\Delta V_{j+1}^n - \Delta V_{j-1}^n), \quad \Delta_{j+1/2} V^n = (\Delta V_{j+1}^n - \Delta V_j^n).$$

Then, our second-order scheme is stated as

$$\begin{aligned} V_j^{n+1} = V_j^n - \frac{\Delta t}{\Delta r} &(\mathcal{F}(r_{j+1/2}, V_j^{n+1/2,R}, V_{j+1}^{n+1/2,L}) - \mathcal{F}(r_{j-1/2}, V_{j-1}^{n+1/2,R}, V_j^{n+1/2,L})) \\ &- \Delta t \frac{2M}{r_j^2} (V_j^2 - 1), \quad (3-8) \end{aligned}$$

in which the numerical flux is still given by (3-5). Here, the two values  $V_{j+1}^{n+1/2,L}$ ,  $V_j^{n+1/2,R}$  are given by

$$\begin{aligned} V_j^{n+1/2,L} &:= V_j^{n,L} - \frac{\Delta t}{2} \left( \frac{(1 - 2M/r_j)V_j^n \Delta_j^n V}{\Delta r} - \frac{2M}{r_j^2} (V_j^{n2} - 1) \right), \\ V_j^{n+1/2,R} &:= V_j^{n,R} - \frac{\Delta t}{2} \left( \frac{(1 - 2M/r_j)V_j^n \Delta_j^n V}{\Delta r} - \frac{2M}{r_j^2} (V_j^{n2} - 1) \right), \end{aligned} \tag{3-9}$$

where, with  $\Delta_j^n V$  defined by (3-7),  $V_j^{n,L} = V_j^n - \Delta_j^n V/2$  and  $V_j^{n,R} = V_j^n + \Delta_j^n V/2$ .

### 4. Numerical experiments for the relativistic Burgers model, I

*Asymptotic-preserving property.* We now present some numerical tests with the proposed finite volume method applied to the relativistic Burgers equation (1-2). As mentioned earlier, we work within the domain  $r > 2M$ , and the mass parameter  $M$  is taken to be  $M = 1$  in all our tests. We work in the space interval  $(r_{\min}, r_{\max})$  with  $r_{\min} = 2M = 2$  and  $r_{\max} = 4$ , and we take 256 points to discretize the space interval.

We begin by showing that the method at both first-order and second-order accuracy preserves the steady state solutions. For positive/negative steady state Burgers solutions  $v = \pm\sqrt{3/4 + 1/(2r)}$ , we see that the initial steady states are exactly conserved by the scheme. We also show that the following steady state shock is preserved by the scheme:

$$v = \begin{cases} \sqrt{3/4 + 1/(2r)}, & 2 < r < 3, \\ -\sqrt{3/4 + 1/(2r)}, & r > 3. \end{cases}$$

We obtain that our finite volume scheme preserves three typical forms for the static solutions, as is illustrated in Figures 2 and 3.

*A moving shock separating two static solutions.* In view of Theorem 2.1, whether the solution to the Riemann problem will move towards the black hole horizon depends only on the behavior of the initial velocity. We take again the space interval to be  $(2.0, 4.0)$  with 256 space mesh points. We take then two kinds of initial data

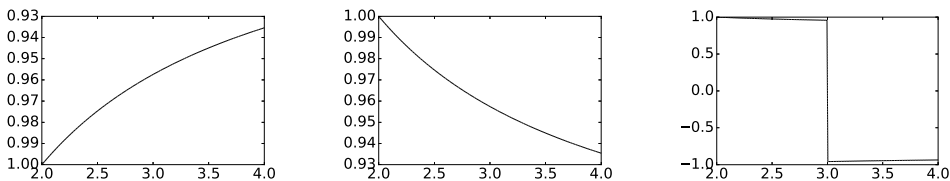
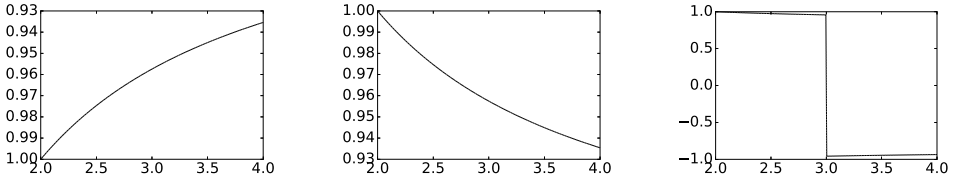
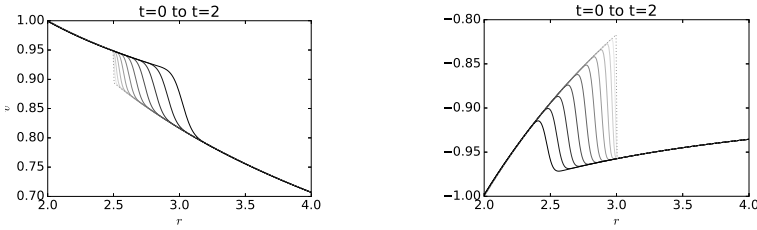


Figure 2. Burgers: three typical behaviors of steady states.

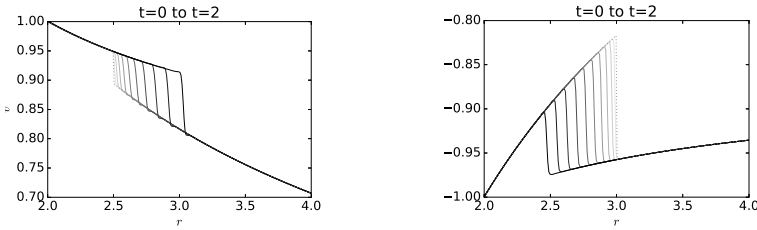




**Figure 3.** Burgers: solution at time  $t = 20$  for three steady state solutions (second-order FVM).



**Figure 4.** Burgers: right- and left-moving shocks (first-order FVM).



**Figure 5.** Burgers: right- and left-moving shocks (second-order FVM).

to be

$$v = \begin{cases} \sqrt{1/2 + 1/r}, & 2 < r < 2.5, \\ \sqrt{2/r}, & r > 2.5, \end{cases} \quad v = \begin{cases} -\sqrt{2/r}, & 2 < r < 2.5, \\ -\sqrt{3/4 + 1/(4r)}, & r > 2.5. \end{cases}$$

The behavior of the two shock solutions obtained with the first-order and second-order accurate versions are shown in Figures 4 and 5.

*Late-time behavior of solutions.* We now study the late-time behavior of solutions whose initial data is given as (2-4), that is, steady state solution with a compactly supported perturbation. We treat the following two kinds of steady state solutions whose values at  $r = 2M$  are  $\pm 1$ , respectively:

$$v = \sqrt{1/2 + 1/r}, \quad v = -\sqrt{1/2 + 1/r},$$

with compactly supported perturbations (see Figure 6).

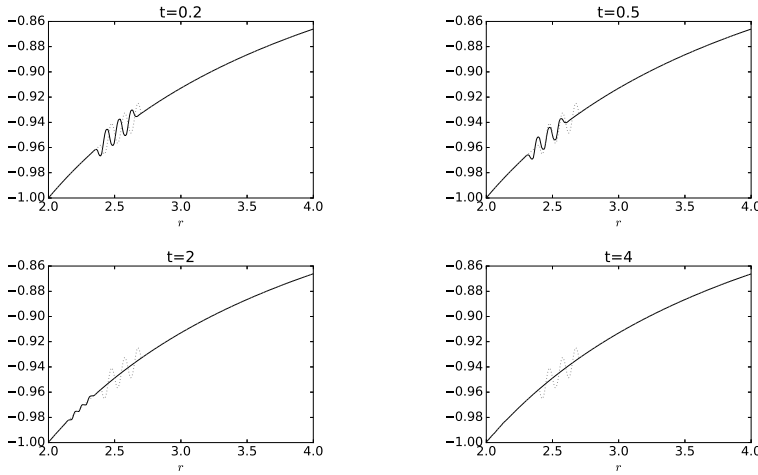


Figure 6. Burgers: evolution of a perturbed steady state (second order FVM).

**5. A generalized random choice scheme for the relativistic Burgers model**

*Explicit solution to the generalized Riemann problem.* In order to construct a Glimm method for the relativistic Burgers model, we need to first introduce the explicit form of the generalized Riemann problem of the relativistic Burgers equation (1-1), which is an initial problem whose initial data  $v_0 = v_0(r)$  is given as

$$v_0(r) = \begin{cases} v_L(r), & 2M < r < r_0, \\ v_R(r), & r > r_0, \end{cases} \tag{5-1}$$

where  $r_0$  is a fixed point in space and  $v_L = v_L(r)$  and  $v_R = v_R(r)$  are two steady state solutions of Burgers equation with explicit form

$$v_L(r) = \text{sgn}(v_L^0) \sqrt{1 - K_L^2 \left(1 - \frac{2M}{r}\right)}, \quad v_R(r) = \text{sgn}(v_R^0) \sqrt{1 - K_R^2 \left(1 - \frac{2M}{r}\right)}, \tag{5-2}$$

where  $K_L, K_R > 0$  are two constants and we denote  $v_L^0 = v_L(r_0)$  and  $v_R(r_0) = v_R^0$ . The existence of the generalized Riemann problem is stated in Theorem 2.1. More precisely, the solution to the Riemann problem  $v = v(t, r)$  can be realized by either a shock wave or a rarefaction wave which is given explicitly by the form

$$v(t, r) = \begin{cases} v_L(r), & r < r_L(t), \\ \tilde{v}(t, r), & r_L(t) < r < r_R(t), \\ v_R(r), & r > r_R(t). \end{cases} \tag{5-3}$$

Here,  $r_L(t)$  and  $r_R(t)$  are bounds of rarefaction regions satisfying

$$R_j(r_j(t)) - R_j(r_0) = t, \tag{5-4}$$

where  $R_j = R_j(r)$  is given by

$$R_j(r) := \frac{R^{v_j}(r)}{2} + \chi_{[v_j^0 < v_k^0]}(r) \frac{R^{v_j}(r)}{2} + \chi_{[v_j^0 < v_k^0]}(r) \frac{R^{v_k}(r)}{2} \quad (5-5)$$

with  $j = L, R$  and  $k = R, L$ ,

$$\chi_{[v_j^0 \geq v_k^0]}(r) = \begin{cases} 1, & v_j^0 \geq v_k^0, \\ 0, & \text{otherwise,} \end{cases}$$

and the function  $R_j^v = R_j^v(r)$  given by

$$\begin{aligned} R^{v_j}(r) := & \operatorname{sgn}(v_j) \frac{1}{(1 - K_j^2)^{3/2}} \left( 2M(1 - K_j^2)^{3/2} \ln(r - 2M) \right. \\ & - 2M(1 - K_j^2)^{3/2} \ln \left( 2r \sqrt{1 - K_j^2} \left( 1 - \frac{2M}{r} \right) + (2M - r)K_j^2 \right) \\ & + 1 \left( r \sqrt{1 - K_j^2} \sqrt{1 - K_j^2} \left( 1 - \frac{2M}{r} \right) \right. \\ & \left. \left. + M(2 - 3K_j^2) \ln \left( r \sqrt{1 - K_j^2} \sqrt{1 - K_j^2} \left( 1 - \frac{2M}{r} \right) + (M - r)K_j^2 + r \right) \right) \right). \end{aligned} \quad (5-6)$$

The function  $\tilde{v} = \tilde{v}(t, r)$  denotes the generalized rarefaction wave

$$\tilde{v}(t, r) = \operatorname{sgn}(r - r_0) \sqrt{1 - K^2(t, r)} \left( 1 - \frac{2M}{r} \right), \quad (5-7)$$

where  $K = K(t, r)$  is characterized by the condition

$$\operatorname{sgn}(r - r_0) = \frac{\tilde{R}(r, K) - \tilde{R}(r_0, K)}{t}, \quad (5-8)$$

where

$$\begin{aligned} \tilde{R}(r, K) := & \frac{1}{(1 - K^2)^{3/2}} \left( 2M(1 - K^2)^{3/2} \ln(r - 2M) \right. \\ & - 2M(1 - K^2)^{3/2} \ln \left( 2r \sqrt{1 - K^2} \left( 1 - \frac{2M}{r} \right) + (2M - r)K^2 \right) \\ & + \left( r \sqrt{1 - K^2} \sqrt{1 - K^2} \left( 1 - \frac{2M}{r} \right) \right. \\ & \left. \left. + M(2 - 3K^2) \ln \left( r \sqrt{1 - K^2} \sqrt{1 - K^2} \left( 1 - \frac{2M}{r} \right) + (M - r)K^2 + r \right) \right) \right). \end{aligned} \quad (5-9)$$

One can check that (5-3) satisfies the Rankine–Hugoniot jump conditions and the entropy inequalities. Importantly, the solution to the generalized Riemann problem is globally defined in time and space.

*A generalized random choice method.* The random choice method is a scheme based on generalized Riemann problems. We use again the time-space grid where the mesh lengths in time and in space are  $\Delta t, \Delta r$  with  $t_n = n\Delta t$  and  $r_j = 2M + j\Delta r$  where we recall  $2M$  is the black hole horizon. Denote by  $V_j^n$  the numerical solution  $V(n\Delta t, 2M + j\Delta r)$ . Let  $(w_n)$  be a sequence equidistributed in  $(-\frac{1}{2}, \frac{1}{2})$ , and write  $r_{n,j} = 2M + (j + w_n)\Delta r$ . We define our Glimm-type approximations as

$$V_j^{n+1} = V_{\mathcal{R}}^{j,n}(t_{n+1}, r_{n,j}), \tag{5-10}$$

where  $V_{\mathcal{R}}^{j,n} = V_{\mathcal{R}}^{j,n}(t, r)$  is the solution to the Riemann problem with the initial data

$$V_0^{j,n} = \begin{cases} V_L^{j,n}(r), & r < r_{j+\text{sgn}(w_n)/2}, \\ V_R^{j,n}(r), & r > r_{j+\text{sgn}(w_n)/2}, \end{cases} \tag{5-11}$$

where the left-hand state  $V_L^{j,n} = V_L^{j,n}(r)$  and the right-hand state  $V_R^{j,n} = V_R^{j,n}(r)$  are steady state solutions to (2-1) with initial conditions

$$\begin{cases} V_L^{j,n}(r_j) = V_j^n, & w_n \geq 0, \\ V_L^{j,n}(r_{j-1}) = V_{j-1}^n, & w_n < 0, \end{cases} \quad \begin{cases} V_R^{j,n}(r_j) = V_j^n, & w_n > 0, \\ V_R^{j,n}(r_{j+1}) = V_{j+1}^n, & w_n \geq 0. \end{cases}$$

We choose a random number only once at each time level  $t = t_n$  rather than in every mesh cell as was done in the original Glimm method.

In order to have an equidistributed sequence, the random values  $(w_n)$  are defined by following Chorin [9]: we give two large prime numbers  $p_1 < p_2$  and define a sequence of integers  $(q_n)$  by

$$q_0 \text{ given } q_0 < p_2, \quad q_n := (p_1 + q_{n-1}) \bmod p_2, \quad n \geq 1. \tag{5-12}$$

Then we define the sequence  $w'_n = (q_n + w_n + \frac{1}{2})/p_2 - \frac{1}{2}$ , which is to be used in our Glimm method instead of  $(w_n)$ . It is direct to see that  $w'_n \in (-\frac{1}{2}, \frac{1}{2})$ .

## 6. Numerical experiments for the relativistic Burgers model, II

*Consistency property.* We now present numerical experiments with the proposed Glimm method for the Burgers equation on a Schwarzschild background (1-1). Recall that  $r > 2M$ , and we choose again  $M = 1$  for the black hole mass. The space interval in consideration is  $(r_{\min}, r_{\max})$  with  $r_{\min} = 2M = 2$  and  $r_{\max} = 4$ . To introduce the random sequence, we fix two prime integers, specifically  $p_1 = 937$  and  $p_2 = 997$  and  $q_0 = 800$ . Since the solution to every local generalized Riemann problem (1-1) with (5-1) is exact, the following observation is immediate.

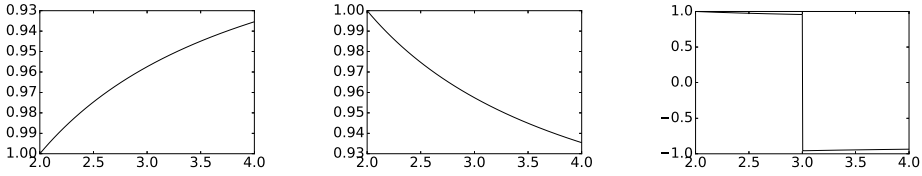


Figure 7. Burgers: evolution at time  $t = 20$  from a steady state initial data (Glimm scheme).

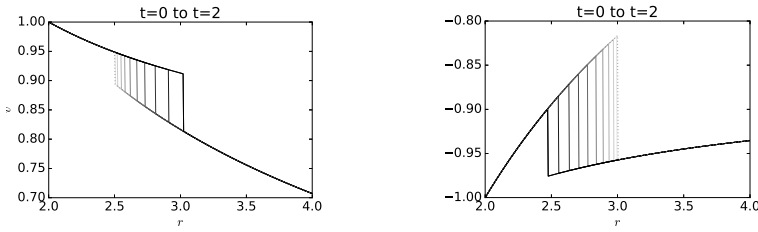


Figure 8. Burgers: right- and left-moving shocks (Glimm scheme).

**Claim 6.1.** Consider a given initial velocity  $v_0 = v_0(r)$  as a steady state solution such that the static Burgers model (2-1) holds. Then the approximate solution to the relativistic Burgers equation (1-1) constructed by the Glimm method (5-10) is exact for such data.

We will still observe the evolution of those three types of solutions shown in Figure 2, that is, the two steady state solutions  $v = \pm\sqrt{3/4 + 1/(2r)}$  and the steady shock

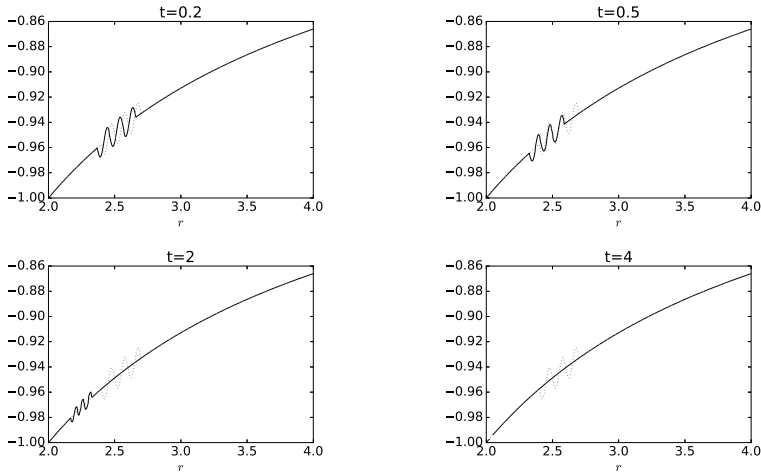
$$v = \begin{cases} \sqrt{3/4 + 1/(2r)}, & 2 < r < 3, \\ -\sqrt{3/4 + 1/(2r)}, & r > 3. \end{cases}$$

*Different types of shocks.* We consider two different shocks whose initial speeds are positive and negative. As was observed by the finite volume method, whether the position of the shock will go toward the black hole horizon is determined uniquely by their initial behavior. We can recover the same conclusion with the Glimm method. Again, we take two kinds of initial data:

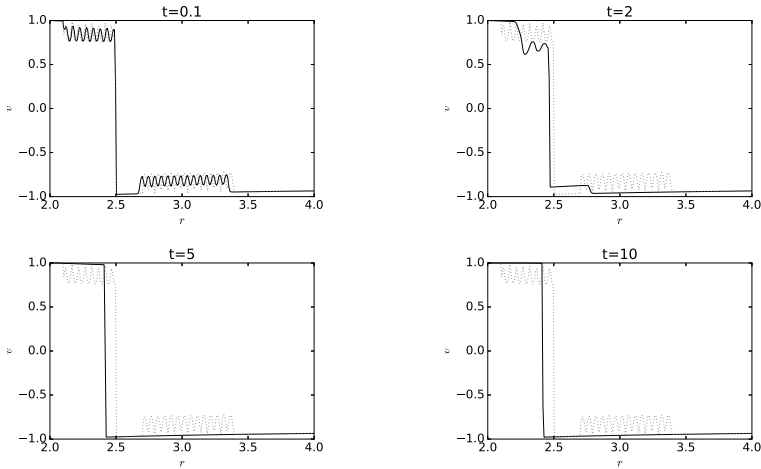
$$v = \begin{cases} \sqrt{1/2 + 1/r}, & 2 < r < 2.5, \\ \sqrt{2/r}, & r > 2.5, \end{cases} \quad v = \begin{cases} -\sqrt{2/r}, & 2 < r < 2.5, \\ \sqrt{3/4 + 1/(4r)}, & r > 2.5. \end{cases}$$

Since our Riemann solver is exact, the numerical solutions contain no numerical diffusion (see Figure 8).

*Asymptotic behavior.* We are now interested in the evolution of solutions whose initial data are given as piecewise steady state solutions satisfying (2-1). As was



**Figure 9.** Burgers: evolution from an initially perturbed steady state (Glimm method).



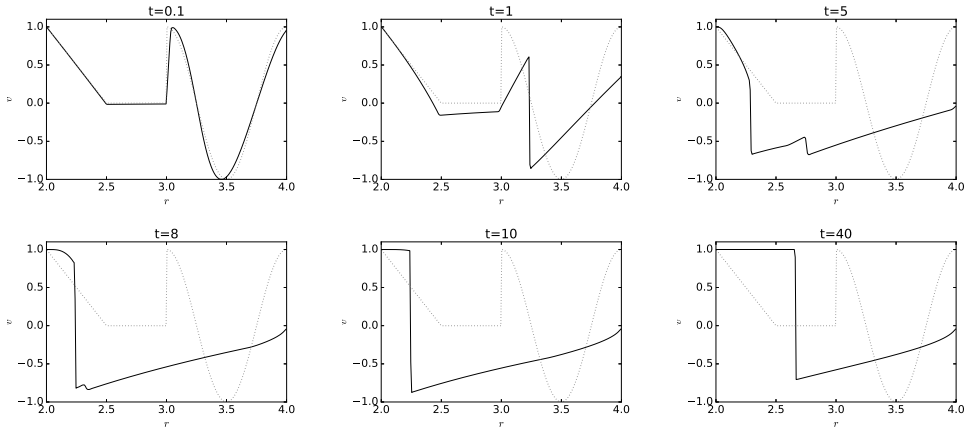
**Figure 10.** Burgers: evolution of a perturbed steady shock (Glimm method).

done earlier, we take into account two kinds of initial data:

$$v = \sqrt{1/2 + 1/r}, \quad v = \begin{cases} \sqrt{1/2 + 1/r}, & 2 < r < 2.5, \\ \sqrt{2/r}, & r > 2.5, \end{cases}$$

perturbed by compactly supported functions (see Figure 9).

*Steady shock with perturbation.* The behavior of a smooth steady state solution to the relativistic Burgers model (1-1) perturbed by a function on a compactly supported function is understood both numerically and theoretically: the solution converges to the same initial steady state solution. The steady shock (2-3) is a



**Figure 11.** Burgers: evolution with prescribed velocity 1 at  $r = 2M$  and at  $r = +\infty$  (Glimm scheme).

solution to the static equation (2-1) in the distribution sense. We are interested in the asymptotic behavior, and our numerical results in Figure 10 lead us to the following:

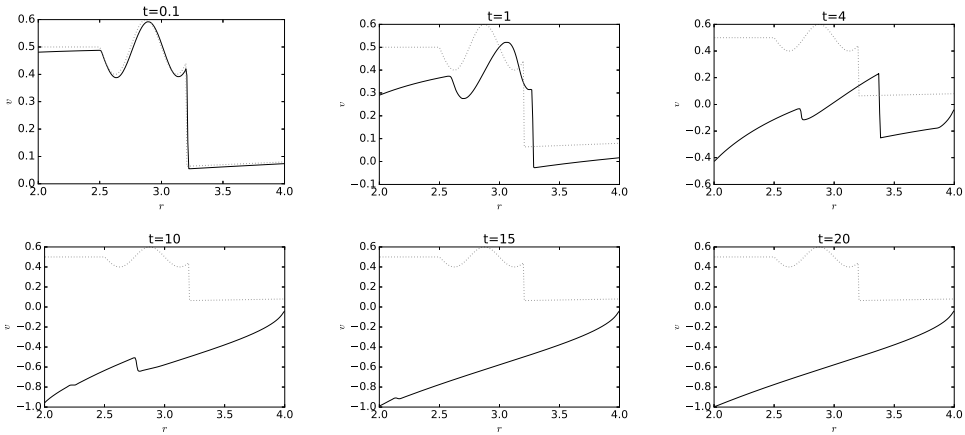
**Conclusion 6.1.** Consider a perturbed steady shock given as (2-3):

$$v_0 = \begin{cases} \sqrt{1 - K^2(1 - 2M/r)}, & 2M < r < r_0, \\ -\sqrt{1 - K^2(1 - 2M/r)}, & r > r_0, \end{cases}$$

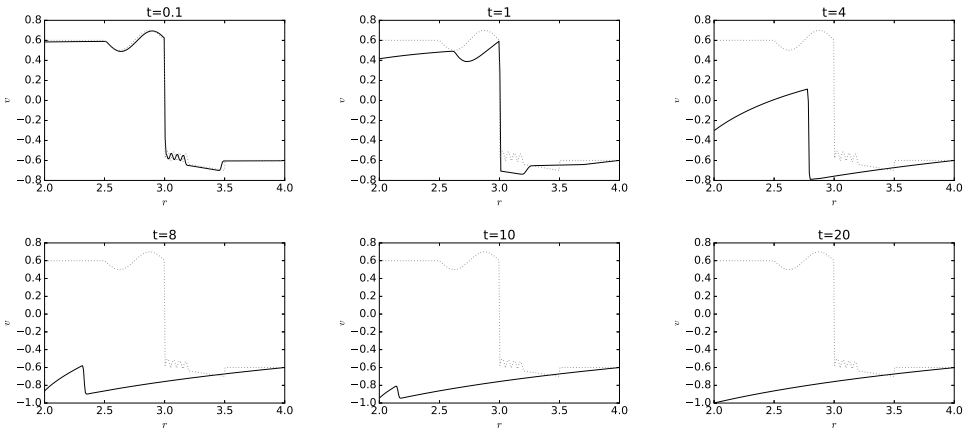
where  $K$  is a given constant and  $r_0 > 2M$  is a fixed radius out of the Schwarzschild black hole region. The solution to the relativistic Burgers model (1-1) converges at some finite time to a solution of the form (with possibly  $r_1 \neq r_0$ )

$$v = \begin{cases} \sqrt{1 - K^2(1 - 2M/r)}, & 2M < r < r_1, \\ -\sqrt{1 - K^2(1 - 2M/r)}, & r > r_1. \end{cases}$$

*Late-time behavior of general solutions.* It is obvious that the steady state solution satisfying (2-1) serves as a solution to the relativistic Burgers equation on a Schwarzschild background. Notice that, on the black hole horizon  $r = 2M$ , the steady state solution equals the light speed, that is, either 1 or  $-1$ , which equals exactly the light speed and obviously their boundary values will not change as time evolves. The value of a steady state solution at infinity is also given explicitly. Observations on the numerical method shows that the asymptotic behavior of the Burgers model (1-1) is mainly determined by the values of the initial data at the black hole horizon  $r = 2M$  and the space infinity  $r = +\infty$ . More precisely, suppose that a given velocity  $v_0 = v_0(r)$  does not satisfy the static Burgers equation (2-1); then we have the following conclusion (see Figures 11, 12, and 13).



**Figure 12.** Burgers: evolution with given velocity less than 1 at  $r = 2M$  and positive at  $r = +\infty$  (Glimm scheme).



**Figure 13.** Burgers: evolution with velocity less than 1 at  $r = 2M$  and negative at  $r = +\infty$  (Glimm scheme).

- Conclusion 6.2.** (1) *If the initial velocity  $\lim_{r \rightarrow 2M} v_0(r) = 1$ , then the solution to the Burgers equation (1-1) satisfies that there exists a time  $t > t_0$  such that for all  $t > t_0$  the solution  $v = v(t, r)$  is a shock with left-hand state 1 and right-hand state  $v_*^-$  with  $v_*^-(r) = -\sqrt{2M/r}$  the negative critical steady solution.*
- (2) *If the initial velocity  $\lim_{r \rightarrow 2M} v_0(r) < 1$  and  $\lim_{r \rightarrow +\infty} v_0(r) > 0$ , there exists a time  $t_0 > 0$  such that the solution to the Burgers equation  $v(t, r) = v_*^-(r)$  for all  $t > t_0$  where  $v_*^-(r) = -\sqrt{2M/r}$  is the negative critical steady state solution to the relativistic Burgers model.*



(3) *If the initial velocity  $\lim_{r \rightarrow 2M} v_0(r) < 1$  and  $\lim_{r \rightarrow +\infty} v_0(r) \leq 0$ , then the solution to the relativistic Burgers model satisfies that*

$$v(t, r) = -\sqrt{1 - (1 - v_0^{\infty 2})\left(1 - \frac{2M}{r}\right)}$$

for  $t > t_0$  for a time  $t_0 > 0$  where  $0 \geq v_0^\infty = \lim_{r \rightarrow +\infty} v_0(r)$ .

### 7. Overview of the theory of the relativistic Euler model

*Continuous and discontinuous steady state solutions.* Steady solutions to the relativistic Euler model on a Schwarzschild background (1-3) are given by the differential system

$$\begin{aligned} \partial_r \left( r(r - 2M) \frac{1}{1 - v^2} \rho v \right) &= 0, \\ \partial_r \left( (r - 2M)^2 \frac{v^2 + k^2}{1 - v^2} \rho \right) &= \frac{M}{r} \frac{(r - 2M)}{1 - v^2} (3\rho v^2 + 3k^2 \rho - \rho - k^2 \rho v^2) + \frac{2k^2}{r} (r - 2M)^2 \rho. \end{aligned} \tag{7-1}$$

Smooth steady states associated with a radius  $r_0 > 2M$ , a density  $\rho_0 > 0$ , and a velocity  $|v_0| < 1$  are given by solving the algebraic system

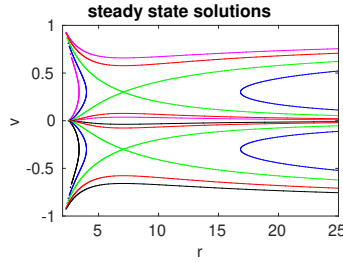
$$\begin{aligned} \text{sgn}(v)(1 - v^2)|v|^{2k^2/(1-k^2)} \frac{r^{4k^2/(1-k^2)}}{(1 - 2M/r)} &= \text{sgn}(v_0)(1 - v_0^2)|v_0|^{2k^2/(1-k^2)} \frac{r_0^{4k^2/(1-k^2)}}{(1 - 2M/r_0)}, \\ r(r - 2M)\rho \frac{v}{1 - v^2} &= r_0(r_0 - 2M)\rho_0 \frac{v_0}{1 - v_0^2}. \end{aligned} \tag{7-2}$$

We have also the expressions of the first-order derivatives

$$\begin{aligned} \frac{d\rho}{dr} &= -\frac{2(r - M)}{r(r - 2M)} \rho - \frac{(1 + v^2)(1 - k^2)}{r(r - 2M)} \rho \left( \frac{2k^2}{1 - k^2} (r - 2M) - M \right) / (v^2 - k^2), \\ \frac{dv}{dr} &= v \frac{(1 - v^2)(1 - k^2)}{r(r - 2M)} \left( \frac{2k^2}{1 - k^2} (r - 2M) - M \right) / (v^2 - k^2). \end{aligned} \tag{7-3}$$

We denote the *critical steady state solution to the relativistic Euler model* (1-3)  $(\rho, v)$  with its velocity  $v = v(r)$  satisfying

$$\begin{aligned} \frac{1 - \epsilon^2 v^2}{1 - 2M/r} (r^2 |v|)^{2\epsilon^2 k^2 / (1 - \epsilon^2 k^2)} &= (1 + 3\epsilon^2 k^2) k^{2\epsilon^2 k^2 / (1 - \epsilon^2 k^2)} \left( \frac{1 + 3\epsilon^2 k^2}{2\epsilon^2 k^2} M \right)^{4\epsilon^2 k^2 / (1 - \epsilon^2 k^2)}. \end{aligned} \tag{7-4}$$



**Figure 14.** Euler: steady state solutions.

Unlike the static Burgers model (2-1), steady state solutions to the relativistic Euler model do not have an explicit form. We recall the following result established in [22]. See Figure 14 for an illustration.

**Theorem 7.1** (smooth steady flows on a Schwarzschild background). *Denoting by  $k \in [0, 1]$  the sound speed and by  $M > 0$  the mass of the black hole, let us consider the relativistic Euler model on a Schwarzschild background (1-3). For any given radius  $r_0 > 2M$ , density  $\rho_0 > 0$ , and velocity  $|v_0| < 1$ , there exists a unique smooth steady state solution  $\rho = \rho(r)$  and  $v = v(r)$  satisfying (7-2) and the initial conditions  $\rho(r_0) = \rho_0$  and  $v(r_0) = v_0$ . Moreover, the velocity component is such that the signs of  $v(r)$  and  $|v(r)| - k$  do not change within the domain of definition of this solution. Two different families of solutions can be distinguished.*

- *If there exists no sonic point at which, by definition, the fluid velocity equals the sound speed, the (smooth) steady state solution is defined globally on the whole interval outside of the black hole horizon  $(2M, +\infty)$ .*
- *Otherwise, the steady state solution cannot be extended as a smooth solution once it reaches the sonic point.*

It is natural to then consider steady shock waves to (1-3), that is, two steady state solutions connected by a standing shock:

$$(\rho, v) = \begin{cases} (\rho_L, v_L)(r), & 2M < r < r_0, \\ (\rho_R, v_R)(r), & r > r_0, \end{cases} \tag{7-5}$$

where  $r_0 > 2M$  is a given radius and  $(\rho_L, v_L)$  and  $(\rho_R, v_R)$  are steady state solutions satisfying (7-2) and

$$v_R(r_0) = \frac{k^2}{v_L(r_0)}, \quad \rho_R(r_0) = \frac{v_L(r_0)^2 - k^4}{k^2(1 - v_L(r_0)^2)} \rho_L(r_0), \quad v_L(r_0) \in (-k, -k^2) \cup (k, 1). \tag{7-6}$$

We refer to such a solution as a *steady shock of the relativistic Euler model*, that is, a function of the form (7-5) and (7-6) satisfying (7-1) in the distributional sense, satisfying the Lax entropy inequality and the Rankine–Hugoniot jump conditions.

Observe that, for a fixed radius  $r_1 \neq r_0$  and  $(\rho_L, v_L), (\rho_R, v_R)$  satisfying (7-5), the following function is *not* a steady shock of the Euler model (1-3):

$$(\rho, v) = \begin{cases} (\rho_L, v_L)(r), & 2M < r < r_1, \\ (\rho_R, v_R)(r), & r > r_1. \end{cases}$$

*Generalized Riemann problem and Cauchy problem.* The generalized Riemann problem for the relativistic Euler system (1-3) is the Cauchy problem with initial data

$$(\rho_0, v_0)(r) = \begin{cases} (\rho_L, v_L)(r), & 2M < r < r_0, \\ (\rho_R, v_R)(r), & r > r_0, \end{cases} \tag{7-7}$$

where  $r = r_0$  is a fixed radius and  $\rho_L = \rho_L(r), v_L = v_L(r), \rho_R = \rho_R(r)$ , and  $v_R = v_R(r)$  are two smooth steady state solutions satisfying the static Euler equations (7-1). Referring to [22], we can construct an approximate solver  $\tilde{U} = (\tilde{\rho}, \tilde{v}) = (\tilde{\rho}, \tilde{v})(t, r)$  of the generalized Riemann problem of the relativistic Euler model (1-3) whose initial data is (7-7) such that:

- $\|\tilde{U}(t, \cdot) - U(t, \cdot)\|_{L^1} = O(\Delta t^2)$  for any fixed  $t > 0$  where  $U = (\rho, v) = (\rho, v)(t, r)$  satisfies (1-3) and (7-7) and  $\Delta t$  is the time step in the construction.
- $\tilde{U} = (\tilde{\rho}, \tilde{v})$  is exact outside the rarefaction fan regions.
- $\tilde{U} = (\tilde{\rho}, \tilde{v})$  (and the exact solution  $U$ ) contains at most three steady states: the two states given in the initial data  $(\rho_L, v_L), (\rho_R, v_R)$  and the uniquely defined intermediate  $(\rho_M, v_M)$  connected by a 1-family wave (either 1-shock or 1-rarefaction) and a 2-family wave (either 2-shock or 2-rarefaction).

**Theorem 7.2** (existence theory of the relativistic Euler model). *Consider the Euler system describing fluid flows on a Schwarzschild geometry (1-3). For any initial density  $\rho_0 = \rho_0(r) > 0$  and velocity  $|v_0| = |v_0(r)| < 1$  satisfying*

$$TV_{[2M+\delta, +\infty)}(\ln \rho_0) + TV_{[2M+\delta, +\infty)}\left(\ln \frac{1 - v_0}{1 + v_0}\right) < +\infty,$$

where  $\delta > 0$  is a constant, there exists a weak solution  $(\rho, v) = (\rho, v)(t, r)$  defined on  $(0, T)$  for any given  $T > 0$  and satisfying the prescribed initial data at the initial time and, with a constant  $C$  independent of time,

$$\begin{aligned} \sup_{t \in [0, T]} \left( TV_{[2M+\delta, +\infty)}(\ln \rho(t, \cdot)) + TV_{[2M+\delta, +\infty)}\left(\ln \frac{1 - v(t, \cdot)}{1 + v(t, \cdot)}\right) \right) \\ \leq TV_{[2M+\delta, +\infty)}(\ln \rho_0) + TV_{[2M+\delta, +\infty)}\left(\ln \frac{1 - v_0}{1 + v_0}\right) e^{CT}. \end{aligned}$$

### 8. A finite volume method for the relativistic Euler model

A *semidiscretized numerical scheme*. We write the relativistic equations on a Schwarzschild background (1-4) as

$$\partial_t U + \partial_r \left( \left( 1 - \frac{2M}{r} \right) F(U) \right) = S(r, U), \quad (8-1)$$

$$U = \begin{pmatrix} U^0 \\ U^1 \end{pmatrix} = \begin{pmatrix} \frac{1+k^2 v^2}{1-v^2} \rho \\ \frac{1+k^2}{1-v^2} \rho v \end{pmatrix}, \quad F(U) = \begin{pmatrix} \frac{1+k^2}{1-v^2} \rho v \\ \frac{v^2+k^2}{1-v^2} \rho \end{pmatrix}, \quad (8-2)$$

with source term

$$S(r, U) = \begin{pmatrix} -\frac{2}{r}(1-2M/r) \frac{1+k^2}{1-v^2} \rho v \\ \frac{-2r+5M}{r^2} \frac{v^2+k^2}{1-v^2} \rho - \frac{M}{r^2} \frac{1+k^2 v^2}{1-v^2} \rho + 2 \frac{r-2M}{r^2} k^2 \rho \end{pmatrix}.$$

The Jacobian matrix

$$D_U F(U) = \begin{pmatrix} 0 & 1 \\ (-v^2+k^2)/(1-k^2 v^2) & 2(1-k^2)v/(1-k^2 v^2) \end{pmatrix} \quad (8-3)$$

admits two real and distinct eigenvalues, denoted  $\mu_{\mp} = (1-2M/r)(v \mp k)/(1 \mp k^2 v)$ .

We also have

$$v = \frac{1+k^2 - \sqrt{(1+k^2)^2 - 4k^2(U^1/U^0)^2}}{2k^2 U^1/U^0} \in (-1, 1)$$

and  $\rho = U^1(1-v^2)/(v(1+k^2))$ .

Denote by  $\Delta t$  and  $\Delta r$  the mesh lengths in time and in space, respectively, and assume the CFL condition

$$\frac{\Delta t}{\Delta x} \max(|\mu_-|, |\mu_+|) \leq \frac{1}{2}. \quad (8-4)$$

We write  $t_n = n\Delta t$  and  $r_j = 2M + j\Delta r$ , and we consider the corresponding mesh points  $(t_n, r_j)$  for all integers  $n \geq 0$  and  $j \geq 0$ . We also set  $\rho(t_n, r_j) = \rho_j^n$ ,  $v(t_n, r_j) = v_j^n$ , and  $U(t_n, r_j) \simeq U_j^n$  where  $U = U(t, r)$  is a solution to (8-1).

We search for our approximations  $U_j^n = (1/\Delta r) \int_{r_{j-1/2}}^{r_{j+1/2}} U(t_n, r) dr$  in the finite volume form

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta r} (F_{j+1/2}^n - F_{j-1/2}^n) + \Delta t S_j^n, \quad (8-5)$$

where the numerical flux is expressed in the form

$$F_{j-1/2}^n = \mathcal{F}_l(r_{j-1/2}, U_{j-1}^n, U_j^n) = \left(1 - \frac{2M}{r_{j-1/2}}\right) \mathcal{F}(U_{j-1/2-}^n, U_{j-1/2+}^n), \quad (8-6)$$

and  $U_{j+1/2\pm}$  and  $U_{j-1/2\pm}$  as well as the source term  $S_j^n = (1/\Delta r) \int_{r_{j-1/2}}^{r_{j+1/2}} S(t, n, r) dr$  must still be determined. For definiteness, we choose the Lax–Friedrichs flux

$$\mathcal{F}(U_L, U_R) = \frac{F(U_L) + F(U_R)}{2} - \frac{1}{\lambda} \frac{U_R - U_L}{2} \quad (8-7)$$

with  $\lambda = \Delta r/\Delta t$ , where  $F$  is the exact flux (8-1).

*Taking the curved geometry into account.* It remains to determine the states  $U_{j+1/2\pm}$  and  $U_{j-1/2\pm}$  as well as the discretized source  $S_j^n$ , which must take into account the Schwarzschild geometry. For a steady state solution  $U = U(r)$ , the equation  $\partial_r((1 - 2M/r)F(U)) = S(r, U)$  holds, where  $U$ ,  $F$ , and the source term  $S$  are given by (8-1). (Equivalently, the solution  $(\rho, v)$  satisfies the static Euler equations (7-1).) We propose to represent the numerical solution in each cell  $(r_{j-1/2}, r_{j+1/2})$  as a steady state solution, whenever such a solution is available. Hence, we require the algebraic relations

$$\begin{aligned} & (1 - v_{j+1/2-}^n)^2 v_{j+1/2-}^{2k^2/(1-k^2)} r_{j+1/2}^{4k^2/(1-k^2)} / (1 - 2M/r_{j+1/2}) \\ & \quad = (1 - v_j^n)^2 v_j^{2k^2/(1-k^2)} r_j^{4k^2/(1-k^2)} / (1 - 2M/r_j), \\ & r_{j+1/2}(r_{j+1/2} - 2M) \rho_{j+1/2-}^n \frac{v_{j+1/2-}^n}{1 - v_{j+1/2-}^n} \\ & \quad = r_j(r_j - 2M) \rho_j^n \frac{v_j^n}{1 - v_j^n}, \\ & (1 - v_{j+1/2+}^n)^2 v_{j+1/2+}^{2k^2/(1-k^2)} r_{j+1/2}^{4k^2/(1-k^2)} / (1 - 2M/r_{j+1/2}) \\ & \quad = (1 - v_{j+1}^n)^2 v_{j+1}^{2k^2/(1-k^2)} r_{j+1}^{4k^2/(1-k^2)} / (1 - 2M/r_{j+1}), \\ & r_{j+1/2}(r_{j+1/2} - 2M) \rho_{j+1/2+}^n \frac{v_{j+1/2+}^n}{1 - v_{j+1/2+}^n} \\ & \quad = r_{j+1}(r_{j+1} - 2M) \rho_{j+1}^n \frac{v_{j+1}^n}{1 - v_{j+1}^n}. \end{aligned} \quad (8-8)$$

A difficulty arises here from the fact that a steady state solution need not be defined globally on the whole interval  $(2M, +\infty)$ , and it is possible that (8-8) does not admit a solution. However, this difficulty can be solved as follows: we simply set  $(\rho_{j+1/2-}^n, v_{j+1/2-}^n) = (\rho_j^n, v_j^n)$  when the first two equations in (8-8) do not

have a solution, while we set  $(\rho_{j+1/2-}^n, v_{j+1/2-}^n) = (\rho_{j+1}^n, v_{j+1}^n)$  when the last two equations in (8-8) do not admit a solution.

Next, integrating (8-5) by parts, we obtain an expression for the source terms, i.e.,

$$\begin{aligned} S_j^n &= \frac{1}{\Delta r} \int_{r_{j-1/2}}^{r_{j+1/2}} S(t_n, r) dr = \frac{1}{\Delta r} \int_{r_{j-1/2}}^{r_{j+1/2}} \partial_r((1-2M/r)F(U(t_n, r))) dr \\ &= \frac{1}{\Delta r} ((1-2M/r_{j+1/2})F(U_{j+1/2-}^n) - (1-2M/r_{j-1/2+})F(U_{j-1/2+}^n)), \end{aligned} \quad (8-9)$$

where the states  $U_{j+1/2-}^n$  and  $U_{j-1/2+}^n$  are determined by (8-8) and  $F(\cdot)$  denotes the flux of the Euler system (8-1). Finally, second-order accuracy in time is achieved in a standard manner via the MUSCL methodology.

**Theorem 8.1.** *The finite volume scheme proposed for the relativistic Euler equations on a Schwarzschild background (1-4) satisfies:*

- *The scheme preserves the steady state solution to the Euler equations (7-1).*
- *The scheme is consistent; that is, for an exact solution  $U = U(t, r)$  and the states  $U_L, U_R \rightarrow U$  and  $r_L, r_R \rightarrow r$ , we have*

$$\mathfrak{F}_r(r_R, U_L, U_R) - \mathfrak{F}_l(r_L, U_L, U_R) = S(r, U)(r_R - r_L) + O((r_R - r_L)^2), \quad (8-10)$$

where  $\mathfrak{F}_l, \mathfrak{F}_r$  are numerical fluxes given by (8-6) and  $S(r, U)$  is the source term given by (8-1).

- *The scheme has second-order accuracy in space and first-order accuracy in time.*

*Proof.* For a steady state given by (7-1), we have  $U_{j+1/2+} = U_{j+1/2-}$ . Hence, the flux of the finite volume method (8-6) satisfies  $F_{j+1/2} = (1-2M/r_{j+1/2})F(U_{j+1/2+}) = (1-2M/r_{j+1/2})F(U_{j+1/2-})$ , which gives

$$\begin{aligned} &\frac{1}{\Delta r} (F_{j+1/2}^n - F_{j-1/2}^n) \\ &= (1-2M/r_{j+1/2})F(U_{j+1/2-}^n) - (1-2M/r_{j-1/2})F(U_{j-1/2+}^n) = S_j^n. \end{aligned}$$

Therefore, the scheme preserves the steady state solutions. Next, according to (8-8) and (8-9), there exist four states  $U_L^l, U_R^l, U_L^r, U_R^r$  such that

$$\begin{aligned}
 & \mathcal{F}_r(r_R, U_L, U_R) - \mathcal{F}_l(r_L, U_L, U_R) \\
 &= (1 - 2M/r_R)\mathcal{F}(U_L^r, U_R^r) - (1 - 2M/r_L)\mathcal{F}(U_L^l, U_R^l) \\
 &= (1 - 2M/r + 2M/r^2(r_R - r) + O(r_R - r)) \\
 &\quad \times (\mathcal{F}(U, U) + \partial_1\mathcal{F}(U, U)(U_R - U) + o(U_R - U)) \\
 &\quad - (1 - 2M/r + 2M/r^2(r_L - r) + O(r_L - r)) \\
 &\quad \times (\mathcal{F}(U, U) + \partial_2\mathcal{F}(U, U)(U_L - U) + o(U_L - U)).
 \end{aligned}$$

By (8-8),  $U_R - U_L = O(r_R - r_L)S(r, U)$ . Moreover, since  $U = U(t, r)$  is exact, we have  $\mathcal{F}(U, U) = F(U)$  and  $\partial_1\mathcal{F}(U, U) = \partial_2\mathcal{F}(U, U) = \partial_U F(U)$ . Therefore,

$$\begin{aligned}
 & \mathcal{F}_r(r_R, U_L, U_R) - \mathcal{F}_l(r_L, U_L, U_R) \\
 &= \frac{2M}{r^2}(r_R - r_L)F(U) + (1 - 2M/r)\partial_U F(U)(U_R - U_L) + O((r_R - r_L)^2) \\
 &= S(r, U)(r_R - r_L) + O((r_R - r_L)^2).
 \end{aligned}$$

Next, a Taylor expansion with respect to time yields us

$$U_j^{n+1} = U_j^n + \partial_t U_j^n \Delta t + \partial_{tt}^2 U_j^n \Delta t^2 + O(\Delta t^3).$$

Recall that our scheme gives

$$\begin{aligned}
 U_j^{n+1} &= U_j^n - \frac{\Delta t}{\Delta r} \left( (1 - 2M/r_{j+1/2})F_{j+1/2}^n - (1 - 2M/r_{j-1/2})F_{j-1/2}^n - \Delta r S_j^n \right). \\
 &= U_j^n - \frac{1}{\lambda} \left( (1 - 2M/r_{j+1/2}) \left( \frac{F(U_{j+1/2+}) - F(U_{j+1/2-})}{2} - \frac{1}{\lambda} \frac{U_{j+1/2+} - U_{j+1/2-}}{2} \right) \right. \\
 &\quad \left. + (1 - 2M/r_{j-1/2}) \left( \frac{F(U_{j-1/2+}) - F(U_{j-1/2-})}{2} + \frac{1}{\lambda} \frac{U_{j-1/2+} - U_{j-1/2-}}{2} \right) \right).
 \end{aligned}$$

According our construction, we have

$$\begin{aligned}
 & \left( 1 - \frac{2M}{r_{j+1/2}} \right) (F(U_{j+1/2+}) - F(U_{j+1/2-})) \\
 &= \left( 1 - \frac{2M}{r_{j+1}} \right) F(U_{j+1}^n) - \left( 1 - \frac{2M}{r_j} \right) F(U_j^n) - \int_{r_j}^{r_{j+1}} S(r, U(t_n, r)) dr.
 \end{aligned}$$

A Taylor expansion to  $\Delta r$  gives us  $U_{j+1/2+} - U_{j+1/2-} = O(\Delta r^3)$  and

$$\begin{aligned} \left(1 - \frac{2M}{r_{j\pm 1}}\right) &= 1 - \frac{2M}{r_j} \pm \frac{2M}{r_j^2} \Delta r - \frac{2M}{r_j^3} \Delta r^2 + O(\Delta r^3), \\ F(U_{j\pm 1}^n) &= F(U_j^n) + \partial_U F(U_j^n)(\pm \partial_r U_j^n \Delta r + \frac{1}{2} \partial_{rr}^2 U_j^n \Delta r^2) \\ &\quad + \frac{1}{2} (\partial_r U_j^n)^T \partial_{UU}^2 F(U_j^n) \partial_r U_j^n \Delta r^2 + O(\Delta r^3), \\ \int_{r_j}^{r_{j+1}} S(r, U(t_n, r)) dr &= S(r_j, U_j^n) \Delta r + \partial_r S(r_j, U_j^n) \Delta r^2 + O(\Delta r^3). \end{aligned}$$

Hence, we conclude  $\partial_t U_j^n + \partial_r((1 - 2M/r_j)F(U_j^n)) - S(r_j, U_j^n) + O(\Delta t + \Delta r^2) = 0$ . □

*Numerical steady state solution.* Recall that the steady state solution to the relativistic Euler model is given by a static Euler system (7-1). Hence, if  $U = U(t, r)$  is a steady state solution, it trivially satisfies  $\int |\partial_r F((1 - 2M/r)U) - S(r, U)| dr = 0$ , where  $F = (F^0, F^1)^T$  is the flux and  $S = (S^0, S^1)^T$  the source term given by (8-1). In order to describe the steady state solution numerically, we define the total variation in time

$$\begin{aligned} E^n := E(t_n) &= \sum_j \sum_{i=0,1} |(1 - 2M/r_{j+1/2})(F^i(U_{j+1/2+}^n) - F^i(U_{j-1/2-}^n)) \\ &\quad - (1 - 2M/r_{j-1/2})(F^i(U_{j-1/2+}^n) - F^i(U_{j-1/2-}^n))|. \end{aligned} \tag{8-11}$$

Clearly, we have the following property.

**Claim 8.2.** *If  $U = U(t, r)$  is a numerical solution to the relativistic Euler model constructed by (8-5)–(8-9), then  $U$  is a steady state solution (smooth or with a shock) for  $t \geq T$  where  $T > 0$  is a finite time if and only if there exists an integer  $N$  such that, for all  $n > N$ , the total variation  $E^n \equiv 0$ .*

### 9. Numerical experiments for the relativistic Euler model

*Nonlinear stability of steady state solutions.* Before studying the stability of steady state solutions, we check that our scheme preserves smooth steady state solutions to the relativistic Euler model (1-4). Recall that  $r > 2M$  with  $M = 1$  being the black hole mass. We work on the space interval  $(r_{\min}, r_{\max})$  with  $r_{\min} = 2M = 2$  and  $r_{\max} = 10$ , and we take 500 points to discretize this interval. We consider the evolution of two steady state solutions satisfying the algebraic relation (7-2) of the Euler model with the density  $\rho(10) = 1.0$  and velocity  $v(10) = 0.6$  and the density  $\rho(10) = 1.0$  and velocity  $v(10) = -0.8$ , respectively. We also provides the evolution of a steady state shock (see Figures 15 and 16).



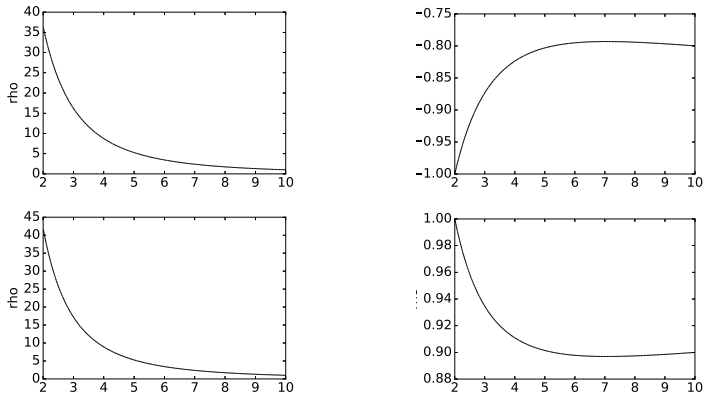


Figure 15. Euler: evolution of steady state solutions plotted at time  $t = 50$ .

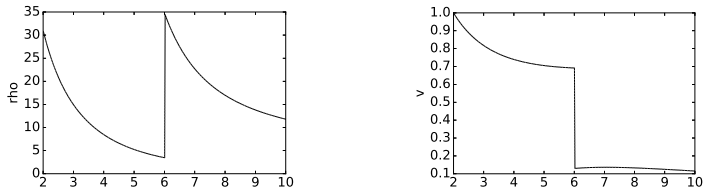


Figure 16. Euler: evolution of a steady shock plotted at time  $t = 50$ .

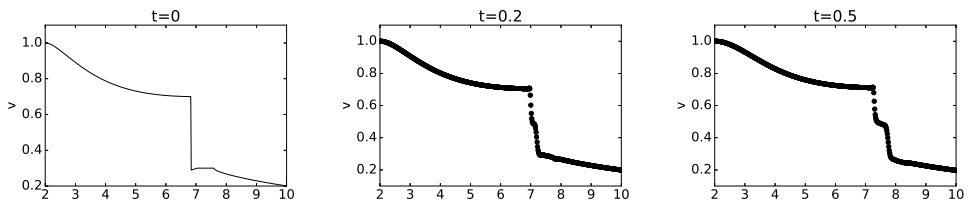


Figure 17. Euler: solution to the Riemann problem (1-shock and 2-shock).

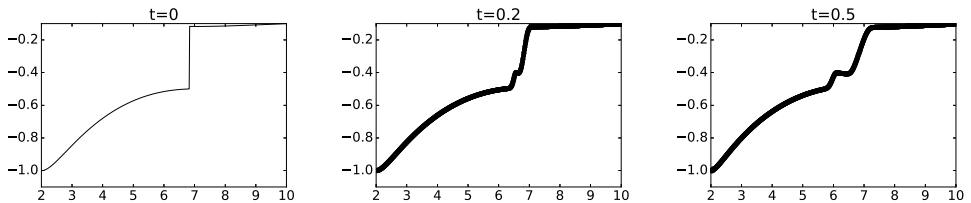
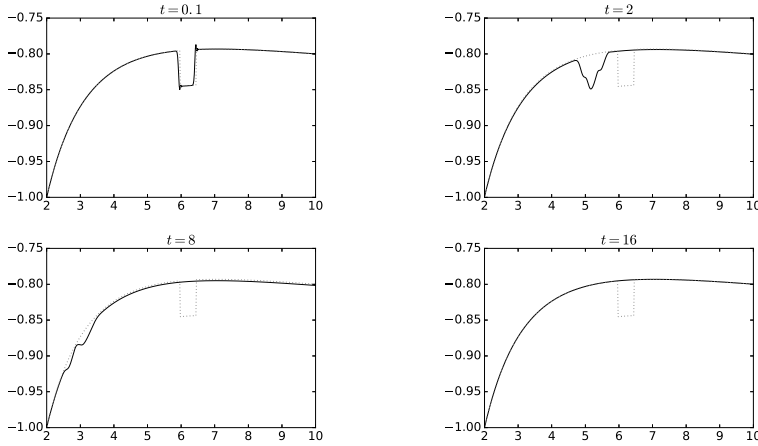
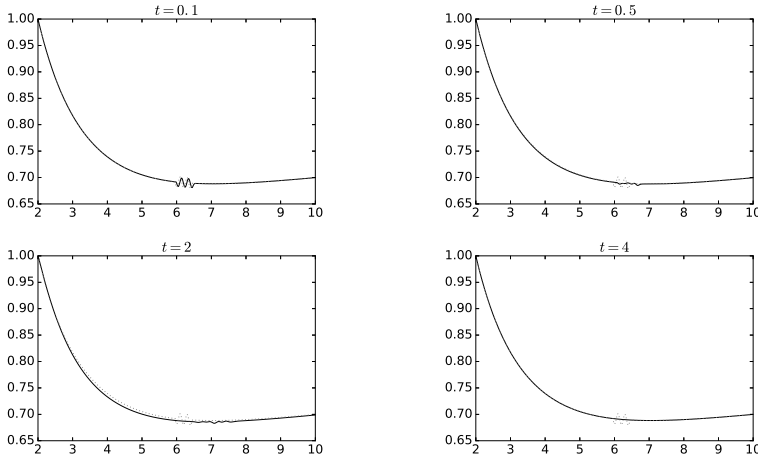


Figure 18. Euler: solution to the Riemann problem (1-rarefaction and 2-rarefaction).

*Propagation of discontinuities.* Referring to [22], we recall that there exists a solution to the generalized Riemann problem (1-3) with (7-7) consisting of at most three steady state solutions. Figures 17 and 18 show the evolution of two



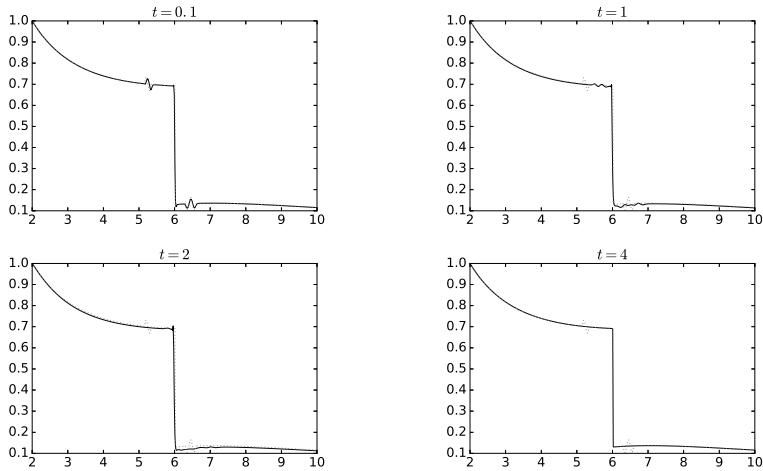
**Figure 19.** Euler: evolution of a perturbed steady state — convergence to the same steady state.



**Figure 20.** Euler: evolution of a perturbed steady state — convergence to the same steady state.

generalized Riemann problems with an initial discontinuity. Furthermore, we are now interested in the late-time behavior of solutions whose initial data is a steady state solution perturbed by a compactly supported solution. Numerical tests lead us to the following result.

**Conclusion 9.1** (stability of smooth steady state solutions to the Euler model). *Let  $(\rho_*, v_*) = (\rho_*, v_*)(r)$ ,  $r > 2M$ , be a smooth steady state solution to the Euler equations (7-1) and  $(\rho_0, v_0) = (\rho_0, v_0)(r) = (\rho_*, v_*)(r) + (\delta_\rho, \delta_v)(r)$  where  $(\delta_\rho, \delta_v) = (\delta_\rho, \delta_v)(r)$  is a function with compact support; then the solution to the relativistic Euler equations on a Schwarzschild background (1-4) denoted by  $(\rho, v) = (\rho, v)(t, r)$  satisfies that  $(\rho, v)(t, \cdot) = (\rho_*, v_*)$  for all  $t > t_0$  where  $t_0 > 0$*



**Figure 21.** Euler: evolution of a perturbed steady state shock — convergence to the same steady state.

is a finite time. Numerical experiments show that there exists a finite time  $t_0 > 0$  such that  $(\rho, v)(t, r) = (\rho_*, v_*)(r)$  for all  $t > t_0$ .

We observe the phenomenon described in Claim 1.3 in Figures 19 and 20, where we have plotted the evolution of different steady state solutions to the Euler model with an initial perturbation. The steady shock given by (7-5) and (7-6) is a weak solution satisfying the static Euler equations (7-1). As is done in the Burgers model, we are also interested in the behavior of steady shocks with perturbations. We summarize our results as follows; see Figure 21.

**Conclusion 9.2.** Consider a steady shock  $(\rho_*, v_*) = (\rho_*, v_*)(r)$ ,  $r > 2M$ , given by (7-5) and (7-6) whose point of discontinuity is at  $r = r_*$ , and we give the initial data  $(\rho_0, v_0) = (\rho_0, v_0)(r) = (\rho_*, v_*)(r) + (\delta_\rho, \delta_v)(r)$  with  $(\delta_\rho, \delta_v) = (\delta_\rho, \delta_v)(r)$  a compactly supported function; then there exists a finite time  $t > t_0$  such that, for all  $t > t_0$ , the solution is a steady state shock.

### References

[1] P. Amorim, P. G. LeFloch, and B. Okutmustur, *Finite volume schemes on Lorentzian manifolds*, Commun. Math. Sci. **6** (2008), no. 4, 1059–1086. MR Zbl

[2] Y. Bakhtin and P. G. LeFloch, *Ergodicity of spherically symmetric fluid flows outside of a Schwarzschild black hole with random boundary forcing*, preprint, 2017, to appear in *Stoch. Part. Diff. Eq. Anal. Comput.* arXiv

[3] A. Beljadid and P. G. LeFloch, *A central-upwind geometry-preserving method for hyperbolic conservation laws on the sphere*, Commun. Appl. Math. Comput. Sci. **12** (2017), no. 1, 81–107. MR

- [4] M. Ben-Artzi, J. Falcovitz, and P. G. LeFloch, *Hyperbolic conservation laws on the sphere: a geometry-compatible finite volume scheme*, J. Comput. Phys. **228** (2009), no. 16, 5650–5668. MR Zbl
- [5] S. Boscarino, R. Bürger, P. Mulet, G. Russo, and L. M. Villada, *Linearly implicit IMEX Runge–Kutta methods for a class of degenerate convection-diffusion problems*, SIAM J. Sci. Comput. **37** (2015), no. 2, B305–B331. MR Zbl
- [6] S. Boscarino, P. G. LeFloch, and G. Russo, *High-order asymptotic-preserving methods for fully nonlinear relaxation problems*, SIAM J. Sci. Comput. **36** (2014), no. 2, A377–A395. MR Zbl
- [7] B. Boutin, F. Coquel, and P. G. LeFloch, *Coupling techniques for nonlinear hyperbolic equations, IV: Well-balanced schemes for scalar multi-dimensional and multi-component laws*, Math. Comp. **84** (2015), no. 294, 1663–1702. MR Zbl
- [8] T. Ceylan, P. G. LeFloch, and O. Bayer, *A finite volume method for the relativistic Burgers equation on a FLRW background spacetime*, Commun. Comput. Phys. **23** (2018), no. 2, 500–519.
- [9] A. J. Chorin, *Random choice solution of hyperbolic systems*, J. Comput. Phys. **22** (1976), no. 4, 517–533. MR Zbl
- [10] C. M. Dafermos and L. Hsiao, *Hyperbolic systems and balance laws with inhomogeneity and dissipation*, Indiana Univ. Math. J. **31** (1982), no. 4, 471–491. MR Zbl
- [11] C. M. Dafermos, *Hyperbolic conservation laws in continuum physics*, 3rd ed., Grundlehren der mathematischen Wissenschaften, no. 325, Springer, 2010. MR Zbl
- [12] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math. **18** (1965), 697–715. MR Zbl
- [13] J. Glimm, G. Marshall, and B. Plohr, *A generalized Riemann problem for quasi-one-dimensional gas flows*, Adv. in Appl. Math. **5** (1984), no. 1, 1–30. MR Zbl
- [14] J. M. Hong and P. G. LeFloch, *A version of the Glimm method based on generalized Riemann problems*, Port. Math. (N.S.) **64** (2007), no. 2, 199–236. MR Zbl
- [15] P. D. Lax, *Hyperbolic systems of conservation laws, II*, Comm. Pure Appl. Math. **10** (1957), 537–566. MR Zbl
- [16] P. D. Lax, *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, no. 11, SIAM, 1973. MR Zbl
- [17] P. G. LeFloch, *Hyperbolic systems of conservation laws: the theory of classical and nonclassical shock waves*, Birkhäuser, 2002. MR Zbl
- [18] ———, *Structure-preserving shock-capturing methods: late-time asymptotics, curved geometry, small-scale dissipation, and nonconservative products*, Advances in numerical simulation in physics and engineering: lecture notes of the XV “Jacques-Louis Lions” Spanish-French School (C. Parés, C. Vázquez, and F. Coquel, eds.), SEMA SIMAI, no. 3, Springer, 2014, pp. 179–222. Zbl
- [19] P. G. LeFloch and H. Makhlof, *A geometry-preserving finite volume method for compressible fluids on Schwarzschild spacetime*, Commun. Comput. Phys. **15** (2014), no. 3, 827–852. MR
- [20] P. G. LeFloch, H. Makhlof, and B. Okutmustur, *Relativistic Burgers equations on curved spacetimes: derivation and finite volume approximation*, SIAM J. Numer. Anal. **50** (2012), no. 4, 2136–2158. MR Zbl
- [21] P. G. LeFloch and P.-A. Raviart, *An asymptotic expansion for the solution of the generalized Riemann problem, I: General theory*, Ann. Inst. H. Poincaré Anal. Non Linéaire **5** (1988), no. 2, 179–207. MR

- [22] P. G. LeFloch and S. Xiang, *Weakly regular fluid flows with bounded variation on the domain of outer communication of a Schwarzschild black hole spacetime*, *J. Math. Pures Appl.* (9) **106** (2016), no. 6, 1038–1090. MR Zbl
- [23] ———, *Weakly regular fluid flows with bounded variation on the domain of outer communication of a Schwarzschild black hole spacetime, II*, preprint, 2017, to appear in *J. Math. Pures Appl.*
- [24] T. T. Li, *A note on the generalized Riemann problem*, *Acta Math. Sci.* (English ed.) **11** (1991), no. 3, 283–289. MR Zbl
- [25] T. T. Li and L. Wang, *The generalized nonlinear initial-boundary Riemann problem for quasi-linear hyperbolic systems of conservation laws*, *Nonlinear Anal.* **62** (2005), no. 6, 1091–1107. MR
- [26] T. P. Liu, *Invariants and asymptotic behavior of solutions of a conservation law*, *Proc. Amer. Math. Soc.* **71** (1978), no. 2, 227–231. MR Zbl
- [27] T. Morales de Luna, M. J. Castro Díaz, and C. Parés, *Reliability of first order numerical schemes for solving shallow water system over abrupt topography*, *Appl. Math. Comput.* **219** (2013), no. 17, 9012–9032. MR Zbl
- [28] G. Russo, *Central schemes for conservation laws with application to shallow water equations*, *Trends and applications of mathematics to mechanics: STAMM 2002* (S. Rionero and G. Romano, eds.), Springer, 2005, pp. 225–246.
- [29] ———, *High-order shock-capturing schemes for balance laws*, *Numerical solutions of partial differential equations*, Birkhäuser, 2009, pp. 59–147. MR
- [30] M. Semplice, A. Coco, and G. Russo, *Adaptive mesh refinement for hyperbolic systems based on third-order compact WENO reconstruction*, *J. Sci. Comput.* **66** (2016), no. 2, 692–724. MR Zbl
- [31] B. van Leer, *On the relation between the upwind-differencing schemes of Godunov, Engquist–Osher and Roe*, *SIAM J. Sci. Statist. Comput.* **5** (1984), no. 1, 1–20. MR Zbl

Received August 11, 2017. Revised April 17, 2018.

PHILIPPE G. LEFLOCH: [contact@philippelefloch.org](mailto:contact@philippelefloch.org)  
*Laboratoire Jacques-Louis Lions, Centre National de la Recherche Scientifique, Sorbonne Université, Paris, France*

SHUYANG XIANG: [xiang@ljl1.math.upmc.fr](mailto:xiang@ljl1.math.upmc.fr)  
*Mathematics Division, Center for Computational Engineering Science, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany*



## A SEMI-IMPLICIT MULTISCALE SCHEME FOR SHALLOW WATER FLOWS AT LOW FROUDE NUMBER

STEFAN VATER AND RUPERT KLEIN

A new large time step semi-implicit multiscale method is presented for the solution of low Froude number shallow water flows. While on small scales which are under-resolved in time the impact of source terms on the divergence of the flow is essentially balanced, on large resolved scales the scheme propagates free gravity waves with minimized diffusion. The scheme features a scale decomposition based on multigrid ideas. Two different time integrators are blended at each scale depending on the scale-dependent Courant number for gravity wave propagation. The finite volume discretization is implemented in the framework of second-order Godunov-type methods for conservation laws. The basic properties of the method are validated by numerical tests. This development is a further step in the construction of asymptotically adaptive numerical methods for the computation of large-scale atmospheric flows.

### 1. Introduction

Modern high-performance computing hardware allows for high-resolution atmospheric flow simulations, which resolve scales ranging from small convective-scale essentially anelastic flows up to large planetary-scale dynamics (see, e.g., [32]). Such simulations are not only demanding in terms of problem size. They also challenge the applied numerical methods, which must correctly resolve the different characteristic flow regimes arising on the different scales captured by the discretization.

An example is the influence of sound waves and the associated compressibility. These waves are usually considered to have little influence in meteorological applications, because the much slower synoptic and planetary wave patterns associated with inertia and advection are most relevant for predicting the weather. This led to so-called approximate “sound-proof” model equations [31; 28; 1; 9], which do not include the fast acoustic waves and have been quite popular to model small-scale atmospheric dynamics. The situation is different for planetary-scale

---

*MSC2010:* 65M08, 86A10.

*Keywords:* shallow water equations, multiscale time integration, asymptotically adaptive numerical methods, large time steps, balanced modes.

dynamics, where long-wave horizontally traveling acoustic modes, i.e., Lamb waves, are sometimes considered nonnegligible. Furthermore, there are indications that effects of compressibility affect large-scale, deep internal wave modes of the atmosphere in a nontrivial fashion [7]. These dynamics are fairly captured by another reduced model, namely the hydrostatic primitive equations (HPEs), which are widely used in current operational general circulation models. At resolutions of only a few kilometers, however, the HPEs lose their validity due to the breakdown of the hydrostatic assumption. Therefore, at least for an accurate representation of large-scale planetary-scale dynamics, the challenge arises of combining large-scale compressible flow representations with essentially sound-proof modeling of the small-scale dynamics.

For the mathematical study of interactions across scales, techniques from multiple-scales asymptotics [16; 22] are increasingly used. These are extensions of the classical single-scale asymptotic method (also known as regular perturbation analysis). In the latter, a small nondimensional parameter of the problem and a special (asymptotic) expansion of the dependent variables are employed to obtain simplified equation sets, which still account for the physical effects characteristic to the specific scale. Examples are the aforementioned anelastic and hydrostatic approximations. In multiple-scales asymptotic analysis the asymptotic expansion is generalized in that the variables artificially depend on more than one space or time scale. This enables the study of effects arising across scales. Since the asymptotic analysis directly relates a reduced model to the full compressible flow equations, it is a natural starting point for the development of numerical methods applicable to the considered singular regimes [20; 22]. In this context, the notion of “asymptotically adaptive numerical methods” was suggested in [18; 19; 24]. Such schemes should be robust, uniformly accurate, and efficient in the vicinity of certain asymptotic regimes and over a variety of relevant applications. The idea is closely related to “asymptotic preserving” schemes (see [5] and references therein).

The aim of this work is to develop such an asymptotically adaptive numerical method that is able to correctly simulate large-scale compressible flow phenomena with high resolution. In this initial attempt not the full dynamics of the atmosphere are considered. Instead, this work deals with the shallow water equations, which describe the vertically averaged motion of an incompressible fluid with a free surface. By “shallow” one refers to the small aspect ratio between the vertical depth and a typical horizontal length scale of the problem, which justifies the hydrostatic assumption, i.e., that the pressure balances the weight of the fluid. However, these equations are not only a good model for representing river flow or large-scale oceanic motions (such as tsunamis). While ignoring the presence of stratification, the shallow water equations incorporate the effects of gravity and can account for the Earth’s rotation and for bottom topography by the addition of appropriate



source terms. Therefore, they are prototypical of the hydrostatic primitive equations and are often used in the development of numerical methods for atmospheric flow problems.

Due to the vertical averaging of the prognostic variables, the shallow water equations only admit external waves. However, the external gravity waves in shallow water flows are the equivalent to Lamb waves in the compressible flow equations [15]. The considered asymptotic regime consists of long-wave acoustic waves (Lamb waves) interacting with slow advection. This is equivalent to the regime of fast gravity waves moving over short-range topography in the shallow water context. The additional atmospheric effect of small-scale flow divergence induced by local diabatic sources is modeled here by a time-dependent bottom topography. Such effects are important when incorporating moist atmospheric processes, such as condensation and rain. In this context, the shallow water model represents a challenging part of the development of numerical methods for the simulation of planetary-scale atmospheric flows at high resolution.

The analysis of the regime of fast gravity waves moving over short-range topography reveals that it essentially consists of long-wave linearized shallow water flow interacting with small-scale flow balancing the influence of the rough topography (see Section 2 for details and [4] by Bresch et al.). Therefore, the new scheme should

- eliminate freely propagating “compressible” short-wave modes that it cannot represent accurately due to temporal under-resolution,
- represent with second-order accuracy the “slaved” dynamics of short-wave solution components induced by slow forcing or arising in the form of high-order corrections to long-wave modes, and
- minimize numerical dispersion for resolved modes.

The first and last points address the specific numerical dispersion behavior of common second-order implicit time discretizations, which usually slow down modes with high wavenumbers [43; 10]. While the decision which modes can be considered to be resolved is certainly subjective and depends on the application, at some point the slowdown of modes with wavenumbers larger than a certain value is unacceptable. These modes should be eliminated over time in a consistent way. On the other hand, long-wave modes, whose oscillation is well resolved at a fixed position, should be well approximated. The second point refers to the balanced flow on the small scale of the regime, which depends on local source terms and the coupling to the large-scale dynamics.

To achieve these goals, a semi-implicit method for the nonlinear shallow water equations is combined with a multilevel approach which has successfully been applied to the linearized equations to model multiscale behavior in [43]. The latter enables the association of different solution components with certain spatial scales

and is based on geometric multigrid ideas. Furthermore, selective to each scale, a proper discretization is applied. The approach results in a robust representation of balanced, slowly forced fast modes on the one hand, and a proper propagation of long-wave gravity waves on the other hand.

The present work extends ideas of multiscale time integration for compressible flows formulated earlier in [18; 12; 30]. These authors already suggested separating the short- and long-wave components of a flow field and to propagate these components in time by different time integration schemes. However, they only allowed for two distinct discrete scales: one representing small-scale solution components and one for long-wave acoustic modes, which are separated from each other by a factor of  $1/M$ , where  $M$  is the Mach number. In contrast, by introducing multigrid decompositions of the flow and a smooth blending of time integrators, we obtain a scheme in this work that allows for much more general data with true multiscale content. Our work extends that of [43] from linear wave propagation in one space dimension to the nonlinear shallow water equations.

This article is structured into the following parts. After the presentation of the governing equations we discuss the asymptotic regime of interest in the next section. The multiscale scheme is then described by a semidiscretization in time in Section 3. In this course, we first extend a zero Froude number projection method to nonzero Froude numbers. The multilevel approach is included in the implicit correction step, which accounts for the correct propagation of gravity waves. Finally, we show the correct behavior of the method by some one-dimensional test cases in Section 4 and give conclusions in Section 5.

## 2. Governing equations

The derivation of the shallow water equations can be found in numerous textbooks (see, e.g., [34; 38]). The case of nonstationary bottom topography was dealt with in [41]. Here, only the resulting equations are presented and the peculiarities concerning time-dependent bottom topography are pointed out. Furthermore, the governing equations are analyzed in the limit of a small Froude number. Particularly, the asymptotic limit regime for long-wave shallow water waves passing over short-range topography as presented in [4] is discussed under the additional assumption of bottom topography changing in time.

**2.1. Shallow water flows with time-dependent bottom topography.** The assumption of a time-dependent bottom topography, which is slightly unusual, is considered to model a source term which acts on the local flow divergence as outlined in the introduction. This generalization neither changes the terms arising in the shallow water equations nor does it introduce additional ones. Therefore, the governing

equations in conservation form are given by

$$\begin{aligned}
 h_t + \nabla \cdot (h\mathbf{u}) &= 0, \\
 (h\mathbf{u})_t + \nabla \cdot (h\mathbf{u} \circ \mathbf{u}) + \frac{1}{2\text{Fr}^2} \nabla(h^2) &= -\frac{1}{\text{Fr}^2} h \nabla b.
 \end{aligned} \tag{1}$$

Here,  $h(t, \mathbf{x})$  is the thickness or depth of the fluid and  $\mathbf{u}(t, \mathbf{x})$  its depth-averaged horizontal velocity, and  $b(t, \mathbf{x})$  denotes the time- and space-dependent bottom topography. The gradient operator  $\nabla$  is acting in the horizontal  $\mathbf{x} = (x, y)$  plane. The “ $\circ$ ” denotes the dyadic product of two vectors. A temporal change in bottom topography either changes the total height  $H = h + b$  or introduces divergence in the momentum field, as can be seen from reformulating the continuity equation to

$$H_t + \nabla \cdot (h\mathbf{u}) = b_t. \tag{2}$$

Furthermore, a change in the gradient of  $b$  directly enters the source term of the momentum equation, leading to a potential disruption of the hydrostatic equilibrium of a previously balanced flow. System (1) is given in nondimensional form introducing the dimensionless characteristic quantity  $\text{Fr} := v_{\text{ref}}/\sqrt{gh_{\text{ref}}}$ , which is known as the Froude number. It defines the ratio between the characteristic flow velocity  $v_{\text{ref}}$  and the gravity wave speed  $\sqrt{gh_{\text{ref}}}$  with  $g$  being the acceleration due to gravity and  $h_{\text{ref}}$  a reference fluid depth. Since we are interested in phenomena associated with the advective time scale of the fluid, we set  $t_{\text{ref}} = \ell_{\text{ref}}/v_{\text{ref}}$  in the dimensional analysis and omitted mentioning of the Strouhal number.

The shallow water equations are mathematically equivalent to the Euler equations of compressible isentropic gas dynamics for an isentropic exponent of  $\gamma = 2$ . In this respect, the Froude number in the shallow water equations takes the role of the Mach number in the Euler equations, the latter being a measure of the compressibility of the fluid. Therefore, effects similar to compressibility can also be modeled by the shallow water equations, where the importance of the “compressibility” depends on the associated scales of fluid motion. In large scale atmospheric applications, a typical flow velocity is 10 m/s and the depth of the atmosphere is given by the pressure scale height, which is approximately 10 km. This results in a Froude number  $\text{Fr} \approx 0.03 \ll 1$ , and the “compressibility” effects associated with the nonlinear nature of external gravity waves plays a minor role in this regime. Note, however, that the shallow water equations intrinsically model an incompressible fluid.

**2.2. Long-wave gravity waves passing over short-range topography.** The regime of particular interest can be characterized by long-wave shallow water waves traveling over rough topography. Consider a multiple-space-scale/single-time-scale analysis for this regime akin to [18; 4; 25]. In addition to the space coordinate  $\mathbf{x}$  defined by nondimensionalization with the reference length  $\ell_{\text{ref}}$ , a second large-scale coordinate  $\xi = \text{Fr}\mathbf{x}$  is introduced, which resolves the distance a gravity wave

traverses on the considered time scale. For the bottom topography  $b = b(t, \mathbf{x}, \boldsymbol{\xi})$  we allow for variations on both space scales. Then, the fluid depth and velocity are expressed in the multiple-scales expansion

$$(h, \mathbf{u})(t, \mathbf{x}; \text{Fr}) = \sum_{i=0}^N \text{Fr}^i (h, \mathbf{u})^{(i)}(t, \mathbf{x}, \boldsymbol{\xi}) + \mathcal{O}(\text{Fr}^N). \quad (3)$$

Note that, by using this ansatz, each spatial derivative of an asymptotic function  $\varphi^{(i)}$  translates into

$$\nabla \varphi^{(i)}|_{\text{Fr}} = \nabla_{\mathbf{x}} \varphi^{(i)} + \text{Fr} \nabla_{\boldsymbol{\xi}} \varphi^{(i)} \quad (4)$$

for fixed Froude number  $\text{Fr}$ . As stated above, this regime has also been discussed in [4] but without time-dependent bottom topography. The leading-order system is separated into two subsystems representing the long-wave and the short-wave components of the flow. They are given by the *long-wave equations for rough topography*

$$\begin{aligned} \overline{(h\mathbf{u})^{(0)}}_t + \overline{h^{(0)}} \nabla_{\boldsymbol{\xi}} h^{(1)} &= \overline{h^{(2)}} \nabla_{\mathbf{x}} h^{(0)}, \\ h_t^{(1)} + \nabla_{\boldsymbol{\xi}} \cdot \overline{(h\mathbf{u})^{(0)}} &= 0 \end{aligned} \quad (5)$$

and the associated *balanced small-scale flow*

$$\begin{aligned} \widetilde{(h\mathbf{u})^{(0)}}_t + \nabla_{\mathbf{x}} \cdot (h\mathbf{u} \circ \mathbf{u})^{(0)} + \widetilde{h^{(0)}} \nabla_{\mathbf{x}} h^{(2)} &= -\widetilde{h^{(0)}} \nabla_{\boldsymbol{\xi}} h^{(1)}, \\ \nabla_{\mathbf{x}} \cdot (h\mathbf{u})^{(0)} &= \nabla_{\mathbf{x}} \cdot \widetilde{(h\mathbf{u})^{(0)}} = \widetilde{b}_t. \end{aligned} \quad (6)$$

The leading-order fluid depth is given by

$$h^{(0)}(t, \mathbf{x}, \boldsymbol{\xi}) = H^{(0)}(t) - b(t, \mathbf{x}, \boldsymbol{\xi}),$$

where  $H^{(0)}$  is the leading-order surface elevation of the fluid and  $dH^{(0)}/dt = \bar{b}_t$ . The next order of the fluid depth  $h^{(1)} = h_1(t, \boldsymbol{\xi})$  is independent of  $\mathbf{x}$ . Here the overbar denotes the average of the pertinent variable in the fast coordinate,  $\mathbf{x}$ , and the wiggly overline indicates the zero-average remainder or fluctuation.

Compared to the linear case (cf. [41]) the two systems (5) and (6) are coupled. The large-scale flow is given by the linearized shallow water equations, which involve nonbalanced free surface waves. It is driven by a source term arising from the small-scale flow in the momentum equation. This source represents the accumulated pressure force, which results from the small-scale flow across the rough topography. In the opposite direction, large-scale gradients of the fluid depth acting on the rough topography induce small-scale momentum. This modifies the otherwise balanced small-scale flow.

In contrast to [4] a source term acting on the local divergence of the flow arises, when considering nonstationary bottom topography. It is generated by local variations in time of the bottom topography. Furthermore, the changes of the mean

in  $b$  over time induce a change in the leading-order surface elevation  $H^{(0)}$ , and the signal speed of the long-wave gravity waves is changing not only in space, but also in time.

Similar asymptotic regimes were studied in [18] concerning weakly compressible flows with small-scale entropy and vorticity, in [26] for modeling ocean flows, and in the context of atmospheric circulation near the equator in [29].

The asymptotic scaling for the velocity in this regime is given by  $\mathbf{u} \sim 1$  as  $Fr \rightarrow 0$ . For the fluid depth we have  $h - h_0(t) \sim Fr$  on the large scale and  $h - h_0(t) \sim Fr^2$  on the small scale. This scaling should be reproduced by a numerical scheme, especially when  $\Delta t \gg \Delta \xi / \sqrt{H_0} = Fr \Delta x / \sqrt{H_0}$ , the latter corresponding to large Courant numbers with respect to gravity waves for the time integration in the present model problem.

**2.3. From zero to low Froude numbers.** To be able to extend the numerical machinery known from projection methods applied to the zero Froude number shallow water equations (also known as “Lake equations”), the shallow water equations must be cast into a similar form. To reformulate system (1), let us decompose the fluid depth into

$$h(t, \mathbf{x}; Fr) = h_0(t, \mathbf{x}) + Fr^2 h'(t, \mathbf{x}) \tag{7}$$

with

$$h_0(t, \mathbf{x}) = H_0(t) - b(t, \mathbf{x}). \tag{8}$$

Here,  $H_0$  is the mean background total elevation, which can only change due to flow over the boundary of the domain or to a change in the mean bottom topography. Therefore,  $h_0$  can only change due to boundary flow or (local) change of bottom topography. The dynamics of the flow are thus given by the perturbation  $h'$  of the fluid depth. This ansatz is justified by the asymptotic analysis of the zero Froude number limit of the governing equations, and we expect that  $h' = \mathcal{O}(1)$  as  $Fr \rightarrow 0$  in the flow regimes of interest. Inserting this into the governing system, the shallow water equations can be rewritten as

$$\begin{aligned} h_t + \nabla \cdot (h\mathbf{u}) &= 0, \\ (h\mathbf{u})_t + \nabla \cdot (h\mathbf{u} \circ \mathbf{u}) + h\nabla h' &= 0, \\ h &= h_0 + Fr^2 h'. \end{aligned} \tag{9}$$

Compared to the zero Froude number equations,  $h'$  takes the role of  $h^{(2)}$ , but is no longer a Lagrange multiplier. Therefore, also the velocity no longer satisfies a strict divergence constraint. However, at low Froude numbers, these fields should be close to their zero Froude number counterparts. This is due to the mathematical equivalence of the shallow water and the Euler equations and related convergence results for the low Mach number limit of the Euler equations (see, e.g., [17]).

### 3. Numerical scheme

The numerical scheme to correctly capture the multiscale behavior of the flow is based on a semi-implicit discretization of the shallow water equations, the latter being an extension of a zero Froude number projection method as in [42]. This construction ensures that the discretization correctly approximates the limit behavior of the equations. A second ingredient is a scale-selective multilevel scheme which was previously derived for the linearized equations [43; 41]. With this addition we account for the characteristic flow behavior on the different scales resolved by the discretization.

The semi-implicit method consists of a predictor step, which solves an auxiliary hyperbolic system. This is followed by a first elliptic correction to adjust the advective flux components. A second elliptic correction accounts for the accurate propagation of gravity waves. This is where we incorporate the multilevel scheme for linearized flows. The multilevel scheme is based on two different time discretizations. A scalewise decomposition of the flow information based on geometric multigrid ideas enables a scale-dependent blending of the two time discretizations. Here, we employ the implicit midpoint rule and the BDF(2) scheme, which are both second-order accurate and need the solution of only one linear system. The implicit midpoint rule conserves energy of all wave modes. While this is advantageous for long waves, it is not desirable for high-wavenumber modes, due to the unfavorable discrete dispersion relation. Backward differentiation (BDF) schemes, on the other hand, are able to filter these short-wave modes in a consistent way. In the present work, only uniform time steps are considered. This simplifies the application of multistep methods, since it is not required to account for the different time step sizes. Often these methods can be generalized to variable time steps as in the case of BDF(2) [8].

Similar to the formulation of a zero Froude number projection method as in [42], the semi-implicit scheme is derived by a semidiscretization in time. The discretization in space is discussed in a second step. The essential difference from the zero Froude number case is that the ansatz (7) for the fluid depth involves the introduction of local time derivatives of this quantity. This leads to the solution of two Helmholtz problems in the correction steps.

**3.1. *Explicit predictor and advective flux correction.*** The auxiliary system solved in the predictor step is given by

$$\begin{aligned} h_t + \nabla \cdot (h\mathbf{u}) &= 0, \\ (h\mathbf{u})_t + \nabla \cdot (h\mathbf{u} \circ \mathbf{u}) &= -(h\nabla h')^{\text{old}}, \end{aligned} \tag{10}$$

where the right-hand side of the momentum equation is treated as a “source term” and computed from an old (known) time level. The homogeneous part of (10) is known as the “pressureless equations” (see [2; 3; 27] and references therein).

The source term is set to  $(h\nabla h')^{\text{old}}(\mathbf{x}) := (h\nabla h')(t^n, \mathbf{x})$ , where  $h'^n$  is computed from  $h^n$  by using (7), i.e.,

$$h'^n = \frac{1}{\text{Fr}^2}(h^n - H_0^n + b^n). \quad (11)$$

Here and in the following  $h(\mathbf{x}, t^n)$  is abbreviated by  $h^n$ , etc.

Integrating the governing equations from time level  $t^n$  to  $t^{n+1} := t^n + \Delta t$  and using the midpoint rule by evaluating the flux terms at the half time levels  $t^{n+1/2} := t^n + \Delta t/2$  yields

$$h^{n+1} = h^n - \Delta t[\nabla \cdot (h\mathbf{u})^{n+1/2}] \quad (12)$$

and

$$(h\mathbf{u})^{n+1} = (h\mathbf{u})^n - \Delta t[\nabla \cdot (h\mathbf{u} \circ \mathbf{u})^{n+1/2} + (h\nabla h')^{n+1/2}], \quad (13)$$

which is second-order accurate. To obtain an accurate and stable approximation of the advective flux terms, the momentum  $(h\mathbf{u})^{*,n+1/2}$  computed by the auxiliary system is modified by a height correction  $\delta h'_{\text{fl}}{}^n$  (where the subscript “fl” refers to the fact that this is a correction to the advective flux components):

$$(h\mathbf{u})^{n+1/2} = (h\mathbf{u})^{*,n+1/2} - \frac{\Delta t}{2} h^n \nabla \delta h'_{\text{fl}}{}^n. \quad (14)$$

Applying the divergence to this equation in combination with the height update (12) leads to an (uncritical) Helmholtz problem for  $\delta h'_{\text{fl}}{}^n$ :

$$-\frac{\text{Fr}^2}{\Delta t} \delta h'_{\text{fl}}{}^n + \frac{\Delta t}{2} \nabla \cdot (h^n \nabla \delta h'_{\text{fl}}{}^n) = \frac{H_0^{n+1} - H_0^n}{\Delta t} - \frac{b^{n+1} - b^n}{\Delta t} - \frac{h^{*,n+1} - h^n}{\Delta t}. \quad (15)$$

The last term on the right-hand side is obtained by substituting the divergence of the auxiliary momentum through the height equation of (10). Note that for  $\text{Fr} = 0$  this equation becomes identical to the first correction of a projection method as in [42]. Using (14), the height at the new time level as given in (12) and the advective components of the momentum flux can be computed.

As written above, for the multiscale method we employ two different time discretizations to correct the remaining nonconvective flux component  $(h\nabla h')^n$  in the momentum equation. The first is based on the implicit midpoint rule as given in (13). By the definition of

$$(h\mathbf{u})_{\text{IMP}}^{**} := (h\mathbf{u})^n - \Delta t[\nabla \cdot (h\mathbf{u} \circ \mathbf{u})^{n+1/2} + (h\nabla h')^n], \quad (16)$$

(the subscript “IMP” referring to the implicit midpoint rule) the momentum at the new time level is obtained by

$$(h\mathbf{u})^{n+1} = (h\mathbf{u})_{\text{IMP}}^{**} - \frac{\Delta t}{2} (\delta h^n \nabla h'^n + h^{n+1/2} \nabla \delta h'_{\text{fl}}{}^n), \quad (17)$$

where  $\delta h^n := h^{n+1} - h^n$  and  $h^{n+1/2} := \frac{1}{2}(h^n + h^{n+1})$ . Here  $\delta h'^n := h'^{n+1} - h'^n$  is the update for the perturbation of the fluid depth computed in the second correction.

The second time discretization utilizes the BDF(2) scheme, where the momentum equation is discretized by

$$(h\mathbf{u})^{n+1} = \frac{4}{3}(h\mathbf{u})^n - \frac{1}{3}(h\mathbf{u})^{n-1} - \frac{2\Delta t}{3}[\nabla \cdot (h\mathbf{u} \circ \mathbf{u})^{n+1} + (h\nabla h')^{n+1}]. \quad (18)$$

Note that the advective flux component  $\nabla \cdot (h\mathbf{u} \circ \mathbf{u})$  is only available at the half time level from the predictor and first correction. Since for the BDF discretization this term is needed at the full time level  $t^{n+1}$ , it is linearly extrapolated from older time levels by

$$(h\mathbf{u} \circ \mathbf{u})^{n+1} := (h\mathbf{u} \circ \mathbf{u})^{n+1/2} + \frac{1}{2}((h\mathbf{u} \circ \mathbf{u})^{n+1/2} - (h\mathbf{u} \circ \mathbf{u})^{n-1/2}). \quad (19)$$

A resulting intermediate momentum update is then given by

$$(h\mathbf{u})_{\text{BDF2}}^{**} := \frac{4}{3}(h\mathbf{u})^n - \frac{1}{3}(h\mathbf{u})^{n-1} - \frac{2\Delta t}{3}[\nabla \cdot (h\mathbf{u} \circ \mathbf{u})^{n+1} + (h\nabla h')^n], \quad (20)$$

and the momentum at the new time level is computed by

$$(h\mathbf{u})^{n+1} = (h\mathbf{u})_{\text{BDF2}}^{**} - \frac{2\Delta t}{3}(\delta h^n \nabla h'^n + h^{n+1} \nabla \delta h'^n). \quad (21)$$

**3.2. Second correction.** For the computation of  $\delta h'^n$  the final momentum updates (17) and (21) are combined with a corresponding discretization of the height equation. Using the implicit midpoint rule and further interpolation of the half time level value by the full time level values yields

$$\nabla \cdot \frac{(h\mathbf{u})^{n+1} + (h\mathbf{u})^n}{2} = -\frac{h^{n+1} - h^n}{\Delta t}. \quad (22)$$

By substitution of (17) into this equation, we obtain an (uncritical) Helmholtz problem for the height update  $\delta h'^n$

$$\begin{aligned} -\frac{2\text{Fr}^2}{\Delta t} \delta h'^n + \frac{\Delta t}{2} \nabla \cdot (\hat{h}^{n+1/2} \nabla \delta h'^n) &= 2 \frac{H_0^{n+1} - H_0^n}{\Delta t} - 2 \frac{b^{n+1} - b^n}{\Delta t} + \nabla \cdot (h\mathbf{u})^n \\ &\quad + \nabla \cdot (h\mathbf{u})_{\text{IMP}}^{**} - \frac{\Delta t}{2} \nabla \cdot (\hat{\delta h}^n \nabla h'^n). \end{aligned} \quad (23)$$

Apart from the last term on the right-hand side, for  $\text{Fr} = 0$  this equation is again essentially equivalent to the zero Froude number case. In the case of the zero Froude number projection method, this last term (without the hat over  $\delta h^n$ ) appears in the intermediate momentum update, since there the height update is given through  $H_0(t)$  and  $b(t, \mathbf{x})$ . In the low Froude number case, however, we have  $\delta h^n = \delta H_0^n + \text{Fr}^2 \delta h'^n$ ,



which means that actually the part  $\text{Fr}^2 \Delta t / 2 \nabla \cdot (\delta h'^n \nabla h'^n)$  should be on the left-hand side of the equation, modifying the solution operator. This issue is solved by using the height update known from the first correction (denoted by the hat), i.e.,

$$\widehat{\delta h}^n := (h^{*,n+1} - h^n) + \frac{\Delta t^2}{2} \nabla \cdot (h^n \nabla \delta h'_{\text{fl}}{}^n), \quad (24)$$

to compute this term. The same is true for the weight of the Laplacian in the Helmholtz operator on the left-hand side, where we also apply the height obtained from the first correction. Note that this does not modify the final momentum update (17), where the solution  $\delta h'^n$  of (23) must be used to determine  $\delta h^n$  in order to get conservation of momentum in the absence of nontrivial bottom topography.

To obtain a BDF(2)-type discretization of the second correction, the height equation is discretized by

$$h^{n+1} = \frac{4}{3}h^n - \frac{1}{3}h^{n-1} - \frac{2\Delta t}{3}[\nabla \cdot (hu)^{n+1}]. \quad (25)$$

Similarly to the discretization using the implicit midpoint rule, the momentum update (21) is then combined with (25) to obtain an equation for  $\delta h'^n$ . This leads to the (uncritical) Helmholtz problem

$$\begin{aligned} -\frac{3\text{Fr}^2}{2\Delta t} \delta h'^n + \frac{2\Delta t}{3} \nabla \cdot (\widehat{h}^{n+1} \nabla \delta h'^n) = & -\frac{\text{Fr}^2}{2\Delta t} \delta h'^{n-1} + \frac{3h_0^{n+1} - 4h_0^n + h_0^{n-1}}{2\Delta t} \\ & + \nabla \cdot (hu)_{\text{BDF2}}^{**} - \frac{2\Delta t}{3} \nabla \cdot (\widehat{\delta h}^n \nabla h'^n), \end{aligned} \quad (26)$$

where  $h_0^n = H_0^n - b^n$ . Here again, the values with the hats are approximations obtained from the height computed in the first correction. To conserve momentum in the absence of nontrivial bottom topography, also in this case the result of (26) must be used in the final momentum update (21) for the calculation of  $\delta h^n$  and  $h^{n+1}$ .

In addition to the two schemes described above, we consider the so-called  $\theta$ -scheme. This means that the nonconvective flux term  $(h \nabla h')^{n+\theta}$  in (13) is approximated at  $t^n + \theta \Delta t$ , and (22) is substituted by

$$\theta \nabla \cdot (hu)^{n+1} + (1 - \theta) \nabla \cdot (hu)^n = -\frac{h^{n+1} - h^n}{\Delta t}, \quad \theta \in [0, 1]. \quad (27)$$

For  $\theta = 1$ , this method becomes the implicit Euler method. While it is of second-order accuracy only for  $\theta = 0.5$  (equivalent to the implicit midpoint rule), the scheme usually stabilizes for  $\theta \in (0.5, 1]$ , since more numerical diffusion is introduced.

**3.3. Multiscale scheme.** With the introduction of the implicit midpoint and the BDF(2)-based time discretizations for the second correction, all ingredients are now at hand to apply the multilevel scheme from [43] as part of a semi-implicit

method to the fully nonlinear shallow water equations. The idea is to define direct scale-dependent splittings of the fields for fluid depth and momentum, i.e.,

$$\delta h' = \sum_{\nu=0}^{\nu_M} \delta h'^{(\nu)} \quad \text{and} \quad (hu) = \sum_{\nu=0}^{\nu_M} (hu)^{(\nu)}. \quad (28)$$

Ideally, this could be a quasispectral or wavelet decomposition, splitting the discrete fields into (local) high-wavenumber and low-wavenumber components. Each scale component should be treated depending on how well it is resolved by the underlying implicit time discretization. For each scale  $\nu$  we introduce a blending parameter  $\mu_\nu$ , which depends on the grid CFL number associated to the scale. It is designed such that for well resolved scales the implicit midpoint rule is used, while for scales which are under-resolved in time it blends towards the BDF(2) scheme.

Since we do not want to solve for separate corrections on each scale, we carefully analyze the formal contribution of the two different time discretizations on each scale. With this information and the application of multigrid prolongation and restriction operators, we derive a multilevel elliptic problem, which yields the correction for our semi-implicit discretization.

By the introduction of projection operators  $\Pi_\nu^h$  and  $\Pi_\nu^{(hu)}$ , which project a height or momentum field to the scale  $\nu$ , the contribution for each scale shall be given by

$$\delta h'^{(\nu)} = (\Pi_\nu^h - \Pi_{\nu-1}^h) \delta h' \quad \text{and} \quad (hu)^{(\nu)} = (\Pi_\nu^{(hu)} - \Pi_{\nu-1}^{(hu)})(hu), \quad (29)$$

where we set  $\Pi_{-1}^h \equiv 0$  and  $\Pi_{-1}^{(hu)} \equiv 0$  for simplicity. The scalewise contribution, which results from blending of the schemes, is then defined as follows. With the application of the two schemes for the semi-implicit solution of the shallow water equations, two different intermediate momentum updates are available after the first correction. For the implicit midpoint time discretization this is (16), whereas for the BDF(2)-based discretization the update is given by (20). With these updates, the right-hand sides of the second correction equations (23) and (26) are given by

$$f_{\text{IMP}}^{\delta h'} = -\frac{2}{\Delta t} \left[ 2 \frac{h_0^{n+1} - h_0^n}{\Delta t} + \nabla \cdot (hu)^n + \nabla \cdot (hu)_{\text{IMP}}^{**} - \frac{\Delta t}{2} \nabla \cdot (\widehat{h}^n \nabla h'^n) \right] \quad (30)$$

and

$$f_{\text{BDF2}}^{\delta h'} = \frac{3Fr^2}{4\Delta t^2} \delta h'^{n-1} - \frac{3}{2\Delta t} \left[ \frac{3h_0^{n+1} - 4h_0^n + h_0^{n-1}}{2\Delta t} + \nabla \cdot (hu)_{\text{BDF2}}^{**} - \frac{2\Delta t}{3} \nabla \cdot (\widehat{h}^n \nabla h'^n) \right]. \quad (31)$$

Here, both correction equations have been normalized, such that the weighted Laplacian is essentially the same in the two resulting Helmholtz operators. Note that

this choice is somehow arbitrary and one could have chosen another normalization. For example in [43] we used a normalization where the terms without derivatives in the Helmholtz operators have the common weight 1. Further analysis revealed, however, that this choice can introduce spurious kinks into the solution for the momentum variable. The Helmholtz operators are then given by

$$A_{\text{IMP}} = \frac{4\text{Fr}^2}{\Delta t^2} \text{id} - \nabla \cdot (\hat{h}^{n+1/2} \nabla) \quad \text{and} \quad A_{\text{BDF2}} = \frac{9\text{Fr}^2}{4\Delta t^2} \text{id} - \nabla \cdot (\hat{h}^{n+1/2} \nabla), \quad (32)$$

where the “id” stands for the identity operator. Note that here we also modified the weight in the Laplacian of the operator for the BDF(2) scheme from  $\hat{h}^{n+1}$  to  $\hat{h}^{n+1/2}$ . Using the projections from (29), a scalewise application and summation over the scales results in a multiscale operator, which is given by

$$A := \sum_{\nu=0}^{\nu_M} (\mu_\nu A_{\text{IMP}} + (1 - \mu_\nu) A_{\text{BDF2}}) (\Pi_\nu^h - \Pi_{\nu-1}^h) \quad (33)$$

or, in particular for the operators defined in (32),

$$A := \frac{\text{Fr}^2}{\Delta t^2} \left[ \sum_{\nu=0}^{\nu_M} (4\mu_\nu + \frac{9}{4}(1 - \mu_\nu)) (\Pi_\nu^h - \Pi_{\nu-1}^h) \right] - \nabla \cdot (\hat{h}^{n+1/2} \nabla). \quad (34)$$

With this operator the elliptic equation of the second correction for the solution of  $\delta h'^n$  becomes

$$A \delta h'^n = \sum_{\nu=0}^{\nu_M} (\mu_\nu f_{\text{IMP}}^{\delta h',(\nu)} + (1 - \mu_\nu) f_{\text{BDF2}}^{\delta h',(\nu)}), \quad (35)$$

which also involves a scale-dependent right-hand side. The momentum at the new time level is then computed according to

$$(hu)^{n+1} = \sum_{\nu=0}^{\nu_M} (\mu_\nu (hu)_{\text{IMP}}^{n+1,(\nu)} + (1 - \mu_\nu) (hu)_{\text{BDF2}}^{n+1,(\nu)}), \quad (36)$$

where the scale-dependent contributions are computed by blending the updates that would be obtained by either the implicit midpoint or the BDF(2) time discretization. They are given by projecting

$$(hu)_{\text{IMP}}^{n+1} = (hu)_{\text{IMP}}^{**} - \frac{\Delta t}{2} (\delta h^n \nabla h'^n + h^{n+1/2} \nabla \delta h'^n) \quad (37)$$

and

$$(hu)_{\text{BDF2}}^{n+1} = (hu)_{\text{BDF2}}^{**} - \frac{2\Delta t}{3} (\delta h^n \nabla h'^n + h^{n+1} \nabla \delta h'^n) \quad (38)$$

to each scale using the projections from (29).

It remains to define how the blending weights for each grid level are determined. As described above, we would like to apply the implicit midpoint rule for scale components, which are well resolved by the discretization. For smaller scales the blending should be shifted successively to the BDF(2) scheme. Since the numerical dispersion heavily depends on the CFL number, in an initial attempt the blending parameter is set to be a function of the grid CFL number. For simplicity, the gravity wave speed  $c = \sqrt{h}/Fr$  is estimated by the square root of the mean height divided by the global Froude number in the conducted numerical simulations. This means that the grid CFL number is given by  $\text{cfl}_v = c\Delta t/\Delta x_v$ , where  $\Delta x_v$  is the grid spacing on the respective grid level  $v$ . The blending parameter is then computed according to

$$\mu_v = \begin{cases} \min(1, (v_M - v)/\lfloor \log_2 \text{cfl} \rfloor) & \text{if } \text{cfl} \geq 2, \\ 1 & \text{otherwise,} \end{cases} \quad (39)$$

where  $\lfloor \cdot \rfloor$  means rounding towards minus infinity. Thus,  $\mu_v$  is chosen such that the scheme associates the implicit midpoint discretization with all gravity wave modes corresponding to coarse grids with grid CFL number  $\text{cfl}_v \leq 1$  ( $\mu_v = 1$ ), while the discretization is nudged towards BDF(2) for modes living on grids with  $\text{cfl}_v > 1$  ( $\mu_v < 1$ ). However, if the fine-grid CFL number is smaller than 2, the scheme would consequently end up with using only the implicit midpoint rule. This choice of blending weights has been also used in the linear case [43].

To summarize the time advancement of the multiscale method, we outline the steps of the algorithm in the following box with a reference to the particular equations.

*Semi-implicit multiscale method*

- (1) Explicit predictor solving the auxiliary system (10) over one time step.
- (2) Solution of elliptic problem (15) for  $\delta h_{\text{fl}}^{\prime,n}$  to compute the advective flux correction (14).
- (3) Computation of intermediate momentum updates for implicit midpoint (16) and BDF(2) discretization (20).
- (4) Computation of RHS for second correction via (30)–(31) and their scale-dependent blending using  $\mu_v$ .
- (5) Solution for  $\delta h^{\prime,n}$  by elliptic multiscale problem (35).
- (6) Computation of full momentum updates (37)–(38) and their scale-dependent blending (36) to obtain the momentum at the new time level.

**3.4. Space discretization.** The space discretization for the semi-implicit method is essentially the same as in the zero Froude number projection method. The major differences are that for nonzero Froude numbers two Helmholtz problems must be

solved instead of Poisson-type problems and that some care needs to be taken in order to get conservation of momentum for constant bottom topography.

The scheme is solved in one space dimension with grid cells  $V_i = [x_{i-1/2}, x_{i+1/2}]$ . Furthermore, a dual discretization is introduced, where each dual grid cell  $\bar{V}_{i+1/2} = [x_i, x_{i+1}]$  is centered around a node  $x_{i+1/2}$  of the primary grid. The whole method is discretized as a finite volume method, which has the form

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{|V_i|} (\mathbf{F}_{i+1/2}^{n+1/2} - \mathbf{F}_{i-1/2}^{n+1/2}) + \Delta t \mathbf{N}_i^{n+1/2}. \quad (40)$$

Here,  $|V_i|$  is the volume of cell  $V_i$ .  $\mathbf{U}_i^n$  represents an approximation to the cell mean of the unknowns  $(h, hu)^T$  in the cell  $V_i$  at time  $t^n$ , and  $\mathbf{F}_{i+1/2}^{n+1/2}$  is the advective part of the numerical flux across the interface at  $x_{i+1/2}$ . The latter approximates the average of the advective flux contribution  $(hu, hu^2)^T$  over one time step  $[t^n, t^{n+1}]$ . The additional nonconservative part  $\mathbf{N}_i^{n+1/2}$  accounts for the gradient in surface elevation and is an approximation to  $(0, -hh'_x)^T$ . The equations are discretized to obtain a scheme which is in conservation form for the height equation. Conservation of momentum is only valid when no bottom topography is present. In this case, momentum should also be conserved on the discrete level. Following the above (semidiscrete) derivation of the scheme, the numerical fluxes are computed in three steps

$$\begin{aligned} \mathbf{F}_{i+1/2}^{n+1/2} &:= \mathbf{F}_{i+1/2}^* + \mathbf{F}_{i+1/2}^{\text{MAC}} + 0, \\ \mathbf{N}_i^{n+1/2} &:= \mathbf{N}_i^* + 0 + \mathbf{N}_i^{\text{P2}}, \end{aligned} \quad (41)$$

which represent contributions from the predictor and the first and second corrections, respectively. Note that the first correction only modifies the advective flux components, while the second correction only modifies the nonconservative part. The detailed contributions are given in the Appendix. For the discretization of the bottom topography  $b$ , a piecewise linear distribution on each primary grid cell which is continuous across the interfaces is assumed. The time derivatives  $b_i^{n+1/2}$  are approximated by the midpoint rule using the values at full time levels.

In the predictor step the auxiliary system (10) is solved using a Godunov-type method for hyperbolic conservation laws [40]. As mentioned above, these are the pressureless equations with the “source term”  $(0, -h^n h_x'^n)^T$  in the momentum equation. Note that this term involves not only the contributions from the bottom topography, but also the nonconvective part of the flux function. For the integration, a semidiscretization in space with second-order reconstruction in the primitive variables and Runge–Kutta time stepping is used [33]. In particular, Heun’s method is applied, which is strong stability preserving (SSP) [35; 13]. The numerical fluxes are evaluated by solving the exact Riemann problem of the pressureless equations at the cell interfaces.

In the first correction, the flux divergence of the auxiliary system is corrected, which is similar to a MAC-type projection [14; 44] in the case of the zero Froude number equations. The height correction  $\delta h_{\tilde{h}}'^n$  is continuous and piecewise linear on the dual grid, which is the 1D analogue as it was used in the solution of an elliptic problem in [36], or in the first correction of the method in [42] in two space dimensions. The fluid depth  $h^n$  in the weighted Laplacian of (15) is interpolated at the nodes of the primary grid by taking the average from the two neighboring cells (cf. [21]).

For the second correction, the divergence on the right-hand side of (23) is applied to each dual control volume. This leads to a 1D divergence defined by

$$\bar{D}_{i+1/2}(u) := \frac{1}{|\bar{V}_{i+1/2}|} (u_{i+1} - u_i). \quad (42)$$

Also the computed correction  $\delta h'^n$  is assumed to be continuous and piecewise linear, but this time on the primary cells. Moreover, it needs to be defined how the fluid depth which enters as weight in the Laplacian on the left-hand side of (23) is discretized. Here we assume that the fluid depth is piecewise constant on each cell. This leads to a piecewise constant distribution of  $h(\delta h'^n)_x$ , and the weighted Laplacian resulting from the divergence (42) is well defined.

Concerning conservation of momentum in the case of flat bottom topography, it must be ensured that the term

$$hh'_x = h_0 h'_x + Fr^2 h' h'_x \quad (43)$$

in the momentum equation can be written as a divergence on the discrete level. Since  $h_0$  is constant in this case, this is no problem for the first term on the right-hand side of (43). For the second term, the equality

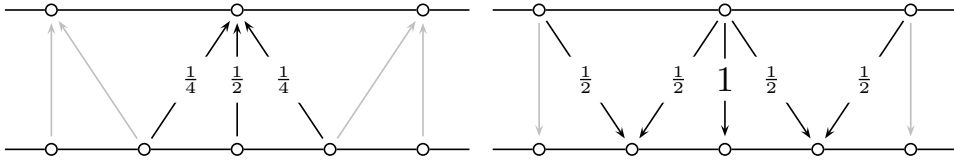
$$h' h'_x = \frac{1}{2} ((h')^2)_x \quad (44)$$

has to be achieved on the discrete level. We realize this by taking

$$(h' h'_x)_i = \left( \frac{h'_{i+1/2} + h'_{i-1/2}}{2} \right) \left( \frac{h'_{i+1/2} - h'_{i-1/2}}{\Delta x} \right) = \frac{(h'_{i+1/2})^2 - (h'_{i-1/2})^2}{2\Delta x}, \quad (45)$$

where the interface values are linearly interpolated from cell mean values.

The spatial discretization of the scale splitting in the second correction of the multiscale scheme is obtained by eliminating every second grid node or, equivalently, by merging two adjacent cells. In this setup the restriction and prolongation operators used in standard multigrid algorithms can be utilized to define the space decomposition. Here we use full weighting (restriction) and linear interpolation (prolongation) [37] for the fluid depth, which can be defined by a stencil. The full



**Figure 1.** One-dimensional versions of full weighting (left) and linear interpolation (right) operators known from standard finite difference geometric multigrid. Arrows indicate mappings between grid functions associated with grid nodes.

weighting is given by

$$R^{(\nu)} = \frac{1}{4} [1 \ 2 \ 1], \tag{46}$$

which means that a variable on the coarse grid node at grid level  $(\nu)$  is derived by averaging over the values at the same node and the two adjacent nodes on the fine grid at grid level  $(\nu + 1)$  with the weights given in the stencil above (see also Figure 1, left). The linear interpolation from grid level  $(\nu)$  to grid level  $(\nu + 1)$  is given by

$$P^{(\nu)} = \frac{1}{2} [1 \ 2 \ 1]. \tag{47}$$

This means that the heights at grid nodes living on the fine grid level, which have a common coarse grid node, obtain the same values as on the coarse grid. The values at grid nodes in between are computed by the average of the values of the adjacent grid nodes (Figure 1, right). Note that  $P^{(\nu)}$  and  $R^{(\nu)}$  are adjoint up to a scaling factor.

Since  $\delta h'$  and  $(hu)$  are staggered in space, the splitting in the momentum field cannot be the same as the one for the height update. Ideally, the splitting should be chosen such that only the portion of the height update associated with the grid level  $(\nu)$  enters the update for the momentum on the same grid level. Revisiting equations (37) and (38) shows that only first derivatives of  $\delta h'$  at different time levels enter the momentum update. Therefore, the splitting in the momentum must match the splitting in  $\partial \delta h' / \partial x$  induced by the  $h$ -splitting [43]. This results in a restriction with stencil

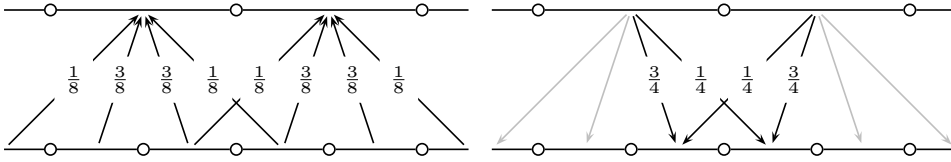
$$\widehat{R}^{(\nu)} = \frac{1}{8} [1 \ 3 \ 3 \ 1] \tag{48}$$

for the momentum (Figure 2, left). The obvious choice for the prolongation operator is a scaled version of the adjoint of the restriction operator  $\widehat{R}^{(\nu)}$ , which results in

$$\widehat{P}^{(\nu)} = \frac{1}{4} [1 \ 3 \ 3 \ 1], \tag{49}$$

which is visualized in Figure 2, right.

A grid function  $\varphi$  can then be decomposed into fractions  $\varphi^{(\nu)}$  associated to different grid levels using the prolongation and restriction operators  $P^{(\nu)}$  and  $R^{(\nu)}$ .



**Figure 2.** One-dimensional versions of restriction (left) and prolongation (right) operators for the momentum variable. Arrows indicate mappings between grid functions associated with grid cells (instead of with grid nodes as in Figure 1).

The grid function on the coarsest level is obtained by the operation

$$\varphi^{(0)} = (R^{(0)} \circ R^{(1)} \circ \dots \circ R^{(v_M-1)})\varphi, \tag{50}$$

and the grid functions on finer levels are computed by

$$\varphi^{(v)} = (I - P^{(v-1)} \circ R^{(v-1)}) \circ (R^{(v)} \circ R^{(v+1)} \circ \dots \circ R^{(v_M-1)})\varphi. \tag{51}$$

An application of the multiscale Helmholtz operator is then realized by decomposing the data into scales, scale-dependent weighting, and rebuilding the full variable. This gives the diagonal component of the operator, which includes the multiscale information. The Laplacian part can just be computed on the finest grid level, since it does not include any multiscale information.

### 4. Numerical results

Having derived the multiscale scheme for computing low Froude number shallow water flows, in this section the performance of the method is evaluated for some test cases. Besides the goal of numerically verifying the second-order accuracy of the method, its asymptotic behavior in the low Froude number regime as described in Section 2.2 is investigated.

The results of the multiscale method are compared to those obtained with the semi-implicit method using the implicit midpoint rule and the BDF(2) discretization in the second correction. With the exception of the last test case, the computations for the BDF(2) and the multiscale schemes are always started with an initial first step by the implicit midpoint rule. With this, enough old time step values can be provided for the BDF(2)-based scheme. As mentioned above, the blending parameter  $\mu_v$  in the multilevel scheme is computed according to (39). However, the precise values are always given for reference in each test case.

Since the presented scheme is semi-implicit, two Courant numbers [6] are considered. The Courant number concerning the maximum propagation speed of information is essentially associated with the propagation of gravity waves in the low Froude number case and denoted by  $\text{cfl}_{\text{grav}}$ . Furthermore, the Courant number concerning advective phenomena (which are mainly computed by the



explicit predictor) is given by  $\text{cfl}_{\text{adv}} := \max_i (|u_i|) \Delta t / \Delta x$ , where  $u_i$  is the velocity computed for each cell.

The linear systems for the solution of  $\delta h_{\text{fl}}^{\prime, n}$  and  $\delta h^{\prime, n}$  are solved using a matrix-free implementation of the Bi-CGSTAB algorithm [39]. In each iteration, the Euclidean norm of the residual vector is calculated, and the algorithm is terminated when either the absolute value or the value relative to the norm of the initial residuum is less than a given tolerance. In the presented calculations, this tolerance is set to  $10^{-10}$ .

**4.1. Weakly nonlinear gravity wave.** The first test case is set up with data, which consists of an initially smooth right-running shallow water simple wave in one space dimension with flat bottom topography. Due to the nonlinearity of the governing equations, a shock develops after some time. While this is one of the most simple setups one can think of, it already reveals some interesting properties of the considered numerical schemes: by the use of the method of characteristics, the exact solution is known until the development of a shock, which is useful for a convergence study. The behavior of the different schemes towards the compressible regime can also be tested, when the exact solution eventually develops a shock. Furthermore, the evolution of long-wave gravity waves can be analyzed, which is relevant for the asymptotic regime described in Section 2.2 and similar to what was investigated for the linearized equations (cf. [43; 41]).

To derive the initial conditions, let us consider the characteristic variables of the shallow water equations. These are given by (see, e.g., [11])

$$p_1 = u - 2c \quad \text{and} \quad p_2 = u + 2c, \quad (52)$$

where  $c = \sqrt{h}/\text{Fr}$  is the gravity wave speed. The definition of a background state  $h_0 = 1$  leads to  $c_0 = 1/\text{Fr}$ . Then, the initial gravity wave speed is given by

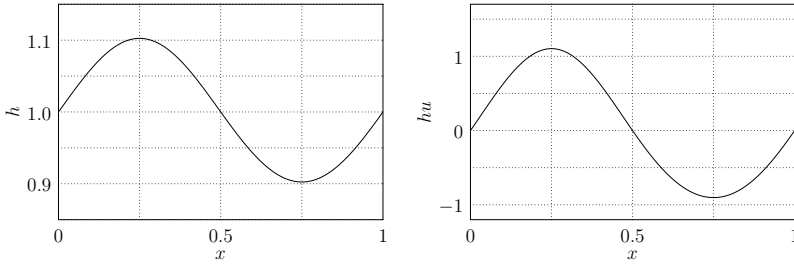
$$c = c_0 + c' = \frac{1}{\text{Fr}} + c'. \quad (53)$$

To obtain a right-running simple wave, the left-running characteristic is set to  $p_1 = \text{const}$ . This constant is chosen to obtain a zero background flow, i.e.,  $p_1 = -2c_0$ , which gives the initial velocity field

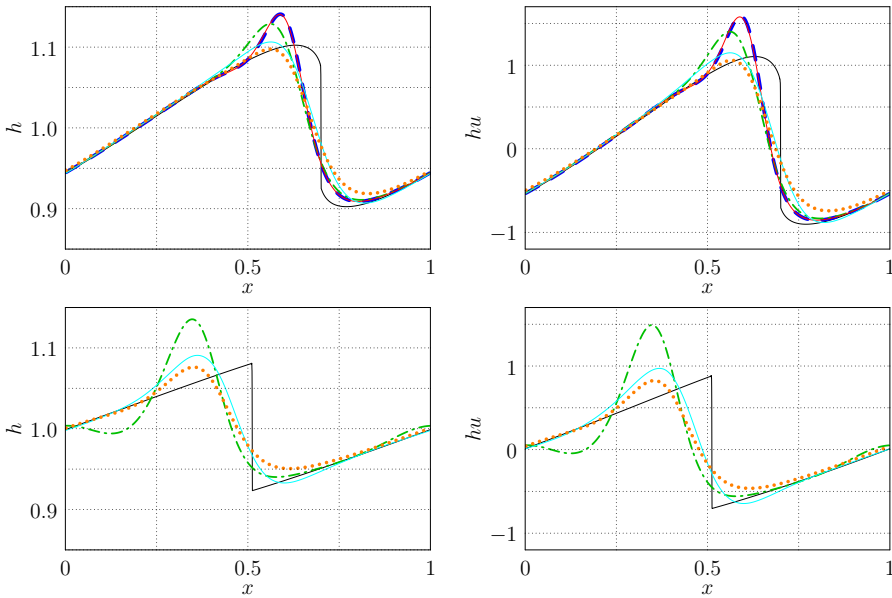
$$u = 2(c - c_0) = 2c'. \quad (54)$$

Therefore, initially the local Froude number ranges from 0 to  $\text{Fr}_{\text{max}} \approx u_{\text{max}}/c_0 = 2\text{Fr} \max_{x \in \Omega} (c'(x))$ . For the performed simulations the perturbation of the gravity wave speed is set to  $c'(x) = \frac{1}{2} \sin(2\pi x)$ . The computational domain is defined by the interval  $\Omega = [0, 1]$  with 256 grid cells and periodic boundary conditions.

In a first setup, the Froude number is set to  $\text{Fr} = 0.1$  and the time step is chosen to be  $\Delta t = 0.003$ , which is equivalent to initial Courant numbers  $\text{cfl}_{\text{adv}} \approx 0.77$



**Figure 3.** Initial conditions for the weakly nonlinear gravity wave test case with  $Fr = 0.1$  on a grid with 256 grid cells. Left: fluid depth. Right: momentum.



**Figure 4.** Solution of the weakly nonlinear gravity wave test case with  $Fr = 0.1$  at times  $t = 0.12$  (top) and  $t = 0.3$  (bottom) computed with  $\text{cfl}_{\text{grav}} \approx 8.83$  on a grid with 256 grid cells. Black: exact solution. Blue dashed: implicit midpoint rule. Green dash-dotted: BDF(2)-type discretization. Orange dotted: off-centered scheme ( $\theta = 0.7$ ). Red: multiscale implicit midpoint/BDF(2) scheme. Cyan: multiscale implicit midpoint/implicit Euler scheme. Note that the implicit midpoint rule and the multiscale implicit midpoint/BDF(2) schemes are only shown for  $t = 0.12$ .

concerning advection and  $\text{cfl}_{\text{grav}} \approx 8.83$  concerning the propagation of gravity waves. In Figure 3 the initial conditions for fluid depth and momentum are given. The solutions of the numerical schemes are given after 40 ( $t = 0.12$ ) and 100 time steps ( $t = 0.3$ ) in Figure 4. At these times the wave has traveled approximately 1.2 and 3 times, respectively, through the domain. Since a shock forms at time  $t_{\text{shock}} = 1/(3\pi)$ , this test shows the performance of the schemes towards the compressible

regime. The multiscale scheme is set up with three grid levels and blending factors  $\mu_v = (1, \frac{1}{2}, 0)$ .

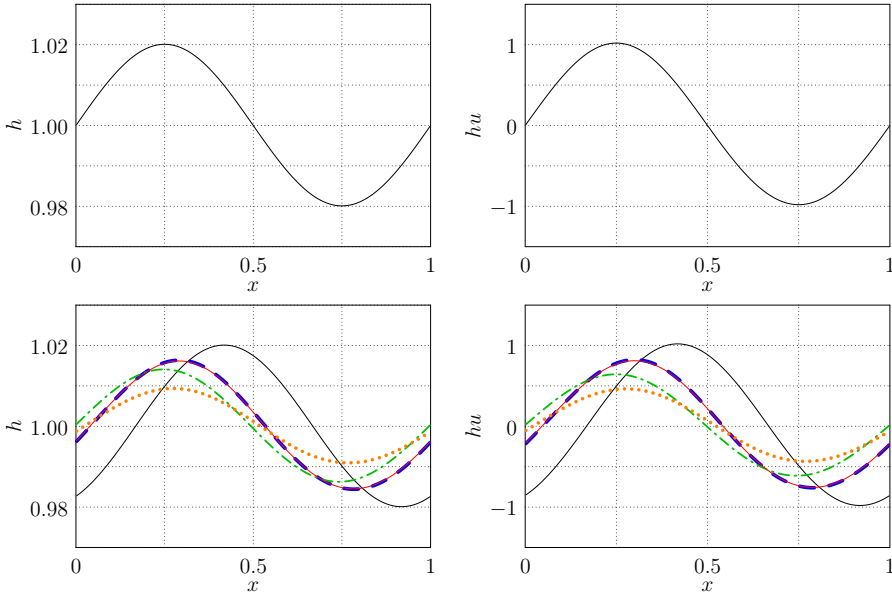
In addition we present results for the  $\theta$ -scheme with  $\theta = 0.7$  and another variant of the multiscale scheme where we switch between the implicit midpoint rule and the implicit Euler method ( $\theta$ -scheme with  $\theta = 1$ ). For the latter we choose six grid levels with blending factors  $\mu_v = (1, \frac{1}{2}, 0, 0, 0, 0)$ . Note that this choice is different from what one would obtain using (39).

As one can see in Figure 4, for  $t = 0.12$  the implicit midpoint rule and multiscale scheme develop an artificial overshoot in the vicinity of the shock, which continuously grows until either the time step has to be reduced or the schemes become unstable (which already happens before the time  $t = 0.3$ ). Since the initial data only consists of long-wave information, and the contributions on the smaller scales are only small corrections, the results for both schemes are almost identical. On the other hand, the  $\theta$ -scheme does not show this behavior, and the discontinuity is smeared out by numerical diffusion. The BDF(2)-based scheme shows a behavior which is in between these two extrema. To show that the multiscale scheme can also be used to suppress the spurious overshoot, we have implemented the version of the multiscale scheme where we switch between the implicit midpoint rule and the implicit Euler. In this case, high wavenumbers are diffused by the first-order method, while the long-wave components are preserved.

The described behavior becomes even more evident at the later time  $t = 0.3$ , where we only show the BDF(2), the  $\theta$ -scheme, and the multiscale implicit midpoint/implicit Euler scheme, due to the stability problem of the implicit midpoint rule. Additionally, all schemes introduce a dispersive error in that they slow down the speed of the simple wave.

To test the evolution of long-wave gravity waves, the Froude number is reduced to  $Fr = 0.02$  in a second setup. This further decreases the nonlinearity of the equations compared to the case with  $Fr = 0.1$ . However, due to the configuration of the initial data, the shock develops at the same time  $t_{\text{shock}} = 1/(3\pi)$  as before. The initial conditions for this test case are shown in Figure 5, top. The time step is again  $\Delta t = 0.003$ , which is equivalent to initial Courant numbers  $\text{cfl}_{\text{adv}} \approx 0.77$  and  $\text{cfl}_{\text{grav}} \approx 39.55$ . The solution at time  $t = 0.024$  is displayed in Figure 5, bottom. At this time, the gravity wave has traveled approximately 1.2 times through the domain, and its shape has not yet been distorted much compared to the initial data. For this test, the multiscale scheme is applied with six levels and blending parameters  $\mu_v = (1, \frac{4}{5}, \frac{3}{5}, \frac{2}{5}, \frac{1}{5}, 0)$ .

At the final time the implicit midpoint rule and the multiscale scheme show the smallest error in amplitude and phase compared to the exact solution. Also in this case the solutions of these schemes are nearly identical. The worst results are produced by the off-centered scheme, which has the biggest phase and amplitude



**Figure 5.** Weakly nonlinear gravity wave test case with  $Fr = 0.02$  computed with  $cf_{\text{grav}} \approx 39.55$  on a grid with 256 grid cells. Top: initial conditions. Bottom: solution at  $t = 0.024$ . Black: exact solution. Blue dashed: implicit midpoint rule. Green dash-dotted: BDF(2)-type discretization. Orange dotted: off-centered scheme ( $\theta = 0.7$ ). Red: multiscale scheme.

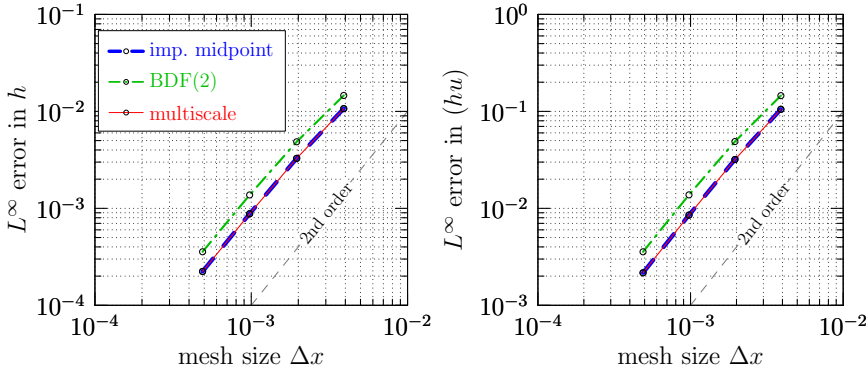
errors. The method with BDF(2) in the second correction produces results which are in between these two extrema.

**4.2. Convergence in one space dimension.** The same initial conditions of a right-running gravity simple wave and for  $Fr = 0.1$  are used in order to undertake a convergence analysis. The solution of the numerical schemes is computed on different grids and compared to the exact solution at time  $t_{\text{end}} = 0.05$ . At this time no shock has developed yet, and the true solution can be computed using the method of characteristics. The numerical solution is computed on grids with 256, 512, 1024, and 2048 cells, and the respective time steps are given by  $\Delta t_{256} = \frac{1}{320}$ ,  $\Delta t_{512} = \frac{1}{640}$ ,  $\Delta t_{1024} = \frac{1}{1280}$ , and  $\Delta t_{2048} = \frac{1}{2560}$ . This corresponds to an advective Courant number  $cf_{\text{adv}} = 0.8$ . For the multiscale method, five grid levels with  $\mu_v = (1, 1, \frac{2}{3}, \frac{1}{3}, 0)$  are used.

For the computation of errors and the convergence rate, the error vectors  $e^N$  in fluid depth and momentum are calculated. For the latter it has elements

$$e_i^N := (hu)_i(t_N) - (hu)_i^N \quad (55)$$

where the cell mean values of the exact solution are compared with those of the simulated data. The global error is measured using discrete versions of the  $L^2$  and



**Figure 6.** Convergence for the one-dimensional simple wave test case.  $L^\infty$  errors in  $h$  and  $(hu)$  for the different variants of the semi-implicit method.

the  $L^\infty$  norms. These are defined by

$$\|\mathbf{e}^N\|_{[2]} := \left( \sum_i |V_i| |e_i^N|^2 \right)^{1/2} \quad \text{and} \quad \|\mathbf{e}^N\|_{[\infty]} := \max_i \{e_i^N\}. \quad (56)$$

The experimental convergence rate  $\gamma$  is calculated by the formula

$$\gamma := \frac{\log(\|\mathbf{e}_c^N\| / \|\mathbf{e}_f^N\|)}{\log(\Delta x_c / \Delta x_f)}. \quad (57)$$

In this definition,  $\mathbf{e}_c^N$  and  $\mathbf{e}_f^N$  are the computed error vectors of the solution on a coarse and a fine grid and  $\Delta x_c$  and  $\Delta x_f$  are the corresponding grid spacings.

The error of the numerical solutions in the  $L^\infty$  norm is summarized in Figure 6. Furthermore, the precise values in the  $L^2$  and  $L^\infty$  norms are given in the Appendix in Tables 1 and 2, where also the convergence rates  $\gamma$  between the grid levels are calculated. On fixed grids, the scheme with implicit midpoint discretization in the second correction produces the smallest errors. The method with a BDF(2)-based second correction produces errors which are about 1.5 times larger. The multiscale scheme produces errors which are comparable with those from the implicit midpoint rule. This is again due to the long-wave nature of the initial conditions. As given by the values of  $\mu_\nu$ , only the finest scales of the BDF(2)-based method are applied, which means that the calculations are nearly identical up to small deviations. The experimental convergence rates suggest for all schemes second-order accuracy.

**4.3. Balanced modes in presence of time-dependent bottom topography.** In a final test case, the schemes are tested for their ability to relax to nontrivial balanced states in the presence of bottom topography varying in time. In order to do so, a test case from [43] (see also [41]) for the linearized equations is extended to the fully nonlinear shallow water equations. The test is defined in one space dimension

on the domain  $\Omega = [0, 100]$ . The bottom topography is given by

$$b(t, x) = \frac{\text{Fr}}{\omega} \sin(\omega t) \tilde{q}(x - x_0), \quad (58)$$

where

$$\tilde{q}(x) = \left[ \frac{2\sigma^2 + \lambda^2\sigma^4 - 4x^2}{\lambda^2\sigma^4} \sin(\lambda x) + \frac{4x}{\lambda\sigma^2} \cos(\lambda x) \right] \exp\left(-\left(\frac{x}{\sigma}\right)^2\right). \quad (59)$$

This means that the term  $b_t(t, x) = \text{Fr} \cos(\omega t) \tilde{q}(x - x_0)$  must be balanced by the production of local divergence. The parameters are given by  $\omega = 0.2\pi$ ,  $x_0 = 50$ ,  $\sigma = 10$ , and  $\lambda = 0.32\pi$ . Initially the fluid is at rest ( $u \equiv 0$ ) with fluid depth  $h \equiv 1$ . When the flow is in balance, the findings from Section 2 imply that for small Froude numbers the perturbations in fluid depth and momentum should also be small, and the dynamics primarily happen in the linear regime. This means that the solution is essentially governed by the asymptotic solution obtained for the linearized shallow water equations. Translated to the given initial value problem and bottom topography, the asymptotic solutions of the perturbation in fluid depth and the velocity are

$$H_{\text{asy}}(t, x) - H_0 = -\frac{\text{Fr}^3}{H_0} \omega \sin(\omega t) \tilde{h}(x - x_0) \quad (60)$$

with  $\tilde{h}(x) = \lambda^{-2} \sin(\lambda x) \exp(-(x/\sigma)^2)$ , and

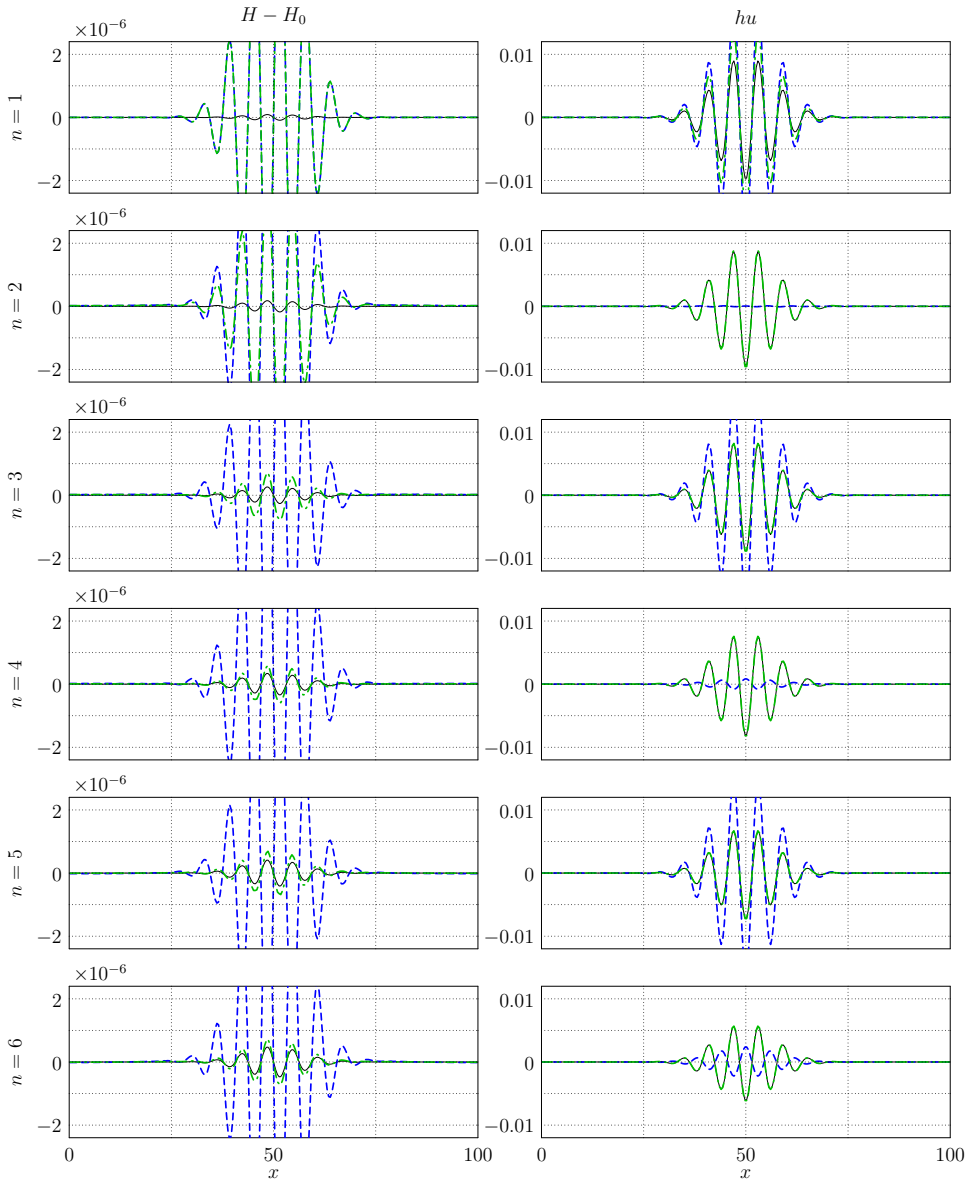
$$u_{\text{asy}}(t, x) = \frac{\text{Fr}}{H_0} \cos(\omega t) \tilde{u}(x - x_0), \quad (61)$$

where  $\tilde{u}(x) = [2x(\sigma\lambda)^{-2} \sin(\lambda x) - \lambda^{-1} \cos(\lambda x)] \exp(-(x/\sigma)^2)$ .

In the presented computations, the Froude number is set to  $\text{Fr} = 0.01$  and the total background height is  $H_0 = 1$ . The computational grid has 256 grid cells, and the fixed time step is given by  $\Delta t = 0.24$ , which corresponds to an advective Courant number  $\text{cfl}_{\text{adv}} \approx 0.006$  when the flow is essentially balanced. The Courant number corresponding to the transport of gravity waves is  $\text{cfl}_{\text{grav}} \approx 61$ .

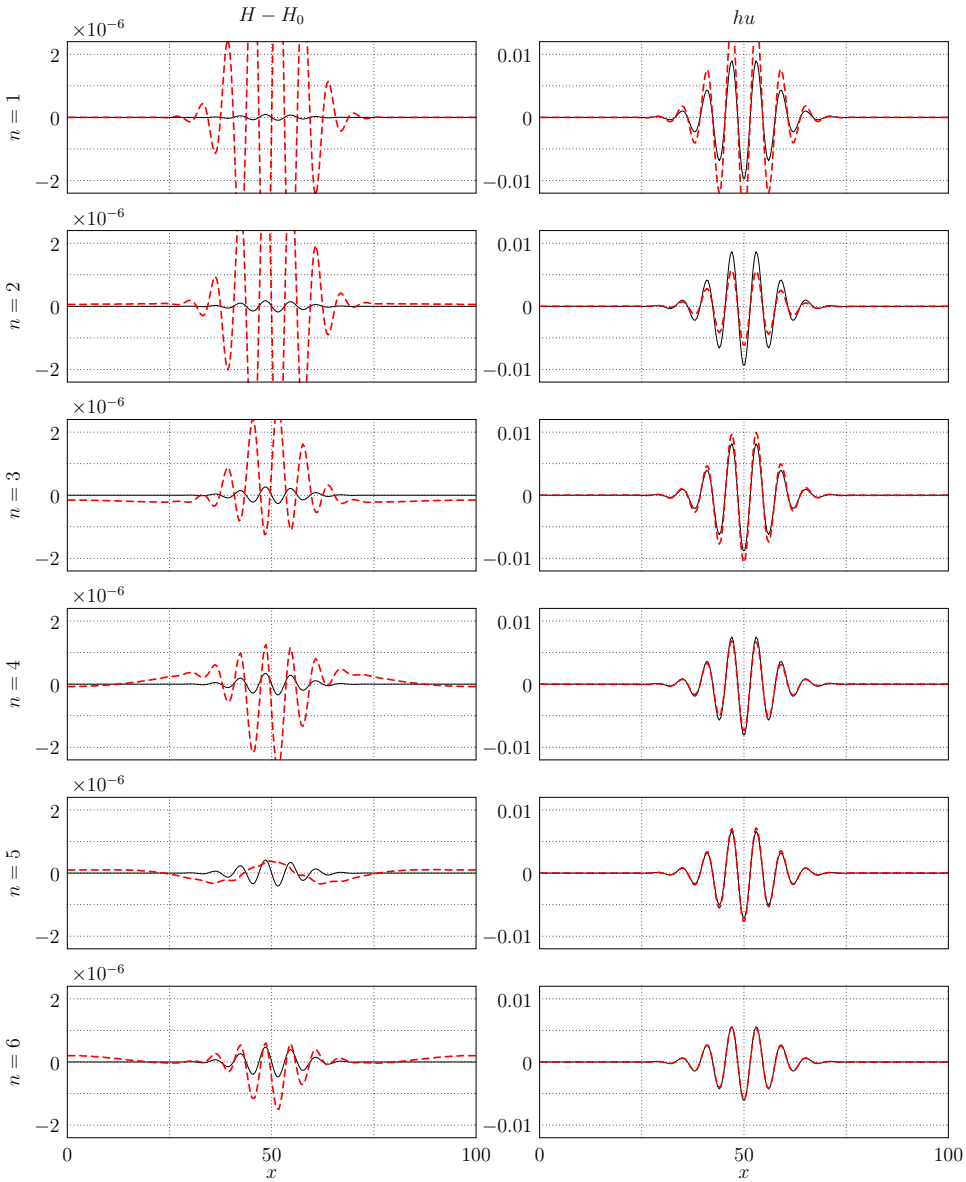
For this test case the BDF(2)-based computations are not initialized with an initial step by the implicit midpoint rule. Instead, the required state at  $t^{-1} = -0.24$  is set to the balanced solution with flat bottom topography. However, compared to an initialization using the implicit midpoint rule, the findings are qualitatively the same. For the multiscale method six grid levels are used with a scale-dependent blending given by  $\mu_v = (1, \frac{4}{5}, \frac{3}{5}, \frac{2}{5}, \frac{1}{5}, 0)$ .

Given the above initial conditions for  $t = 0$ , the fluid depth is in balance with the initial bottom topography. However, the temporal change of the latter introduces divergence into the velocity field, which, in turn, results in higher-order perturbations in the fluid depth. In Figure 7, the numerical results are displayed together with



**Figure 7.** Numerical solution of the balancing test case after the first six time steps using the implicit midpoint rule (blue dashed) and the BDF(2) scheme (green dash-dotted) on a grid with 256 cells and  $Fr = 0.01$ . Left column: perturbation in fluid depth. Right column: momentum. Each step  $n$  is one row. The asymptotic solution is plotted as a black line.

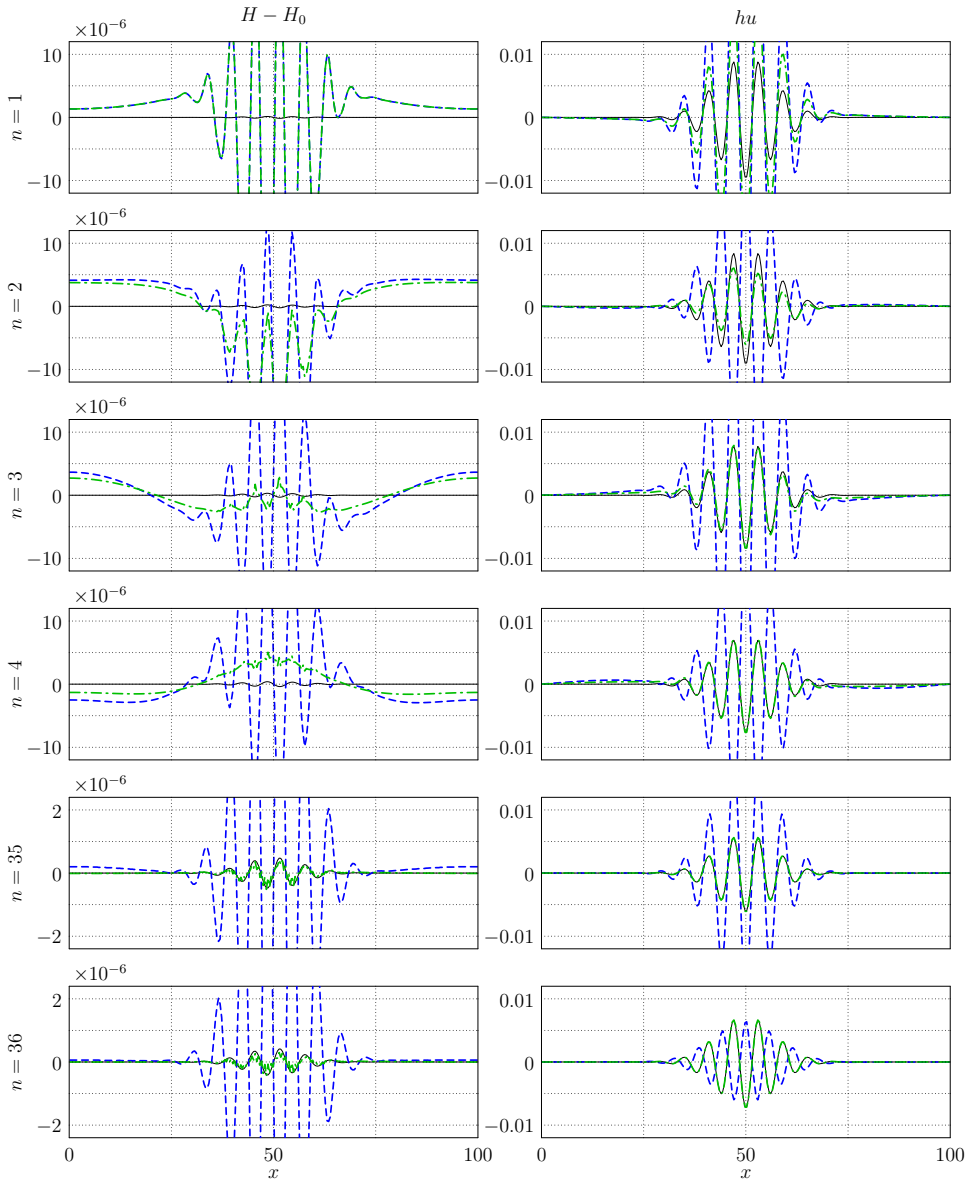
the asymptotic solution for the first six time steps using the implicit midpoint rule and BDF(2)-based discretization. Using the implicit midpoint rule, both the computed perturbations in the fluid depth and the momentum field oscillate around



**Figure 8.** Same as Figure 7, but using the multiscale scheme (red dashed).

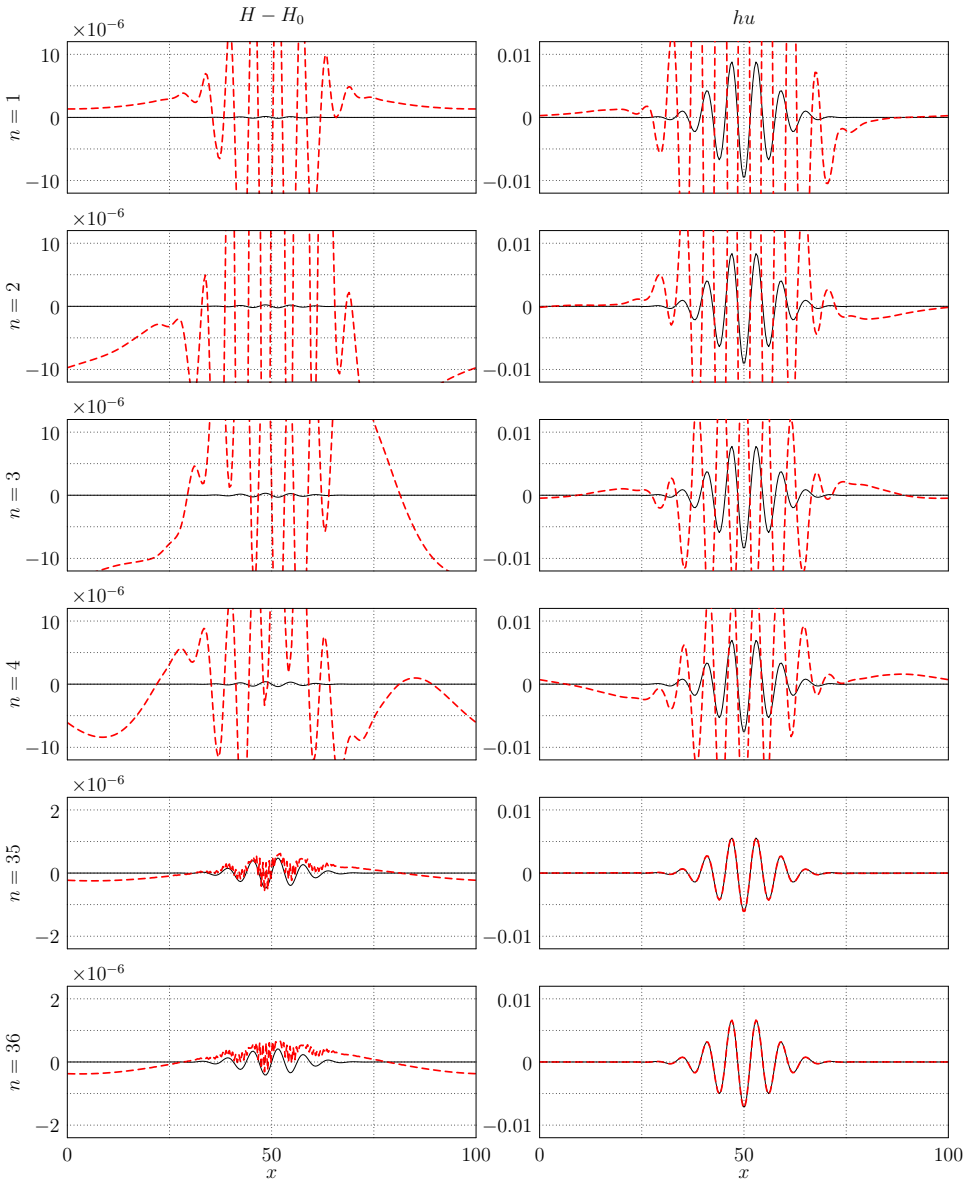
the balanced state, but they do not relax to it. Additionally, the amplitude of the numerically calculated perturbations in the fluid depth are about 8 times larger compared to the asymptotic solution. The BDF(2)-based discretization results in a completely different behavior. Here, the initial deviations from the balanced state vanish after only a few time steps. After the fourth time step the numerical solution





**Figure 9.** Numerical solution of the balancing test case using completely unbalanced initial data after the first 4 time steps and time steps 35 and 36 using the implicit midpoint rule (blue dashed) and the BDF(2) discretization (green dash-dotted) on a grid with 256 cells and  $Fr = 0.01$ . The asymptotic solution is plotted as a black line.

is nearly indistinguishable from the asymptotic solution. This behavior is also reproduced by the multiscale method, for which the results are given in Figure 8.



**Figure 10.** Same as Figure 9, but using the multiscale scheme (red dashed).

These results are in good agreement with the findings for the linearized shallow water equations [43; 41].

In a second run, the simulation is started at  $t = 0.15$ , and the bottom topography is assumed to be flat before this time. At this time, when the bottom topography switches instantaneously to another state, both fluid depth and momentum are not in

balance. This leads to much bigger initial deviations from the asymptotic solution, as can be seen in Figure 9 for the implicit midpoint rule and the BDF(2)-type discretization (note the different scaling in the  $y$ -axis for the perturbation in fluid depth for the first four time steps). To evaluate the long-term behavior, the numerical solution is additionally plotted for the time steps 35 and 36. Also in this case the solution of the implicit midpoint rule does not relax to the balanced state, but rather oscillates around it. Only the long-wave perturbations are diminished with time. Here, the perturbations in fluid depth computed by the numerical scheme are about two orders of magnitude larger than those predicted by the asymptotic solution. For the momentum, the amplitude of the numerical solution is also about three times larger than the predicted balanced state.

The BDF(2)-based method, on the other hand, shows a behavior similar to the first setup. After initial deviations, which are of the same order as for the implicit midpoint rule, the numerical solutions essentially relax to the balanced state predicted by the asymptotic solution. Only in the fluid depth, very high-wave-number small-amplitude deviations persist. Additional tests (not shown) suggest that these artifacts are due to the fact that the explicit predictor cannot cope with too high-wave-number modes at these large Courant numbers. In this part of the scheme, a two-stage Runge–Kutta method is used for the time discretization. Since the gravity waves are generated by the “source term” of the predictor, which is always evaluated at the old time level, high-wave-number gravity waves get very much distorted in the second stage of the Runge–Kutta scheme. This can eventually lead to instabilities, if these parts of the solution become too large.

The results of the multiscale method are given in Figure 10. Qualitatively, the behavior is similar to the BDF(2)-based second correction. However, the scale-dependent blending of the two methods leads to even larger very high-wave-number deviations, but whose amplitude is of the order of the perturbations in fluid depth. Also some long-wave perturbations persist, which cannot propagate away due to the periodic boundary conditions.

## 5. Conclusion

In this work, a new multiscale semi-implicit method for the numerical solution of low Froude number shallow water flows is introduced. It is motivated by significant shortcomings of classical semi-implicit large time step integration schemes applied in current atmospheric codes. A principal feature of the new method is the diverse treatment of long- and short-wave solution components in accordance with the asymptotic regime of fast gravity waves traveling over short-range topography. This is achieved through a multilevel approach borrowing ideas from multigrid schemes

method	norm	256	rate $\gamma$	512	rate $\gamma$	1024	rate $\gamma$	2048
trapezoidal rule	$L^2$	$3.2801 \times 10^{-3}$	1.846	$9.1251 \times 10^{-4}$	1.955	$2.3530 \times 10^{-4}$	1.991	$5.9190 \times 10^{-5}$
	$L^\infty$	$1.0686 \times 10^{-2}$	1.705	$3.2770 \times 10^{-3}$	1.898	$8.7942 \times 10^{-4}$	1.977	$2.2342 \times 10^{-4}$
BDF(2)	$L^2$	$4.7937 \times 10^{-3}$	1.763	$1.4127 \times 10^{-3}$	1.912	$3.7548 \times 10^{-4}$	1.975	$9.5495 \times 10^{-5}$
	$L^\infty$	$1.4599 \times 10^{-2}$	1.587	$4.8593 \times 10^{-3}$	1.822	$1.3743 \times 10^{-3}$	1.947	$3.5642 \times 10^{-4}$
multiscale method	$L^2$	$3.2793 \times 10^{-3}$	1.846	$9.1193 \times 10^{-4}$	1.956	$2.3512 \times 10^{-4}$	1.991	$5.9157 \times 10^{-5}$
	$L^\infty$	$1.0661 \times 10^{-2}$	1.703	$3.2748 \times 10^{-3}$	1.898	$8.7882 \times 10^{-4}$	1.977	$2.2328 \times 10^{-4}$

**Table 1.** Errors and convergence rates in  $h$  for the different variants of the semi-implicit method.

for elliptic equations. The scheme is second-order accurate and admits time steps depending essentially on the flow velocity.

The multiscale scheme is able to properly propagate long-wave gravity waves, and their dispersion and amplitude errors are minimized as much as the considered base schemes admit. However, some artifacts can be observed in the fluid depth, which are probably related to the explicit predictor of the semi-implicit method. But these should be acceptable in practical applications. In the presence of bottom topography, which varies slowly in time, the balanced state is attained after a reasonable number of time steps.

The ultimate goal of this work is to develop a multiscale multiply blended scheme that does not only account for the scale-dependent propagation properties of the various wave modes in the atmosphere, thereby creating the numerical analogue of the blended model formulation of [23].

The source code for the method and tests are available upon request from the authors.

## Appendix

**A.1. Numerical fluxes of the finite volume scheme.** As outlined in (41), the numerical fluxes are computed in three steps. Here, the particular terms using the trapezoidal rule in the second correction are given. The case using the BDF(2) discretization uses the same spatial operators, but has some differences in the particular terms.  $F_I^*$  and  $N_i^*$  are the numerical fluxes approximating the flux function and “source term” of the auxiliary system, respectively. These are

$$F_I^* = \begin{pmatrix} (hu)^{n+1/2} \\ (hu)^{n+1/2} u^{n+1/2} \end{pmatrix} \quad \text{and} \quad N_i^* = \begin{pmatrix} 0 \\ -(hh'_x)^n \end{pmatrix}. \quad (62)$$

The second flux term

$$F_I^{\text{MAC}} := -\frac{\Delta t}{2} \begin{pmatrix} h^n (\delta h'_{\text{fl}})^n_x \\ (hu)^{*,n+1/2} (\delta h'_{\text{fl}})^n_x + h^n (\delta h'_{\text{fl}})^n_x u^{*,n+1/2} \end{pmatrix}_I \quad (63)$$

method	norm	256	rate $\gamma$	512	rate $\gamma$	1024	rate $\gamma$	2048
trapezoidal rule	$L^2$	$3.2422 \times 10^{-2}$	1.864	$8.9047 \times 10^{-3}$	1.961	$2.2875 \times 10^{-3}$	1.991	$5.7556 \times 10^{-4}$
	$L^\infty$	$1.0527 \times 10^{-1}$	1.722	$3.1899 \times 10^{-2}$	1.904	$8.5226 \times 10^{-3}$	1.977	$2.1654 \times 10^{-3}$
BDF(2)	$L^2$	$4.7676 \times 10^{-2}$	1.740	$1.4277 \times 10^{-2}$	1.910	$3.8002 \times 10^{-3}$	1.976	$9.6614 \times 10^{-4}$
	$L^\infty$	$1.4534 \times 10^{-1}$	1.573	$4.8843 \times 10^{-2}$	1.826	$1.3778 \times 10^{-2}$	1.952	$3.5620 \times 10^{-3}$
multiscale method	$L^2$	$3.2404 \times 10^{-2}$	1.865	$8.8982 \times 10^{-3}$	1.961	$2.2855 \times 10^{-3}$	1.990	$5.7521 \times 10^{-4}$
	$L^\infty$	$1.0494 \times 10^{-1}$	1.720	$3.1864 \times 10^{-2}$	1.904	$8.5157 \times 10^{-3}$	1.976	$2.1639 \times 10^{-3}$

**Table 2.** Errors and convergence rates in  $(hu)$  for the different variants of the semi-implicit method.

corresponds to the first correction computed by (15). As stated above, with this correction the new time level fluid depth can be determined. The third contribution in (41) is given by

$$N_i^{P2} := \begin{pmatrix} 0 \\ -\frac{1}{2}(\delta h^n h'_x{}^n + h^{n+1/2} \delta h'_x{}^n) \end{pmatrix}_i \tag{64}$$

and represents the correction computed by the second Helmholtz equation (23).

**A.2. “Simple wave” test case.** The computed errors and convergence rates (cf. Section 4.2) are given in Tables 1 and 2.

### Acknowledgements

The authors acknowledge partial support by Deutsche Forschungsgemeinschaft through the Collaborative Research Center 1114 “Scaling cascades in complex systems”, Project A02, and by Einstein Stiftung Berlin through the Einstein Visiting Fellowship “Multiscale atmospheric flows: rigorous analysis, ensemble downscaling, and data assimilation” (Professor Edriss Titi). This work benefited greatly from free software products. Without these tools — such as L<sup>A</sup>T<sub>E</sub>X and the Linux operating system — a lot of tasks would not have been so easy to realize. It is our pleasure to thank all developers for their excellent products.

### References

[1] P. R. Bannon, *On the anelastic approximation for a compressible atmosphere*, J. Atmos. Sci. **53** (1996), no. 23, 3618–3628.

[2] F. Bouchut, *On zero pressure gas dynamics*, Advances in kinetic theory and computing (B. Perthame, ed.), Ser. Adv. Math. Appl. Sci., no. 22, World Sci. Publ., River Edge, NJ, 1994, pp. 171–190. MR Zbl

[3] F. Bouchut, S. Jin, and X. Li, *Numerical approximations of pressureless and isothermal gas dynamics*, SIAM J. Numer. Anal. **41** (2003), no. 1, 135–158. MR Zbl

[4] D. Bresch, R. Klein, and C. Lucas, *Multiscale analyses for the shallow water equations*, Computational science and high performance computing IV (E. Krause, Y. Shokin, M. Resch, D.

- Kröner, and N. Shokina, eds.), Notes Numer. Fluid Mech. Multidiscip. Des., no. 115, Springer, 2011, pp. 149–164. MR
- [5] F. Cordier, P. Degond, and A. Kumbaro, *An asymptotic-preserving all-speed scheme for the Euler and Navier–Stokes equations*, J. Comput. Phys. **231** (2012), no. 17, 5685–5704. MR Zbl
- [6] R. Courant, K. Friedrichs, and H. Lewy, *Über die partiellen Differenzgleichungen der mathematischen Physik*, Math. Ann. **100** (1928), no. 1, 32–74. MR Zbl
- [7] T. Davies, A. Staniforth, N. Wood, and J. Thuburn, *Validity of anelastic and other equation sets as inferred from normal-mode analysis*, Q. J. Roy. Meteor. Soc. **129** (2003), no. 593, 2761–2775.
- [8] P. Deuffhard and F. Bornemann, *Scientific computing with ordinary differential equations*, Texts in Applied Mathematics, no. 42, Springer, 2002. MR Zbl
- [9] D. R. Durran, *Improving the anelastic approximation*, J. Atmos. Sci. **46** (1989), no. 11, 1453–1461.
- [10] ———, *Numerical methods for fluid dynamics: with applications to geophysics*, 2nd ed., Texts in Applied Mathematics, no. 32, Springer, 2010. MR Zbl
- [11] G. Erbes, *A semi-Lagrangian method of characteristics for the shallow-water equations*, Mon. Weather Rev. **121** (1993), no. 12, 3443–3452.
- [12] K. J. Geratz, *Erweiterung eines Godunov-Typ-Verfahrens für zwei-dimensionale kompressible Strömungen auf die Fälle kleiner und verschwindender Machzahl*, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 1997.
- [13] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev. **43** (2001), no. 1, 89–112. MR Zbl
- [14] F. H. Harlow and J. E. Welch, *Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface*, Phys. Fluids **8** (1965), no. 12, 2182–2189. MR Zbl
- [15] E. Kalnay, *Atmospheric modeling, data assimilation and predictability*, Cambridge University, 2002.
- [16] J. Kevorkian and J. D. Cole, *Multiple scale and singular perturbation methods*, Applied Mathematical Sciences, no. 114, Springer, 1996. MR Zbl
- [17] S. Klainerman and A. Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math. **34** (1981), no. 4, 481–524. MR Zbl
- [18] R. Klein, *Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics, I: One-dimensional flow*, J. Comput. Phys. **121** (1995), no. 2, 213–237. MR Zbl
- [19] ———, *Asymptotic analyses for atmospheric flows and the construction of asymptotically adaptive numerical methods*, ZAMM Z. Angew. Math. Mech. **80** (2000), no. 11-12, 765–777. MR Zbl
- [20] ———, *An applied mathematical view of meteorological modelling*, Applied mathematics entering the 21st century (J. M. Hill and R. Moore, eds.), SIAM, 2004, pp. 227–269. MR Zbl
- [21] ———, *Asymptotics, structure, and integration of sound-proof atmospheric flow equations*, Theor. Comp. Fluid Dyn. **23** (2009), no. 3, 161–195. Zbl
- [22] ———, *Scale-dependent models for atmospheric flows*, Annu. Rev. Fluid Mech. **42** (2010), 249–274. MR Zbl
- [23] R. Klein and T. Benacchio, *A doubly blended model for multiscale atmospheric dynamics*, J. Atmos. Sci. **73** (2016), no. 3, 1179–1186.

- [24] R. Klein, N. Botta, T. Schneider, C. D. Munz, S. Roller, A. Meister, L. Hoffmann, and T. Sonar, *Asymptotic adaptive methods for multi-scale problems in fluid mechanics*, J. Engrg. Math. **39** (2001), no. 1-4, 261–343. MR Zbl
- [25] R. Klein, S. Vater, E. Paeschke, and D. Ruprecht, *Multiple scales methods in meteorology*, Asymptotic methods in fluid mechanics: survey and recent advances (H. Steinrück, ed.), CISM Courses and Lectures, no. 523, Springer, 2010, pp. 127–196. Zbl
- [26] O. Le Maître, J. Levin, M. Iskandarani, and O. M. Knio, *A multiscale pressure splitting of the shallow-water equations, I: Formulation and 1D tests*, J. Comput. Phys. **166** (2001), no. 1, 116–151. MR
- [27] R. J. Leveque, *The dynamics of pressureless dust clouds and delta waves*, J. Hyperbolic Differ. Equ. **1** (2004), no. 2, 315–327. MR Zbl
- [28] F. B. Lipps and R. S. Hemler, *A scale analysis of deep moist convection and some related numerical calculations*, J. Atmos. Sci. **39** (1982), no. 10, 2192–2210.
- [29] A. J. Majda and R. Klein, *Systematic multiscale models for the tropics*, J. Atmos. Sci. **60** (2003), no. 2, 393–408.
- [30] C.-D. Munz, S. Roller, R. Klein, and K. J. Geratz, *The extension of incompressible flow solvers to the weakly compressible regime*, Comput. & Fluids **32** (2003), no. 2, 173–196. MR Zbl
- [31] Y. Ogura and N. A. Phillips, *Scale analysis of deep and shallow convection in the atmosphere*, J. Atmos. Sci. **19** (1962), no. 2, 173–179.
- [32] W. Ohfuchi, H. Nakamura, M. K. Yoshioka, T. Enomoto, K. Takaya, X. Peng, S. Yamane, T. Nishimura, Y. Kurihara, and K. Ninomiya, *10-km mesh meso-scale resolving simulations of the global atmosphere on the Earth Simulator: preliminary outcomes of AFES (AGCM for the Earth Simulator)*, J. Earth Simulator **1** (2004), 8–34.
- [33] S. Osher, *Convergence of generalized MUSCL schemes*, SIAM J. Numer. Anal. **22** (1985), no. 5, 947–961. MR Zbl
- [34] J. Pedlosky, *Geophysical fluid dynamics*, 2nd ed., Springer, 1987. Zbl
- [35] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys. **77** (1988), no. 2, 439–471. MR Zbl
- [36] E. Süli, *Convergence of finite volume schemes for Poisson’s equation on nonuniform meshes*, SIAM J. Numer. Anal. **28** (1991), no. 5, 1419–1430. MR Zbl
- [37] U. Trottenberg, C. W. Oosterlee, and A. Schüller, *Multigrid*, Academic, 2001. MR Zbl
- [38] G. K. Vallis, *Atmospheric and oceanic fluid dynamics: fundamentals and large-scale circulation*, Cambridge, 2006. Zbl
- [39] H. A. van der Vorst, *Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput. **13** (1992), no. 2, 631–644. MR Zbl
- [40] B. van Leer, *Towards the ultimate conservative difference scheme, V: A second-order sequel to Godunov’s method*, J. Comput. Phys. **32** (1979), no. 1, 101–136. Zbl
- [41] S. Vater, *A multigrid-based multiscale numerical scheme for shallow water flows at low Froude number*, Ph.D. thesis, Freie Universität Berlin, 2013.
- [42] S. Vater and R. Klein, *Stability of a Cartesian grid projection method for zero Froude number shallow water flows*, Numer. Math. **113** (2009), no. 1, 123–161. MR Zbl
- [43] S. Vater, R. Klein, and O. M. Knio, *A scale-selective multilevel method for long-wave linear acoustics*, Acta Geophys. **59** (2011), no. 6, 1076–1108. Zbl

- [44] J. E. Welch, F. H. Harlow, J. P. Shannon, and B. J. Daly, *The MAC method: a computing technique for solving viscous, incompressible, transient fluid-flow problems involving free surfaces*, Tech. Report LA-3425, Los Alamos Scientific Laboratory, 1965.

Received December 1, 2017. Revised June 27, 2018.

STEFAN VATER: [stefan.vater@math.fu-berlin.de](mailto:stefan.vater@math.fu-berlin.de)

*Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany*

RUPERT KLEIN: [rupert.klein@math.fu-berlin.de](mailto:rupert.klein@math.fu-berlin.de)

*Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany*



## Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at [msp.org/camcos](http://msp.org/camcos).

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language.** Articles in CAMCoS are usually in English, but articles written in other languages are welcome.

**Required items.** A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format.** Authors are encouraged to use  $\LaTeX$  but submissions in other varieties of  $\TeX$ , and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of Bib $\TeX$  is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to [graphics@msp.org](mailto:graphics@msp.org) with details about how your graphics were generated.

**White space.** Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

**Proofs.** Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

# *Communications in Applied Mathematics and Computational Science*

vol. 13

no. 2

2018

---

- A numerical study of the extended Kohn–Sham ground states of atoms 139  
ERIC CANCÈS and NAHIA MOURAD
- An equation-by-equation method for solving the multidimensional moment  
constrained maximum entropy problem 189  
WENRUI HAO and JOHN HARLIM
- Symmetrized importance samplers for stochastic differential equations 215  
ANDREW LEACH, KEVIN K. LIN and MATTHIAS MORZFELD
- Efficient high-order discontinuous Galerkin computations of low Mach  
number flows 243  
JONAS ZEIFANG, KLAUS KAISER, ANDREA BECK, JOCHEN SCHÜTZ  
and CLAUS-DIETER MUNZ
- A numerical study of the relativistic Burgers and Euler equations on a  
Schwarzschild black hole exterior 271  
PHILIPPE G. LEFLOCH and SHUYANG XIANG
- A semi-implicit multiscale scheme for shallow water flows at low Froude  
number 303  
STEFAN VATER and RUPERT KLEIN



1559-3940(2018)13:2;1-B