SIAM J. SCI. COMPUT.
Vol. 42, No. 3, pp. B816–B845

HIGH ORDER ASYMPTOTIC PRESERVING DEFERRED CORRECTION IMPLICIT-EXPLICIT SCHEMES FOR KINETIC MODELS*

RÉMI ABGRALL† AND DAVIDE TORLO‡

Abstract. This work introduces an extension of the residual distribution (RD) framework to stiff relaxation problems. The RD is a class of schemes which is used to solve a hyperbolic system of partial differential equations. To our knowledge, it has been used only for systems with mild source terms, such as gravitation problems or shallow water equations. What we propose is an implicit-explicit (IMEX) version of the RD schemes that can resolve stiff source terms, without refining the discretization up to the stiffness scale. This can be particularly useful in various models, where the stiffness is given by topological or physical quantities, e.g., multiphase flows, kinetic models, or viscoelasticity problems. We will focus on kinetic models that are BGK approximation of hyperbolic conservation laws. The extension to more complicated problems will be carried out in future works. The provided scheme is able to catch different relaxation scales automatically, without losing accuracy; we prove that the scheme is asymptotic preserving and this guarantees that, in the relaxation limit, we recast the expected macroscopic behavior. To get a high order accuracy, we use an IMEX time discretization combined with a deferred correction procedure, while naturally RD provides high order space discretization. Finally, we show some numerical tests in one and two dimensions for stiff systems of equations.

Key words. residual distribution, IMEX, relaxation, deferred correction, asymptotic preserving, kinetic model

AMS subject classifications. 65M12, 65L04, 65M60

DOI. 10.1137/19M128973X

1. Introduction.

In many models, such as kinetic models, multiphase flows, viscoelasticity, or relaxing gas flows, we have to deal with hyperbolic systems with relaxation terms. The relaxation term is often led by a parameter ε, the relaxation parameter, that can represent the mean free path, the average distance between two collisions of particles, the time needed to reach the equilibrium between two phases, etc. Expanding these equations asymptotically with respect to ε, one can find the limit equations that describe the average, effective, or macroscopic physical behavior [9, 20, 23].

In particular, we focus on the kinetic model proposed by Aregba-Driollet and Natalini in [9, 10]. This model is able to solve any hyperbolic system of equation, through a BGK relaxation, which leads to a linear advection system with a relaxation source term. It can be used to test classical hyperbolic systems in the relaxation limit case. This model must be subjected to a generalization of Whitham's subcharacteristic condition [9, 20], which ensures the stability of the model. We use this model to approximate the transport linear equation and the Euler equation in one and two

*Submitted to the journal's Computational Methods in Science and Engineering section September 26, 2019; accepted for publication (in revised form) May 11, 2020; published electronically June 29, 2020.

https://doi.org/10.1137/19M128973X

Funding: This work was supported by the ITN ModCompShock project funded by the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie grant agreement 642768.

†Institut für Mathematik, Winterthurstrasse 190, CH 8057 Zürich, Switzerland (remi.abgrall@math.uzh.ch).

‡Corresponding author. Institut für Mathematik, Winterthurstrasse 190, CH 8057 Zürich, Switzerland (davide.torlo@math.uzh.ch).

dimensions. There are various other models and physical problems which behave similarly to this kinetic model. In the future, the perspective is to extend the method to other problems, such as the multiphase flows Baer–Nunziato model or viscoelasticity problems.

We use the residual distribution (RD) framework [3, 6, 15, 24] to discretize our space. This class of schemes is a generalization of finite element methods (FEM); they use compact stencils, they do not need Riemann solvers, and they are easily generalizable. Indeed, many well-known FEM, finite volume, and discontinuous Galerkin schemes can be rewritten into the RD distribution framework as shown in [5]. The main steps of the scheme are three: we have to compute total residuals for each cell; then, we have to distribute each residual to degrees of freedom belonging to the cell; and finally, we sum all contributions for each degree of freedom. In order to get a high order scheme, the RD is coupled with a deferred correction (DeC) iterative method to have high order time integrator [4, 16, 21]. It needs two operators: the first is a low order method, but easy to invert, while the second must be higher order, but we do not need to solve it directly. The coupling of these two operators allows us to reach the high order through a few iterative intermediate steps. Thanks to this, we can produce a scheme which is fast, high order, and stable. To our knowledge, RD was used only for hyperbolic equations with mild source terms, such as in gravitation problems or shallow water equations, but never on strongly stiff source terms.

To deal with the stiffness of the relaxation term, we have to introduce some special treatments. An explicit scheme with CFL conditions tuned on the macroscopic regime would, indeed, present instabilities, because of the stiff relaxation term. It is natural to choose an implicit or semi-implicit formulation, which guarantees the stability of the scheme. We use an implicit-explicit (IMEX) scheme to treat implicitly the relaxation term and explicitly the advection part [20, 23]. Nevertheless, we can obtain a computationally explicit scheme, thanks to some properties of the considered model. Then, we introduce an IMEX discretization for the DeC RD schemes with the details of its implementation. Furthermore, we prove that the new DeC RD IMEX scheme is asymptotic preserving (AP). AP schemes allow us to preserve the asymptotic behavior of the model from the microscopic regime to the macroscopic one. These schemes solve the microscopic equations, avoiding the coupling of different models, and automatically are able to solve the asymptotic macroscopic limit in a robust way. In the appendix, we also provide a proof of the accuracy of the total scheme.

We show the performance of the high order scheme on some tests. In particular, we simulated different examples in one and two dimensions for the linear transport equation and the Euler equation. Thanks to these results, we validate the accuracy of our method and the capability of shock limiting along discontinuities.

The outline of the manuscript is as follows. In section 2 we present the kinetic model we want to solve, the conditions under which it is stable, and some examples. In section 3 we introduce a first order IMEX scheme that preserves the AP property of the analytical model. In section 4 we describe the RD schemes for the spatial discretization with the DeC high order time discretization. In section 5, we adjust the time discretization of the DeC according to the IMEX scheme and we prove the AP property of the whole scheme. One can find more details about the RD scheme in Appendix A and the proof of high order accuracy in Appendix B. We show numerical results for one-dimensional (1D) and 2D problems in section 6.

2. Kinetic relaxation model for hyperbolic systems. In this section, we present the kinetic model that will be the object of this work. This family of kinetic models was introduced by Aregba-Driollet and Natalini in [9, 10]. Starting from a

hyperbolic system of conservation laws, the *macroscopic model*, they build an artificial kinetic model, the relaxed *microscopic model* we will actually solve. The scheme we propose in this work solves this artificial model, where no physical meaning is involved in the kinetic model, but only in the *macroscopic* limit. The aim is to test the properties and the quality of the scheme before applying it to more involved problems. In the future, we aim to develop the method for the Baer–Nunziato multiphase equations model, Boltzmann equations, and lattice Boltzmann models.

Let us introduce the two models we will consider. Let $\Omega \subset \mathbb{R}^D$ be a bounded smooth spatial domain and let $u : \Omega \times \mathbb{R}^+ \to \mathbb{R}^S$ be a weak solution of the macroscopic model that is defined by the following hyperbolic system of S conservation laws:

$$\partial_t u(x,t) + \sum_{d=1}^{D} \partial_{x_d} A_d(u(x,t)) = 0 \quad \forall x \in \Omega,\ \forall t \in \mathbb{R}^+. \tag{2.1}$$

Here, t defines the time, x_d defines the variable in the different dimensions, and ∂ represents the partial derivative in a specified variable. $A_d : \mathbb{R}^S \to \mathbb{R}^S$, for $d = 1, \dots, D$, are some Lipschitz continuous functions and $u_0 : \Omega \to \mathbb{R}^S$ are the initial conditions and B an operator representing the boundary conditions. The kinetic model proposed in [9] is a relaxed version of this system. Let $f^\varepsilon : \Omega \times \mathbb{R}^+ \to \mathbb{R}^L$ be the solution of the following microscopic kinetic model, where $L > S$ is to be defined:

$$f^\varepsilon(x,t)_t + \sum_{d=1}^{D} \Lambda_d \partial_{x_d} f^\varepsilon(x,t) = \frac{1}{\varepsilon}\left(\mathcal{M}(P f^\varepsilon(x,t)) - f^\varepsilon(x,t)\right) \quad \forall x \in \Omega,\ \forall t \in \mathbb{R}^+, \tag{2.2}$$

where $\Lambda_d \in R^{L \times L}$ are constant diagonal matrices and the source term is the difference between the microscopic variable f^ε and the equilibrium state given by the Maxwellian $\mathcal{M} : \mathbb{R}^S \to \mathbb{R}^L$, which embeds a macroscopic variable into the microscopic space, and $P \in \mathbb{R}^{L \times S}$ is a projection matrix that compresses information from the microscopic variables to the macroscopic ones. The relaxation parameter $\varepsilon \in \mathbb{R}^+$ can be a physical parameter or an artificial one, and, as $\varepsilon \to 0$, the kinetic model (2.2) tends formally to the macroscopic one (2.1). Again, f_0 are initial conditions and boundary conditions must be imposed. All the operators and the domain and the codomain spaces are summarized in Figure 2.1.

There are two fundamental hypotheses on the operators $\mathcal{M}$, P and the functions A_d and Λ_d, which allow us to prove the convergence of the kinetic model to the macroscopic one:

$$P(\mathcal{M}(u)) = u \quad \forall u \in \mathbb{R}^S, \tag{2.3}$$

$$P\Lambda_d \mathcal{M}(u) = A_d(u) \quad \forall u \in \mathbb{R}^S. \tag{2.4}$$

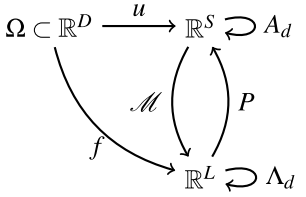


FIG. 2.1. *Relaxation functions.*

The first property (2.3) tells us that the projection P of the Maxwellian $\mathcal{M}$ is the identity matrix $I \in \mathbb{R}^{S\times S}$ or, in other words, that if we take a macroscopic variable u, we embed it in the microscopic space, and then we project it back, we obtain the original state. The second property (2.4) is necessary to guarantee that the limit of the kinetic model will preserve the original macroscopic fluxes.

What we will consider in this work is one specific model, the so-called diagonal relaxation method (DRM) [9]. In this model we choose $L := (D+1)\cdot S$, $P := (I,\dots,I) \in \mathbb{R}^{S\times L}$ as the juxtaposition of $D+1$ identity matrices $I \in R^{S\times S}$. We introduce a constant parameter $\lambda > 0$ to define the flux matrices

$$\Lambda_d := \operatorname{diag}\left(C_1^{(d)},\dots,C_{D+1}^{(d)}\right) \quad \forall d = 1,\dots,D, \quad C_j^{(d)} := \begin{cases} -\lambda I_S, & j = d, \\ \lambda I_S, & j = D+1, \\ 0 & \text{else.} \end{cases} \tag{2.5}$$

The Maxwellian functions are defined in blocks of dimension S each, $\mathcal{M}_j : \mathbb{R}^S \to \mathbb{R}^S$ with $j = 1,\dots,D+1$, so that the original Maxwellian function can be reinterpreted as $\mathcal{M} = (\mathcal{M}_1,\dots,\mathcal{M}_{D+1})^T : \mathbb{R}^S \to \mathbb{R}^L$, as follows:

$$\begin{cases} \mathcal{M}_{D+1}(u) := \left(u + \dfrac{1}{\lambda}\displaystyle\sum_{d=1}^{D} A_d(u)\right)/(D+1), \\ \mathcal{M}_j(u) := -\dfrac{1}{\lambda}A_j(u) + \mathcal{M}_{D+1}(u) \quad \text{for } j = 1,\dots,D. \end{cases} \tag{2.6}$$

These definitions verify the hypotheses (2.3) and (2.4).

Example 2.1 (Jin–Xin relaxation system). If we consider a 1D scalar example, $D = 1$, $S = 1$, as the macroscopic equation

$$\partial_t u + \partial_x a(u) = 0, \tag{2.7}$$

the DRM for the relaxed model leads for the variable $f := (f_1, f_2)^T$ to the kinetic model

$$\partial_t \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} + \begin{pmatrix} -\lambda & 0 \\ 0 & \lambda \end{pmatrix} \partial_x \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \frac{1}{2\varepsilon} \begin{pmatrix} -f_1 + f_2 - a(f_1+f_2)/\lambda \\ f_1 - f_2 + a(f_1+f_2)/\lambda \end{pmatrix}. \tag{2.8}$$

If we apply a change of variables to the previous system and we define $u^\varepsilon := Pf^\varepsilon = f_1^\varepsilon + f_2^\varepsilon$ and $v^\varepsilon := \lambda(f_2^\varepsilon - f_1^\varepsilon)$, we can rewrite the previous system as

$$\begin{cases} \partial_t u^\varepsilon + \partial_x v^\varepsilon = 0, \\ \partial_t v^\varepsilon + \lambda^2 \partial_x u^\varepsilon = \frac{a(u^\varepsilon) - v^\varepsilon}{\varepsilon}, \end{cases} \tag{2.9}$$

also known as the Jin–Xin relaxation system proposed in [20]. In this small case, one can easily perform a Chapman–Enskog expansion and see that

$$\partial_t u^\varepsilon + \partial_x a(u^\varepsilon) = \varepsilon\left(\lambda^2 - (a'(u^\varepsilon))^2\right)\partial_{xx} u^\varepsilon + \mathcal{O}(\varepsilon^2). \tag{2.10}$$

We observe that the macroscopic model appears as the 0th term of the Chapman–Enskog expansion, while the first term is a diffusion operator if the Whitham's subcharacteristic condition is fulfilled, i.e., $\lambda^2 \geq (a'(u^\varepsilon))^2$.

Example 2.2 (Euler system 1D). Suppose we have the system of equations

$$\partial_t \left(\rho, \rho u, E\right) + \partial_x \left(\rho u, \rho u^2 + p, u(E+p)\right) = 0, \tag{2.11}$$

where ρ is the density, u the speed, p the pressure, and E the total energy and they are linked by the closure equation of state (EOS) $p = (\gamma - 1)(E - 0.5\rho u^2)$. Then, we denote the different components of the microscopic variable as $f^\varepsilon = (\rho_1, \rho_1 u_1, E_1, \rho_2, \rho_2 u_2, E_2)^T$. The kinetic model reads

$$\partial_t \begin{pmatrix} \rho_1 \\ \rho_1 u_1 \\ E_1 \\ \rho_2 \\ \rho_2 u_2 \\ E_2 \end{pmatrix} + \partial_x \begin{pmatrix} -\lambda\rho_1 \\ -\lambda\rho_1 u_1 \\ -\lambda E_1 \\ \lambda\rho_2 \\ \lambda\rho_2 u_2 \\ \lambda E_2 \end{pmatrix} = \frac{1}{2\varepsilon} \begin{pmatrix} -\frac{\rho_1 u_1 + \rho_2 u_2}{\lambda} - \rho_1 + \rho_2 \\ -\frac{\rho_1 u_1^2 + p_1 + \rho_2 u_2^2 + p_2}{\lambda} - \rho_1 u_1 + \rho_2 u_2 \\ -\frac{u_1(E_1+p_1) + u_2(E_2+p_2)}{\lambda} - E_1 + E_2 \\ \frac{\rho_1 u_1 + \rho_2 u_2}{\lambda} + \rho_1 - \rho_2 \\ \frac{\rho_1 u_1^2 + p_1 + \rho_2 u_2^2 + p_2}{\lambda} + \rho_1 u_1 - \rho_2 u_2 \\ \frac{u_1(E_1+p_1) + u_2(E_2+p_2)}{\lambda} + E_1 - E_2 \end{pmatrix}. \tag{2.12}$$

Example 2.3 (scalar 2D). Let us consider a scalar equation in two dimensions:

$$\partial_t u + \partial_x a(u) + \partial_y b(u) = 0. \tag{2.13}$$

The microscopic unknown will be denoted by $f^\varepsilon = (f_1, f_2, f_3)^T$ and let us define $u^\varepsilon := Pf = f_1 + f_2 + f_3$. Thus, the model will be

$$\partial_t \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} + \partial_x \begin{pmatrix} -\lambda f_1 \\ 0 \\ \lambda f_3 \end{pmatrix} + \partial_y \begin{pmatrix} 0 \\ -\lambda f_2 \\ \lambda f_3 \end{pmatrix} = \frac{1}{3\varepsilon} \begin{pmatrix} (-2f_1 + f_2 + f_3) + \frac{-2a(u^\varepsilon) + b(u^\varepsilon)}{\lambda} \\ (f_1 - 2f_2 + f_3) + \frac{a(u^\varepsilon) - 2b(u^\varepsilon)}{\lambda} \\ (f_1 + f_2 - 2f_3) + \frac{a(u^\varepsilon) + b(u^\varepsilon)}{\lambda} \end{pmatrix}. \tag{2.14}$$

2.1. Chapman–Enskog expansion. Inspired by the Jin–Xin example, Example 2.1, we develop the Chapman–Enskog for the general kinetic system (2.2), with the only additional properties (2.3) and (2.4), as proposed in [9].

PROPOSITION 2.4. *Assume that f^ε, a solution of* (2.2)*, converges to f, in some strong topology, as $\varepsilon \to 0$. And suppose, furthermore, that the initial conditions f_0^ε are such that $Pf_0^\varepsilon \to u_0$. Then the projection of the solution of the kinetic model* (2.2) *converges to the macroscopic solution u of the system* (2.1)*, i.e., $Pf^\varepsilon \to u$.*

Proof. Define the auxiliary variables as in the Jin–Xin example, Example 2.1:

$$u^\varepsilon := Pf^\varepsilon, \quad v_d^\varepsilon := P\Lambda_d f^\varepsilon \quad \forall\, d = 1, \dots, D. \tag{2.15}$$

Then we have from (2.2) that

$$\begin{cases} \partial_t u^\varepsilon + \displaystyle\sum_{j=1}^{D} \partial_{x_j} v_j^\varepsilon = 0, \\ \partial_t v_d^\varepsilon + \displaystyle\sum_{j=1}^{D} \partial_{x_j} (P\Lambda_j \Lambda_d f^\varepsilon) = \frac{1}{\varepsilon}(A_d(u^\varepsilon) - v_d^\varepsilon) \quad \forall\, d \in \{1, \dots, D\}. \end{cases} \tag{2.16}$$

Applying the Chapman–Enskog expansion, we get that

$$\partial_t u^\varepsilon + \sum_{d=1}^{D} \partial_{x_d} A_d(u^\varepsilon) = \varepsilon \sum_{d=1}^{D} \partial_{x_d} \left(\sum_{j=1}^{D} B_{dj}(u^\varepsilon) \partial_{x_j} u^\varepsilon \right) + \mathcal{O}(\varepsilon^2), \tag{2.17}$$

$$v_d^\varepsilon = A_d(u^\varepsilon) - \varepsilon \left(\partial_t v_d^\varepsilon + \sum_{j=1}^{D} \partial_{x_j} (P \Lambda_d \Lambda_j \mathcal{M}(u^\varepsilon)) \right) + \mathcal{O}(\varepsilon^2), \tag{2.18}$$

$$\text{with } B_{dj}(u) := P \Lambda_d \Lambda_j \mathcal{M}'(u) - A_d'(u) A_j'(u) \in \mathbb{R}^{S \times S} \quad \forall\, d, j = 1, \dots, D. \tag{2.19}$$ ∎

If we want the microscopic limit to be a stable approximation of the original equation, we have to impose a generalized Whitham's subcharacteristic condition on the final result (2.17) as stated in [9, 20, 10]. It must hold that

$$\sum_{j,d=1}^{D} (B_{dj} \xi_j, \xi_d) \geq 0 \qquad \forall \xi_1, \dots, \xi_D \in \mathbb{R}^S. \tag{2.20}$$

This condition can be interpreted as an imposition of positive diffusion to the equation (2.17).

2.2. AP property. The asymptotic behavior given by the Chapman–Enskog expansion is the property that we would like to maintain also at the discrete level. Schemes that verify this limit are called asymptotic preserving, or AP. This property can be summarized in the diagram of Figure 2.2. The macroscopic and microscopic analytical models are respectively denoted by $\mathcal{F}^0$ and $\mathcal{F}^\varepsilon$, meaning that $\mathcal{F}^0 := \lim_{\varepsilon \to 0} \mathcal{F}^\varepsilon$. The discretization of the kinetic model given by the scheme is defined as $\mathcal{F}_\Delta^\varepsilon$. The limit of this model is defined as $\mathcal{F}_\Delta^0$. We can say that a scheme is AP if $\lim_{\Delta \to 0} \mathcal{F}_\Delta^0 = \mathcal{F}^0$. In order to verify this property, we have to build a scheme that, in the discrete Chapman-Enskog expansion, behaves analogously to the analytical one.

3. AP IMEX first order scheme. In order to obtain a stable and AP scheme, we have to be careful in the time discretization. A natural choice for this class of problems is the IMEX schemes. They are particularly suited for the model (2.2), because, as ε vanishes, the source term becomes stiff. Classically, one should take discretization scales of the same order of the relaxation parameter, $\Delta t \sim h \sim \varepsilon$, where Δt is the size of a timestep and $h := \max_{\mathrm{E} \in \Omega} d(\mathrm{E})$ is the maximum diameter of an element of the domain. Obviously, this is not feasible as $\varepsilon \to 0$. Therefore, we need to treat the stiff term in an implicit way. The flux part will be discretized in an

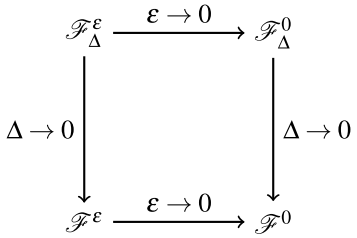


FIG. 2.2. *AP schemes.*

explicit way. The resulting IMEX discretization in time we obtain, after some initial condition $f^{0,\varepsilon} = f_0^\varepsilon(x)$, is the following:

$$\frac{f^{n+1,\varepsilon} - f^{n,\varepsilon}}{\Delta t} + \sum_{d=1}^{D} \Lambda_d \partial_{x_d} f^{n,\varepsilon} = \frac{1}{\varepsilon}\left(\mathcal{M}(Pf^{n+1,\varepsilon}) - f^{n+1,\varepsilon}\right), \tag{3.1}$$

where the superscript n indicates the known explicit timestep t^n or the unknown implicit timestep t^{n+1}.

Remark 3.1 (CFL conditions). Since the flux is explicitly discretized, we need to impose some restrictions on the timestep size, such that $\Delta t \leq \lambda^{-1}\mathrm{CFL}h$, where CFL is a number smaller than 1 that depends on the used polynomials. Here, λ is the convection coefficient in (2.5) and the spectral radius of Λ_d. The choice of this parameter is led by the Whitham's subcharacteristic condition (2.20), knowing that it is necessary that λ is bigger than the spectral radius of the original fluxes $\lambda \geq \rho(A_d)$, $d = 1, \ldots, D$, to verify the condition. This does not allow us to choose better CFL conditions than the ones of the macroscopic problem.

In the general case, the source may depend nonlinearly on the variable f^{n+1} and the solution of this system (of dimension L) must be found with nonlinear solvers such as the Newton–Raphson method. In the specific case of this kinetic model (2.2), we can exploit the property (2.3) to write the projection of the previous time discretization (3.1) and obtain

$$\frac{u^{n+1,\varepsilon} - u^{n,\varepsilon}}{\Delta t} + \sum_{d=1}^{D} P\Lambda_d \partial_{x_d} f^{n,\varepsilon} = 0, \tag{3.2}$$

where $u^{n,\varepsilon} := Pf^{n,\varepsilon}$. This resulting time discretization is totally explicit in time, so we can compute $u^{n+1,\varepsilon}$ without recurring to the nonlinear solver. Once we obtain this value, we can substitute it in (3.1) and collect all the $f^{n+1,\varepsilon}$ on the left-hand side, leading to the following explicit scheme:

$$f^{n+1,\varepsilon} = \frac{\varepsilon}{\Delta t + \varepsilon} f^{n,\varepsilon} - \frac{\varepsilon \Delta t}{\Delta t + \varepsilon} \sum_{d=1}^{D} \Lambda_d \partial_{x_d} f^{n,\varepsilon} + \frac{\Delta t}{\Delta t + \varepsilon} \mathcal{M}(u^{n+1,\varepsilon}). \tag{3.3}$$

We notice that ε never appears alone at the denominator, so for any value of ε the scheme will be stable. Moreover, if we let $\varepsilon \to 0$, using the property (2.4), the scheme is converging to

$$\begin{cases} u^{n+1} = u^n + \Delta t \sum_{d=1}^{D} A_d(u^n), \\ f^{n+1} = \mathcal{M}(u^{n+1}). \end{cases} \tag{3.4}$$

This coincides with the explicit Euler scheme for the macroscopic model (2.1).

Clearly, this scheme is only first order accurate in time, since the discretization has been done only at the previous or at the new timestep. We introduce a high order accurate discretization in space (residual distribution) and the DeC procedure to achieve high order accuracy in time.

4. Residual distribution schemes. In this section we introduce the spatial and time discretization given by RD schemes [1, 15] and the DeC approach [4, 16].

4.1. Notation. To simplify the notation, we rewrite (2.2) as

$$\partial_t f + \sum_{d=1}^{D} \partial_{x_d} \Lambda_d f - S(f) = 0, \tag{4.1}$$

where f is the variable of the equation and S is the source term. The RD framework is based on the FEM discretization, so we proceed defining a triangulation Ω_h on our domain Ω, denoting by E the generic element of the mesh and by h the characteristic mesh size (implicitly supposing some regularity on the mesh). Following the ideas of the Galerkin FEM, we use an approximation space V_h for the solutions given by globally continuous piecewise polynomials of degree p:

$$V_h := \{f \in \mathcal{C}^0(\Omega_h),\ f|_{\mathrm{E}} \in \mathbb{P}^p \quad \forall \mathrm{E} \in \Omega_h\}. \tag{4.2}$$

Now we can rewrite the numerical solution $f_h(x) \approx f(x)$ as a linear combination of compactly supported basis functions $\varphi_\sigma \in V_h$ through the coefficients f_σ for every degree of freedom $\sigma \in D_h$. This can be written as

$$f_h(x) = \sum_{\sigma \in D_h} f_\sigma \varphi_\sigma(x) = \sum_{\mathrm{E} \in \Omega_h} \sum_{\sigma \in \mathrm{E}} f_\sigma \varphi_\sigma|_{\mathrm{E}}(x) \quad \forall x \in \Omega, \tag{4.3}$$

where D_h is the set of all the degrees of freedom of Ω_h, so that $\{\varphi_\sigma : \sigma \in D_h\}$ is a basis for V_h, and the coefficient f_σ must be found with a numerical method.

4.2. Residual distribution algorithm. RD schemes can be summarized as follows and as sketched in Figure 4.1.

1. Define $\forall \mathrm{E} \in \Omega_h$ a fluctuation term (total residual)[1]

$$\phi^{\mathrm{E}} := \int_{\mathrm{E}} \left(\sum_{d=1}^{D} \partial_{x_d} \Lambda_d f_h - S(f_h) \right) dx = \int_{\partial \mathrm{E}} \sum_{d=1}^{D} \Lambda_d f_h \cdot \mathbf{n} d\Gamma - \int_{\mathrm{E}} S(f_h) dx. \tag{4.4}$$

2. Split the total residual ϕ^{E} into nodal residuals ϕ^{E}_σ for every degree of freedom σ not vanishing in the cell E, i.e.,

$$\phi^{\mathrm{E}} = \sum_{\sigma \in \mathrm{E}} \phi^{\mathrm{E}}_\sigma \quad \forall \mathrm{E} \in \Omega_h. \tag{4.5}$$

In Appendix A or in [2, 7] one can find more details on possible definitions of the nodal residuals.

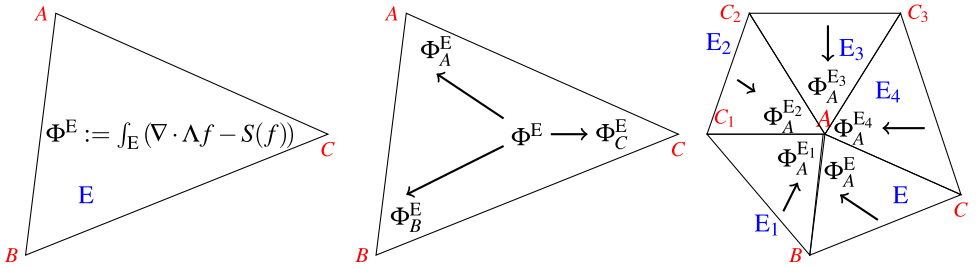


FIG. 4.1. *Defining total residual and nodal residuals and building the RD scheme.*

[1]The second formulation of (4.4) can be used to rewrite the DG or FV numerical flux into the RD framework as in [5].

3. The resulting scheme is obtained for each degree of freedom σ by summing all the nodal residual contributions from different elements E, that is,

$$f_\sigma^{n+1} = f_\sigma^n - \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \phi_\sigma^{\mathrm{E}} \quad \forall \sigma \in D_h. \tag{4.6}$$

The key of the scheme is the definition of nodal residuals. This choice is the actual definition of the spatial discretization. Equation (4.5) is guaranteeing the conservation of the scheme. The high order accuracy in space can be achieved choosing higher order polynomial basis functions and consistent nodal residuals with high order artificial diffusion. The stability must be reached with some stabilization terms that must be added to the nodal residuals, always maintaining (4.5). In [1, 4, 5] it has been shown that well-known FEM or finite volume schemes (such as SUPG, DG, FV-WENO, etc.) can be rewritten in terms of RD, just choosing the proper nodal residuals.

Details and some examples of the schemes can be found in Appendix A. In particular, we will use the residual distributions, and hence the schemes, defined and tested in [6].

4.3. Time discretization. In this section, we will introduce the explicit DeC algorithm of [4, 6]. This is a preliminary step to understand the relative IMEX version that we will present in section 5. To introduce the DeC algorithm, we have to follow a particular discretization of the variables in time. Following the idea of many one-step time integration schemes, such as Runge–Kutta (RK), arbitrary high order using derivatives, and so on, we build a high order approximation of the time evolution through stages in the time interval. To do so, we discretize the timestep $[t^n, t^{n+1}]$ into M subtimesteps $[t^{n,0}, t^{n,1}], \dots, [t^{n,M-1}, t^{n,M}]$ and the variable f_h in time at each subtimestep $f_h^{n,m}$ as in Figure 4.2.

The Picard–Lindelöf theorem proves the existence and uniqueness of the solution of an ODE, making use of the so-called Picard iterations. We follow the statement result of the theorem writing for $m = 1, \dots, M$

$$f_h^{n,m} = f_h^n - \int_{t^n}^{t^{n,m}} (\nabla \cdot A(f_h(x,s)) - S(f_h(x,s)))\, ds. \tag{4.7}$$

More precisely, the scheme that we want to solve is a system of equations, where each entry is the discretization of (4.7) for a different $m = 1, \dots, M$. For the flux and source terms, we use the discretization produced with the RD method, while the finite difference of the time derivative is simply approached with a Galerkin residual. Let us define $\underline{f} := (f^0, \dots, f^M)$ the vector of variables for all the subtimesteps, avoiding the obvious index of the timestep n and the discretization index h. In practice, for all the degrees of freedom $\sigma \in D_h$, we can write the operator $\mathcal{L}^2$ that we are interested in as

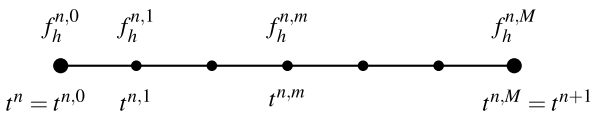


FIG. 4.2. *Subtimesteps.*

(4.8)
$$\mathcal{L}^2_\sigma(\underline{f}) := \begin{pmatrix} \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{\mathrm{E}} \varphi_\sigma (f^1 - f^0)dx + \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{t^{n,0}}^{t^{n,1}} \mathcal{I}_M(\phi^{\mathrm{E}}_\sigma(f^0),\dots,\phi^{\mathrm{E}}_\sigma(f^M),s)ds \\ \dots \\ \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{\mathrm{E}} \varphi_\sigma (f^M - f^0)dx + \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{t^{n,0}}^{t^{n,M}} \mathcal{I}_M(\phi^{\mathrm{E}}_\sigma(f^0),\dots,\phi^{\mathrm{E}}_\sigma(f^M),s)ds \end{pmatrix}.$$

The $\mathcal{L}^2$ operator is composed of M equations with M unknowns $f^1,\dots,f^M$, the function $\mathcal{I}_M$ is an interpolation polynomial in nodes $\{t^{n,m}\}_{m=0}^M$, and the time integration is computed using quadrature formulas in the same interpolation points. After applying the quadrature rule, the time integration of the flux and source can be rewritten as

$$\int_{t^{n,0}}^{t^{n,m}} \mathcal{I}_M(\phi^{\mathrm{E}}_\sigma(f^0),\dots,\phi^{\mathrm{E}}_\sigma(f^M),s)ds = \Delta t \sum_{r=0}^{M} \theta^m_r \phi^{\mathrm{E}}_\sigma(f^r). \tag{4.9}$$

What we aim for is the solution of the system $\mathcal{L}^2(\underline{f}^*) = 0$. This is a system containing many implicit, in general, nonlinear terms and can be seen as an implicit RK method. We do not want to make use of nonlinear solvers to find the solution of this system of $M \times |D_h|$ equations. Nevertheless, the solution $\underline{f}^*$ is an approximation of the exact solution with an accuracy of order $M+1$ in time and $p+1$ in space, where p is the degree of the utilized polynomials.

The core of the DeC algorithm, as presented in [4], is an iterative procedure that uses two operators, one high order and one low order explicit or easy to solve. So, we introduce a first order approximation of the scheme $\mathcal{L}^2$ presented in [4, 6] that we will call $\mathcal{L}^1$:

(4.10)
$$\mathcal{L}^1_\sigma(\underline{f}) := \begin{pmatrix} (f^1_\sigma - f^0_\sigma) \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{\mathrm{E}} \varphi_\sigma dx + \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{t^{n,0}}^{t^{n,1}} \mathcal{I}_0(\phi^{\mathrm{E}}_\sigma(f^0),\dots,\phi^{\mathrm{E}}_\sigma(f^M),s)ds \\ \dots \\ (f^M_\sigma - f^0_\sigma) \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{\mathrm{E}} \varphi_\sigma dx + \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{t^{n,0}}^{t^{n,M}} \mathcal{I}_0(\phi^{\mathrm{E}}_\sigma(f^0),\dots,\phi^{\mathrm{E}}_\sigma(f^M),s)ds \end{pmatrix}.$$

The first simplification applied is a mass lumping on the derivative in time, where we pass from the integral of the $\mathcal{L}^2$ operator of $\int_{\mathrm{E}} \varphi_\sigma f^m = \sum_j \int_{\mathrm{E}} \varphi_\sigma \varphi_j f^m_j$ to $\int_{\mathrm{E}} \varphi_\sigma f^m_\sigma$, that produces a diagonal mass matrix. The inversion of this mass matrix is only possible if $|\mathrm{E}_\sigma| := \sum_{\mathrm{E}} \int_{\mathrm{E}} \varphi_\sigma(x)dx > 0$ for all the degrees of freedom. For this reason, we will always consider Bernstein polynomials $\mathbb{B}^p$, which are nonnegative on the cells of interest, instead of Lagrange polynomial $\mathbb{P}^p$, as basis functions for every cell E. This choice and its practical implementation are explained in detail in [6]. In particular, the usage of barycentric coordinates and a map to a reference element are crucial in this procedure. The mass lumping introduces an error with respect to the previous method of the order $\mathcal{O}(h)$.

The second simplification is in the residual part, where we substituted the high order interpolant $\mathcal{I}_M$ with the left Riemann sum, that consists of the constant interpolant $\mathcal{I}_0$ in the beginning stage $f^{n,0}$, resulting in an explicit right-hand side. The

final first order scheme $\mathcal{L}^1(\underline{f}) = 0$ is, hence, explicit and easy to solve. The considered interpolant polynomial can be rewritten as

$$\int_{t^{n,0}}^{t^{n,m}} \mathcal{I}_0\left(\phi_\sigma^{\mathrm{E}}(f^0), \dots, \phi_\sigma^{\mathrm{E}}(f^M), s\right) = \Delta t \beta^m \phi_\sigma^{\mathrm{E}}(f^0), \tag{4.11}$$

where $\beta^m := \frac{t^{n,m}-t^{n,0}}{t^{n+1}-t^n}$. This approximation in time is a first order approximation and brings an error of order $\mathcal{O}(\Delta t^2)$ with respect to the $\mathcal{L}^2$ formulation if the solution is regular enough.

To incorporate the properties of the AP IMEX time discretization we have studied in section 3, we need to redefine the interpolant $\mathcal{I}_0$. The details will be given in section 5.

4.4. Deferred correction algorithm. Now, we present the DeC algorithm. It was introduced by Dutt, Greengard, and Rokhlin in [16] and then an implicit version was proposed by Minion in [21]. In [4] the DeC is used to obtain a mass-matrix-free scheme and, doing so, it rewrites the same DeC algorithm in a slightly different formulation with two operators $\mathcal{L}^1$ and $\mathcal{L}^2$. This allows us to easily prove that the proposed method verifies the hypothesis of the DeC algorithm. The DeC scheme is an algorithm that allows us to obtain a high order scheme starting from a low order one in a general way. It has already been used for implicit schemes in ODE and PDE contexts, as well as in combination with RK schemes; see, for example, [11, 22].

The method consists in an iterative procedure that mimics the Picard iterations and reduces at each step the error between the iteration variables and the solution of the high order method. In our case, the high order method that we want to approximate is $\mathcal{L}^2(\underline{f}^*) = 0$ given by (4.8). We will denote the iteration coefficient as (k) and the variables related to the iteration with the superscript (k) as $f^{(k)}$, while K is the maximum number of iterations. We keep the notation for the subtimesteps m without brackets, e.g., $f^{m,(k)}$ denotes the discretized variable f at the subtimestep $m \in \{0, \dots, M\}$ at the iteration $k \in \{0, \dots, K\}$. We omit the timestep index n for clarity of the notation. The algorithm proceeds as follows:

$$\begin{aligned} &f^{m,(0)} := f(t^n) \quad \forall m = 1, \dots, M; \\ &f^{0,(k)} := f(t^n) \quad \forall k = 1, \dots, K; \\ &\mathcal{L}^1(\underline{f}^{(k)}) = \mathcal{L}^1\left(\underline{f}^{(k-1)}\right) - \mathcal{L}^2\left(\underline{f}^{(k-1)}\right) \text{ with } k = 1, \dots, K. \end{aligned} \tag{4.12}$$

Given the DeC procedure (4.12), we can state the following proposition as in [4].

PROPOSITION 4.1. *Let $\mathcal{L}^1$ and $\mathcal{L}^2$ be two operators defined on V_h^m, which depend on the discretization scale $\Delta \sim h \sim \Delta t$, such that*

- *$\mathcal{L}^1$ is coercive with respect to a norm, i.e., $\exists \alpha_1 > 0$ independent of Δ, such that we have that*

$$\alpha_1 ||\underline{f} - \underline{g}|| \le ||\mathcal{L}^1(\underline{f}) - \mathcal{L}^1(\underline{g})|| \quad \forall\, \underline{f}, \underline{g},$$

- *$\mathcal{L}^1 - \mathcal{L}^2$ is Lipschitz with constant $\alpha_2 > 0$ uniformly with respect to Δ, i.e.,*

$$||(\mathcal{L}^1_\Delta(\underline{f}) - \mathcal{L}^2_\Delta(\underline{f})) - (\mathcal{L}^1_\Delta(\underline{g}) - \mathcal{L}^2_\Delta(\underline{g}))|| \le \alpha_2 \Delta ||\underline{f} - \underline{g}|| \quad \forall\, \underline{f}, \underline{g}.$$

We also assume that there exists a unique $\underline{f}^*_\Delta$ such that $\mathcal{L}^2(\underline{f}^*_\Delta) = 0$. Then, if $\eta := \frac{\alpha_2}{\alpha_1}\Delta < 1$, the DeC is converging to $\underline{f}^*$ and after K iterations the error $||\underline{f}^{(K)} - \underline{f}^*||$ is smaller than $\eta^K ||\underline{f}^{(0)} - \underline{f}^*||$.

Proof. By definition, we know that $\mathcal{L}^1(\underline{f}^*) = \mathcal{L}^1(\underline{f}^*) - \mathcal{L}^2(\underline{f}^*)$, so that

$$\mathcal{L}^1\left(\underline{f}^{(k+1)}\right) - \mathcal{L}^1(\underline{f}^*) = \left(\mathcal{L}^1(\underline{f}^{(k)}) - \mathcal{L}^1(\underline{f}^*)\right) - \left(\mathcal{L}^2(\underline{f}^{(k)}) - \mathcal{L}^2(\underline{f}^*)\right), \tag{4.13}$$

$$\begin{aligned} &\alpha_1 ||\underline{f}^{(k+1)} - \underline{f}^*|| \le ||\mathcal{L}^1(\underline{f}^{(k+1)}) - \mathcal{L}^1(\underline{f}^*)|| \\ &\quad = ||\mathcal{L}^1(\underline{f}^{(k)}) - \mathcal{L}^2(\underline{f}^{(k)}) - (\mathcal{L}^1(\underline{f}^*) - \mathcal{L}^2(\underline{f}^*))|| \le \alpha_2 \Delta ||\underline{f}^{(k)} - \underline{f}^*||. \end{aligned} \tag{4.14}$$

Hence, we can write

$$||\underline{f}^{(k+1)} - \underline{f}^*|| \le \left(\frac{\alpha_2}{\alpha_1}\Delta\right) ||\underline{f}^{(k)} - \underline{f}^*|| \le \left(\frac{\alpha_2}{\alpha_1}\Delta\right)^{k+1} ||\underline{f}^{(0)} - \underline{f}^*||. \tag{4.15}$$

After k iterations we have an error at most of $\eta^k \cdot ||\underline{f}^{(0)} - \underline{f}^*||$. ∎

The proof of the properties of $\mathcal{L}^1$ and $\mathcal{L}^2$, which depend on their definitions, can be found for our specific case in Appendix B.1.

Proposition 4.1 states that at each iteration we gain one order of accuracy with respect to the previous correction $(k-1)$. Notice that we always solve the equations for the unknown variable $\underline{f}^{(k)}$ which appears only in the $\mathcal{L}^1$ formulation, the one that can be easily solved, while $\mathcal{L}^2$ is only applied to already computed predictions of the solution $\underline{f}^{(k-1)}$.

Remark 4.2 (computational costs and order of accuracy). Proposition 4.1 tells us that if the method $\mathcal{L}^2$ is accurate with order of accuracy z, namely it has $M = z - 1$ subtimesteps, then we should perform $K = z$ iterations for every timestep of the method. For space accuracy, we will use $p = z - 1$ polynomial order for the basis functions. For example, $\mathbb{B}^1$ basis functions and $K = 2$ iterations of the DeC with 1 subtimestep ($t^{n,0} = t^n$, $t^{n,1} = t^{n+1}$) amount to an RK2 method; see [24]. In all our test cases we will use the same number of degrees of polynomials, corrections-1, and subtimesteps, i.e., $p = K - 1 = M$.

Remark 4.3 (comparison with RK schemes). First of all, the presented DeC scheme does not make use of mass matrices, sparing the cost of its inversion and the multiplication, passing from a cost of $\mathcal{O}(|D_h|^2)$ to $\mathcal{O}(|D_h|)$. Any high order RK method without mass matrix would require extra efforts in the formulation of the scheme to compensate for this fact; see [15, 24]. Nevertheless, a z-order DeC scheme can be written as an RK scheme with $(M-1) \times K = z(z-1) \approx z^2$ stages, but the M subtimesteps are independent one from another and can be performed in parallel, reducing the time cost to just $K = z$ corrections for any order of accuracy, which is faster than or comparable to RK where the stages are bigger than or equal to z. Moreover, the coefficients of the time integration are automatically given by the polynomials used, so it does not require a different definition for different orders, resulting in an arbitrary high order accurate schemes.

Remark 4.4 (distribution of subtimestep points in DeC). In this work, we considered equidistributed subtimestep points $t^{n,m} = t^n + \frac{m}{M}\Delta t$ both to define the polynomials in time and as quadrature points in time. Other choices may have more advantages and stability properties, as shown in [13], for example, Gauss–Legendre points were already used in [16]. It is also possible not to include the start and end points t^m, t^{n+1} and extrapolate the final point with interpolation polynomials. This choice varies the stability properties of the time integration scheme. It has been shown that the schemes generated by other distributions, e.g., Chebyshev or Gauss–Legendre, give

better stability properties for very high orders. Since we consider at most order 4, we have not noticed remarkable differences in results between distributions. Hence, for ease of computation, we consider the equispaced points.

Example 4.5 (explicit DeC). We present an example of the explicit DeC procedure for second order accuracy. Take $M = 1$ subtimestep $t^n = t^0$, $t^{n+1} = t^1$ and $K = 2$ iterations. Recalling that $f^{0,(0)} = f^{1,(0)}$, the scheme for the first iteration reads

$$f^{0,(0)} = f^{1,(0)} = f^{0,(1)} = f^{0,(2)} = f^n, \tag{4.16a}$$

$$\mathcal{L}^1(\underline{f}^{(1)}) = \mathcal{L}^1(\underline{f}^{(0)}) - \mathcal{L}^2(\underline{f}^{(0)}), \tag{4.16b}$$

$$\begin{aligned} f_\sigma^{1,(1)} - f_\sigma^n + \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi_\sigma^{\mathrm{E}}(f^n) &= f_\sigma^{1,(0)} - f_\sigma^n + \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi_\sigma^{\mathrm{E}}(f^n) - f_\sigma^{1,(0)} \\ &\quad + f_\sigma^n - \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{r=0}^{1}\theta_r^1\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi_\sigma^{\mathrm{E}}(f^{r,(0)}) \end{aligned} \tag{4.16c}$$

$$\iff f_\sigma^{1,(1)} = f_\sigma^n - \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi_\sigma^{\mathrm{E}}(f^n). \tag{4.16d}$$

The second and last iteration reads

$$\mathcal{L}^1(\underline{f}^{(2)}) = \mathcal{L}^1(\underline{f}^{(1)}) - \mathcal{L}^2(\underline{f}^{(1)}), \tag{4.17a}$$

$$\begin{aligned} f_\sigma^{1,(2)} - f_\sigma^n + \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi_\sigma^{\mathrm{E}}(f^n) &= f_\sigma^{1,(1)} - f_\sigma^n + \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi_\sigma^{\mathrm{E}}(f^n) - f_\sigma^{1,(1)} \\ &\quad + f_\sigma^n - \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{r=0}^{1}\theta_r^1\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi_\sigma^{\mathrm{E}}(f^{r,(1)}) \end{aligned} \tag{4.17b}$$

$$\iff f^{n+1} = f_\sigma^{1,(2)} = f_\sigma^n - \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{r=0}^{1}\theta_r^1\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi_\sigma^{\mathrm{E}}(f^{r,(1)}). \tag{4.17c}$$

For this simple second order case, the scheme coincides with the strong stability preserving RK method of second order [18].

5. IMEX DeC kinetic scheme. Now we want to combine the time discretization of the IMEX scheme (3.1) and the DeC method. The IMEX discretization is a first order discretization, thus, it can only affect the $\mathcal{L}^1$ operator. On the contrary, to get high order of accuracy through the DeC procedure, the $\mathcal{L}^2$ operator must remain the same as (4.8). To modify $\mathcal{L}^1$, we have to introduce a few new terms. In particular, we have to treat separately the time derivative, the fluxes, and the source term. This implies a new definition of total (4.4) and nodal (4.5) residuals of the RD scheme.

As in (3.1), we want the zero order interpolant $\mathcal{I}_0$ to be explicit in the fluxes and implicit in the source term. In the subtimestep context of the DeC formulation, this means that the source term is evaluated constantly at the end of the subtimestep, namely in $t^{n,m}$, while the fluxes are evaluated at the beginning of the timestep $t^{n,0}$. Moreover, in order to invert the system, we apply a mass lumping also on the source term, as we did for the time derivative in $\mathcal{L}^1$ (4.10). This leads to the following definitions:

(5.1)
$$\phi^{\mathrm{E}}_{source} := \int_{\mathrm{E}} \frac{\mathcal{M}\left(Pf^{n,m,\varepsilon}\right) - f^{n,m,\varepsilon}}{\varepsilon} dx, \quad \phi^{\mathrm{E}}_{source,\sigma} := \int_{\mathrm{E}} \varphi_\sigma(x) \frac{\mathcal{M}\left(Pf^{n,m,\varepsilon}_\sigma\right) - f^{n,m,\varepsilon}_\sigma}{\varepsilon} dx,$$

(5.2)
$$\phi^{\mathrm{E}}_{ad} = \int_{\mathrm{E}} \sum_{d=1}^{D} \Lambda_d \partial_{x_d} f^{n,0,\varepsilon} dx.$$

With definition (5.1), we can collect the degrees of freedom of the source outside the integral and have a linear dependency on the unknown $f^{n,m,\varepsilon}_\sigma$, thanks to the projection trick explained in (3.2). The total advection residuals (5.2), on the contrary, behave as before, while the nodal residuals $\phi^{\mathrm{E}}_{ad,\sigma}$ can be defined in many ways, according to the scheme we want to achieve; see Appendix A.

So, if we rewrite the $\mathcal{L}^1$ operator explicitly, we get

(5.3)
$$\begin{aligned}\mathcal{L}^1_\sigma(f^{n,0},\dots,f^{n,M}) &= \mathcal{L}^1_\sigma(\underline{f}) \\ &:= \begin{pmatrix} |\mathrm{E}_\sigma|(f^{n,1}_\sigma - f^{n,0}_\sigma) + \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \beta^1 \Delta t \phi^{\mathrm{E}}_{ad,\sigma}(f^{n,0}) + |\mathrm{E}_\sigma| \frac{\beta^1 \Delta t}{\varepsilon}\left(\mathcal{M}(Pf^{n,1}_\sigma) - f^{n,1}_\sigma\right) \\ \cdots \\ |\mathrm{E}_\sigma|(f^{n,M}_\sigma - f^{n,0}_\sigma) + \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \beta^M \Delta t \phi^{\mathrm{E}}_{ad,\sigma}(f^{n,0}) + |\mathrm{E}_\sigma| \frac{\beta^M \Delta t}{\varepsilon}(\mathcal{M}(Pf^{n,M}_\sigma) - f^{n,M}_\sigma) \end{pmatrix}.\end{aligned}$$

The system $\mathcal{L}^1 = 0$ can be solved without recurring to any nonlinear solver if we use projection P on the whole operator, defining the *u auxiliary operator* $\mathcal{L}^{1,m}_{\sigma,u} := P\mathcal{L}^{1,m}_\sigma$. Indeed, what we get is the following operators for each subtimestep $m = 1,\dots,M$, defining $\Delta t^m := \beta^m \Delta t$:

(5.4a)
$$\mathcal{L}^{1,m}_{\sigma,u}(\underline{f}) = |\mathrm{E}_\sigma|(Pf^m_\sigma - Pf^0_\sigma) + \Delta t^m \sum_{\mathrm{E}|\sigma\in\mathrm{E}} P\phi^{\mathrm{E}}_{ad,\sigma}(f^0);$$

(5.4b)
$$\mathcal{L}^{1,m}_\sigma(\underline{f}) = |\mathrm{E}_\sigma|\left(1 + \frac{\Delta t^m}{\varepsilon}\right) f^m_\sigma - |\mathrm{E}_\sigma| f^0_\sigma + \Delta t^m \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \phi^{\mathrm{E}}_{ad,\sigma}(f^0) - |\mathrm{E}_\sigma| \frac{\Delta t^m}{\varepsilon} \mathcal{M}(Pf^m_\sigma).$$

Equation (5.4a) can be solved explicitly for Pf^m; then, we can substitute it in the Maxwellian term of (5.4b), which is given by (5.3) collecting all the unknown terms f^m. Given this, we can solve $\mathcal{L}^1 = 0$ for f^m explicitly, from a computational point of view. Moreover, as before, we can see that (5.4b) does not lead to terms with ε alone at the denominator. Indeed, it can be rewritten as

(5.4c)
$$\frac{\varepsilon \cdot \mathcal{L}^{1,m}_\sigma(\underline{f})}{|\mathrm{E}_\sigma|(\varepsilon + \Delta t^m)} = f^m_\sigma - \frac{\varepsilon \cdot f^0_\sigma}{\varepsilon + \Delta t^m} + \frac{\varepsilon \Delta t^m}{|\mathrm{E}_\sigma|(\varepsilon + \Delta t^m)} \sum_{K|\sigma\in K} \phi^K_{ad,\sigma}(f^0) - \frac{\Delta t^m}{\varepsilon + \Delta t^m} \mathcal{M}(Pf^m_\sigma).$$

This guarantees that, as $\varepsilon \to 0$, we are not facing any stiffness.

Finally, we can write a general term of the correction DeC procedure for the $(k+1)$th correction and the mth subtimestep. With the *auxiliary equation* we find $Pf^{m,(k+1)}$,

(5.5a)

$$\begin{aligned}&\mathcal{L}^{1,m}_{\sigma,u}(\underline{f}^{(k+1)}) - \mathcal{L}^{1,m}_{\sigma,u}(\underline{f}^{(k)}) + \mathcal{L}^{2,m}_{\sigma,u}(\underline{f}^{(k)}) \stackrel{!}{=} 0,\\ &\quad |\mathrm{E}_\sigma|(Pf^{m,(k+1)}_\sigma - Pf^{m,(k)}_\sigma)\\ &\qquad + \sum_{\mathrm{E}|\sigma\in\mathrm{E}}\left[\int_{\mathrm{E}}\varphi_\sigma(x)(u^{m,(k)}(x) - u^{0,(k)}(x))dx + \Delta t\sum_{r=0}^{M}\theta^m_r P\phi^{\mathrm{E}}_\sigma(f^{r,(k)})\right] \stackrel{!}{=} 0,\end{aligned}$$

and, then, the equation for the kinetic unknown $f^{m,(k+1)}$,

(5.5b)

$$\begin{aligned}&\mathcal{L}^{1,m}_{\sigma}(\underline{f}^{(k+1)}) - \mathcal{L}^{1,m}_{\sigma}(\underline{f}^{(k)}) + \mathcal{L}^{2,m}_{\sigma}(\underline{f}^{(k)}) \stackrel{!}{=} 0,\\ &\quad |\mathrm{E}_\sigma|\left(1+\frac{\Delta t^m}{\varepsilon}\right)(f^{m,(k+1)}_\sigma - f^{m,(k)}_\sigma) - |\mathrm{E}_\sigma|\frac{\Delta t^m}{\varepsilon}\left(\mathcal{M}\left(Pf^{m,(k+1)}_\sigma\right) - \mathcal{M}\left(Pf^{m,(k)}_\sigma\right)\right)\\ &\qquad + \sum_{\mathrm{E}|\sigma\in\mathrm{E}}\left[\int_{\mathrm{E}}\varphi_\sigma(x)(f^{m,(k)}(x) - f^{0,(k)}(x))dx + \Delta t\sum_{r=0}^{M}\theta^m_r\phi^{\mathrm{E}}_\sigma(f^{r,(k)})\right] \stackrel{!}{=} 0.\end{aligned}$$

Again, thanks to the factor $\left(1+\frac{\Delta t^m}{\varepsilon}\right)$ in front of the unknown $f^{m,(k+1)}$, we are sure not to have any stiff terms, even in the source of the residuals $\phi^E_\sigma(f^{r,(k)})$ of $\mathcal{L}^2$.

Example 5.1 (IMEX DeC scheme). We show an example of the second order scheme of the IMEX DeC algorithm, where we have $M = 1$ subtimestep and $K = 2$ DeC iterations. The variables for any subtimesteps m at the correction (0) are initialized as $f^{m,(0)} := f^n$ and the beginning steps for all corrections k as well $f^{0,(k)} := f^n$. Then, we proceed solving the projected operator. At the first iteration, it coincides with explicit Euler, i.e.,

$$Pf^{1,(1)}_\sigma := Pf^0 - \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{\mathrm{E}|\sigma\in\mathrm{E}} P\phi^{\mathrm{E}}_{\sigma,ad}(f^0). \tag{5.6a}$$

Then, we can use this result to solve (5.5b) for $f^{1,(1)}$ inverting the beginning coefficient, i.e.,

$$f^{1,(1)}_\sigma := f^0 + \frac{\Delta t}{\Delta t+\varepsilon}(\mathcal{M}(Pf^{1,(1)}_\sigma) - \mathcal{M}(Pf^0_\sigma)) - \frac{\varepsilon\Delta t}{|\mathrm{E}_\sigma|(\Delta t+\varepsilon)}\sum_{\mathrm{E}|\sigma\in\mathrm{E}}\phi^{\mathrm{E}}_\sigma(f^0). \tag{5.6b}$$

Note that the nodal residuals of the $\mathcal{L}^2$ operators contain source terms that are an $\mathcal{O}(\frac{1}{\varepsilon})$ but that part is premultiplied by ε itself, leading to a stable approximation. At the moment, we have a first order approximation of the solution. Performing the second correction, we obtain a second order approximation, i.e.,

(5.6c)

$$Pf^{1,(2)}_\sigma := Pf^{1,(1)} - \sum_{\mathrm{E}|\sigma\in\mathrm{E}}\left(\int_{\mathrm{E}}\frac{\varphi_\sigma}{|\mathrm{E}_\sigma|}(Pf^{1,(1)} + Pf^0)dx - \frac{\Delta t}{|\mathrm{E}_\sigma|}\sum_{r=0}^{1}\theta^1_r P\phi^{\mathrm{E}}_{\sigma,ad}(f^{r,(1)})\right).$$

Here, we used the fact that for one subtimestep $\theta^1_0 = \theta^1_1 = \frac{1}{2}$. What we obtain is essentially a strong stability preserving second order RK, with a correction term for the mass matrix that we lumped. The last step for the final kinetic variable $f^{1,(2)}$ is

(5.6d)
$$f_\sigma^{n+1} = f_\sigma^{1,(2)} := f^{1,(1)} + \frac{\Delta t}{\Delta t+\varepsilon}(\mathcal{M}(Pf_\sigma^{1,(2)}) - \mathcal{M}(Pf_\sigma^{1,(1)})) - \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \frac{\varepsilon}{|\mathrm{E}_\sigma|(\Delta t+\varepsilon)}\left(\int_{\mathrm{E}} \varphi_\sigma(f^{1,(1)} - f^0)dx + \Delta t\frac{\phi_\sigma^{\mathrm{E}}(f^0) + \phi_\sigma^{\mathrm{E}}(f^{1,(1)})}{2}\right).$$

As before, the source terms in the nodal residuals of $\mathcal{L}^2$ are controlled by the ε in front of them. Finally, we have a second order approximation for the microscopic variable.

5.1. AP property of the IMEX DeC scheme. As for the first order scheme, we have to prove that the whole IMEX DeC discretization is AP. This means that when we let the relaxation term vanish, we should recast a scheme consistent with the macroscopic model (2.1). We will expand all the terms in ε and we will keep track also of the $\mathcal{O}(\Delta)$. Notice that ε goes to 0 before Δ, in other words, $\mathcal{O}(\frac{\varepsilon}{\Delta t}) = \mathcal{O}(\varepsilon)$; see also Figure 2.2.

THEOREM 5.2 (IMEX DeC is AP). *Suppose that at t^n the variable f^n is such that*

$$f^n = \mathcal{M}(Pf^n) + \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta); \tag{5.7}$$

then, at each subtimestep $m = 1, \dots, M$ and for every correction $k = 0, \dots, K$ and every degree of freedom $\sigma \in D_h$

$$\frac{Pf_\sigma^{m,(k)} - Pf_\sigma^0}{\Delta t^m} + \sum_{d=1}^{D} \partial_{x_d} A_d(Pf^0) + \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta) = 0, \tag{5.8a}$$

$$f^{m,(k)} = \mathcal{M}(Pf^{m,(k)}) + \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta). \tag{5.8b}$$

Proof. We will prove the statement by induction on the corrections $k = 0, \dots, K$. For the correction $k = 0$ we know from the initial conditions that the theses hold. So, given that (5.8a) and (5.8b) hold for k and for any $m = 1, \dots, M$, we have to prove the same properties for $k+1$. Let us consider the projection of the DeC scheme (5.5a). We will split it into $\mathcal{L}_u^{1,m}(\underline{f}^{(k+1)})$ and $\mathcal{L}_u^{1,m}(\underline{f}^{(k)}) - \mathcal{L}_u^{2,m}(\underline{f}^{(k)})$. The first term gives us

$$\mathcal{L}_u^{1,m}(\underline{f}^{(k+1)}) = \frac{Pf_\sigma^{m,(k+1)} - Pf_\sigma^0}{\Delta t^m} + \beta^m \sum_{d=1}^{D} \partial_{x_d} P\Lambda_d f^0 \tag{5.9a}$$

$$= \frac{Pf_\sigma^{m,(k+1)} - Pf_\sigma^0}{\Delta t^m} + \beta^m \sum_{d=1}^{D} \partial_{x_d} P\Lambda_d \mathcal{M}(Pf^0) + \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta) \tag{5.9b}$$

$$= \frac{Pf_\sigma^{m,(k+1)} - Pf_\sigma^0}{\Delta t^m} + \beta^m \sum_{d=1}^{D} \partial_{x_d} A_d(Pf^0) + \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta). \tag{5.9c}$$

Here, we used in (5.9b) the initial hypothesis (5.7) and in (5.9c) we have used the property (2.4). The second term gives us

$$\mathcal{L}_u^{1,m}(\underline{f}^{(k)}) - \mathcal{L}_u^{2,m}(\underline{f}^{(k)}) \tag{5.9d}$$

$$= \frac{Pf_\sigma^{m,(k)} - Pf_\sigma^0}{\Delta t^m} + \beta^m \sum_{d=1}^{D} \partial_{x_d} P\Lambda_d f^0 - \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{\mathrm{E}} \frac{\varphi_\sigma}{|\mathrm{E}_\sigma|} \frac{Pf^{m,(k)} - Pf^0}{\Delta t^m} \tag{5.9e}$$

$$- \sum_{d=1}^{D} \sum_{r=0}^{M} \theta_r^m \partial_{x_d} P\Lambda_d f^{r,(k)}.$$

The two time derivatives differ by a mass lumping that leads to an $\mathcal{O}(\Delta)$ error. For the property (2.4), we can write

$$\mathcal{L}_u^{1,m}(\underline{f}^{(k)}) - \mathcal{L}_u^{2,m}(\underline{f}^{(k)}) \tag{5.9f}$$

$$= \beta^m \sum_{d=1}^{D} \partial_{x_d} P\Lambda_d \mathcal{M}(Pf^0) - \sum_{d=1}^{D} \sum_{r=0}^{M} \theta_r^m \partial_{x_d} P\Lambda_d \mathcal{M}(Pf^{r,(k)}) \tag{5.9g}$$

$$+ \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta)$$

$$= \beta^m \sum_{d=1}^{D} \partial_{x_d} P\Lambda_d \mathcal{M}(Pf^0) - \sum_{d=1}^{D} \beta^m \partial_{x_d} P\Lambda_d \mathcal{M}(Pf^0) + \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta) \tag{5.9h}$$

$$= \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta).$$

In the last step, we have used the induction hypothesis (5.8a) that gives us an $\mathcal{O}(\Delta) + \mathcal{O}(\varepsilon)$. Now, if we sum the two contributions $\mathcal{L}_u^{1,m}(\underline{f}^{(k+1)}) - \mathcal{L}_u^{1,m}(\underline{f}^{(k)}) - \mathcal{L}_u^{2,m}(\underline{f}^{(k)}) = 0$, which is the first step of the DeC scheme, we obtain the property (5.8a) for $(k+1)$ and any m.

To prove the second property (5.8b) for $(k+1)$, we have to expand similarly the second step of the IMEX DeC scheme (5.5b). We start again from $\mathcal{L}_\sigma^{1,m}(\underline{f}^{(k+1)})$. We can collect already the unknown $f_\sigma^{m,(k+1)}$ and see what is a $\mathcal{O}(\varepsilon)$,

$$\mathcal{L}_\sigma^{1,m,(k)}(\underline{f}^{(k+1)}) = \left(1 + \frac{\Delta t^m}{\varepsilon}\right) \left(f_\sigma^{m,(k+1)} - \frac{\Delta t^m}{\Delta t^m + \varepsilon} \mathcal{M}(Pf_\sigma^{m,(k+1)}) + \mathcal{O}(\varepsilon)\right). \tag{5.10a}$$

The second term must be multiplied by the inverse of $1 + \frac{\Delta t^m}{\varepsilon}$, which is $\frac{\varepsilon}{\varepsilon + \Delta t^m}$. Thanks to this factor, we consider only terms with an ε at the denominator. So, we write

$$\frac{\varepsilon}{\varepsilon + \Delta t^m} \left(\mathcal{L}_\sigma^{1,m}(\underline{f}^{(k)}) - \mathcal{L}_\sigma^{2,m}(\underline{f}^{(k)})\right) \tag{5.10b}$$

$$= f_\sigma^{m,(k)} - \mathcal{M}(Pf_\sigma^{m,(k)}) - \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \int_{\mathrm{E}} \frac{\varphi_\sigma}{|\mathrm{E}_\sigma|} \left(f^{m,(k)} - \sum_{r=0}^{M} \theta_r^m \mathcal{M}(Pf^{r,(k)})\right) dx + \mathcal{O}(\varepsilon) \tag{5.10c}$$

$$= \mathcal{O}(\Delta) + \mathcal{O}(\varepsilon).$$

Again, the last step is just due to the mass lumping and the time integration. There we get an extra $\mathcal{O}(\Delta)$. If we sum the terms together and solve the scheme $\mathcal{L}_\sigma^{1,m}(\underline{f}^{(k+1)}) - \mathcal{L}_\sigma^{1,m}(\underline{f}^{(k)}) + \mathcal{L}_\sigma^{2,m}(\underline{f}^{(k)}) = 0$, we obtain the second property (5.8b) of the induction step. Hence, we proved the theorem. ∎

Summarizing, the proposed IMEX DeC scheme is an AP scheme that can solve with high order accuracy kinetic models in the form (2.2). In this section we proved that the scheme is AP and, thus, can resolve the small scales of ε without refining the

discretization scales. The proof of the high order accuracy of the scheme is given in Appendix B.

Remark 5.3 (comparison with high order IMEX schemes). As pointed out in Remark 4.3, with respect to higher RK IMEX schemes, our scheme is mass matrix free and the weights of time integration are automatically defined by the polynomial choice.

One can also think of combining a high order RK IMEX procedure with the DeC algorithm, as done in [11]. In this case, we face the same problems presented above. Anyway, this approach should lead to an increase of the order convergence in each correction step of the DeC procedure. Namely, if we use an IMEX RK2 scheme as $\mathcal{L}^1$ formulation, we will get 2 orders of accuracy more at each DeC corrections. Overall, there is no improvement in the computational costs between IMEX RK DeC and an IMEX DeC. Moreover, it has been shown in [13] that this approach leads also to some problems of smoothness of the error behavior and consequently in a drop in the order accuracy.

6. Numerical simulations. In this section, we validate the theoretical results through some numerical tests. We will focus on scalar equations and Euler systems of equations as macroscopic model, in both one and two dimensions. In all the simulations, we will introduce the macroscopic equation (2.1) and we will run the simulation on the related kinetic model generated by (2.2). In all the tests we will use the presented IMEX DeC scheme.

Some parameters must be chosen in each simulation. In particular, the relaxation parameter ε will be chosen according to what we are interested in. Most of the time we want to check the macroscopic limit, so we will choose $\varepsilon \ll \Delta t$. As imposed by the Whitham's subcharacteristic conditions (2.20), we have to choose the convection parameter bigger than the spectral radius of the macroscopic Jacobian of the flux, i.e., $\lambda > \rho(JA(u)), \forall\ u$ in the domain of interest.

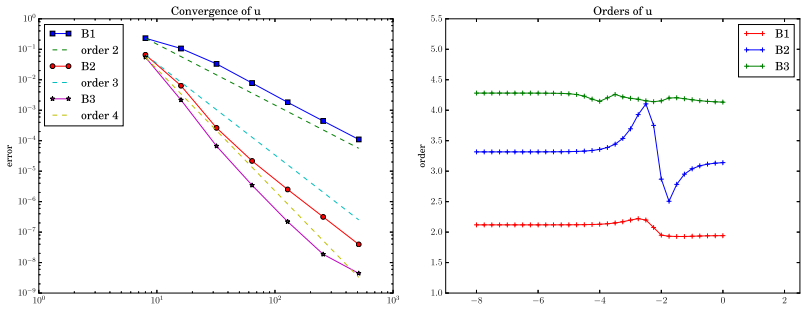
In the nodal residual definitions, more parameters play a role in order to stabilize the solution. We will make use of different schemes presented in [6] and reported in Appendix A. In particular, we will specify the choice of the coefficients θ_i of the penalty terms for the jump of the derivatives on the boundaries; see Appendix A.

6.1. 1D numerical tests.

6.1.1. Convergence for linear transport equation. To start, we test the IMEX DeC scheme with the scalar linear equation $u_t + u_x = 0$ as a macroscopic equation; see Example 2.1. The nodal distribution that we will use for smooth test cases is a Galerkin approximation stabilized by jump penalty terms proposed by Burman and Hansbo [12]. The scheme is defined in Appendix A in (A.3). The initial conditions are $u_0(x) = e^{-80(\sin(\pi(x-0.4))/\pi)^2}$ and $f_0 = \mathcal{M}(u_0)$. All the other parameters are in Figure 6.1(c). The number of subtimesteps M is the same as the degree of the polynomials in $\mathbb{B}^p$ and the corrections are $K = p + 1 = M + 1$.

As we can see in Figure 6.1(a), the convergence of the scheme is the theoretical one.

In Figure 6.1(b) we test the scheme varying the relaxation parameter ε. The order of accuracy is the expected one. There are slight oscillations in particular for $\mathbb{B}^2$ solutions. This is a well-known problem of order reduction as ε is approaching the magnitude of Δ, which affects several schemes, including some RK methods, as stated in [11]. Anyway, we can say that the scheme is getting an order of accuracy bigger than or equal to the expected one, except for few midrange values of ε. Moreover, this proves stability, for any value of ε.

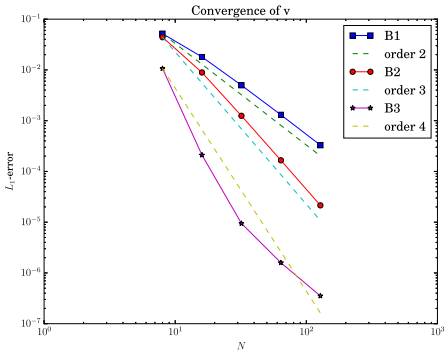


(a) Scalar 1D convergence (b) Varying relaxation parameter ε

Ω	T	λ	ε	CFL	BC		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
$[0,1]$	0.12	1.5	10^{-9}	0.1	periodic	θ_1	1	1	1	θ_2	0	0	5

(c) Parameters for transport tests

FIG. 6.1. *Scalar linear* 1*D test.*



(a) 1D Euler convergence

Ω	T	λ	ε	CFL	BC		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
$[-1,1]$	0.1	3	10^{-9}	0.2	periodic	θ_1	1	1	1	θ_2	0	0	5

(b) Parameters for isentropic Euler 1D

FIG. 6.2. *Convergence on Euler equations in one dimension.*

6.1.2. Euler equation—isentropic flow. Now, we solve the Euler equations

$$\big(\rho, \rho v, E\big)_t + \big(\rho v, \rho v^2 + p, (E+p)v\big)_x = 0, \tag{6.1}$$

$$p = (E - 0.5\rho v^2)(\gamma - 1), \tag{6.2}$$

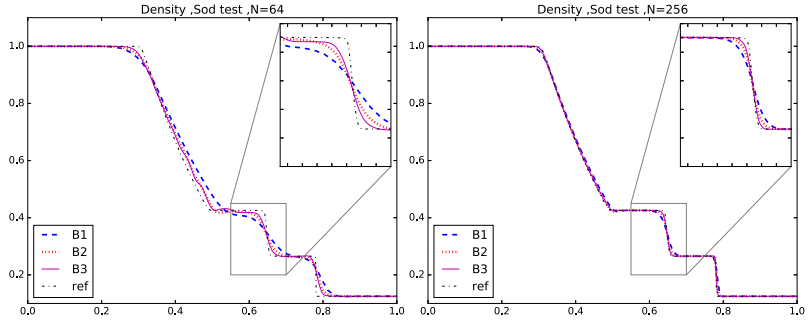
where ρ is the density, v the speed, p the pressure, and E the total energy. The quantities are linked by the EOS (6.2). To test the convergence of the scheme on 1D Euler equations, we use the case of isentropic flow, when $\gamma = 3$ and $p = \rho^\gamma$, with initial conditions $\big(\rho_0, v_0, p_0\big) = \big(1 + 0.5 \cdot \sin(\pi x), 0, \rho_0^\gamma\big)$. The parameters used for the scheme are in Figure 6.2(b). As we can see in Figure 6.2(a), the order of convergence is what we expected.

6.1.3. Euler equation—Sod shock test. Now we test the IMEX DeC scheme on not-smooth solutions. We begin with the Euler Sod test case. The Sod test case is solving (6.1) on domain $[0,1]$, with EOS (6.2), where $\gamma = 1.4$. The initial

conditions are $(\rho_0, v_0, p_0) = (1, 0, 1)$ for $x \leq 0.5$ and $(\rho_0, v_0, p_0) = (0.125, 0, 0.1)$ for $x > 0.5$. The nodal residual definition in this nonsmooth test case is in Appendix A in (A.8), where a convex combination between a Rusanov scheme and a limitation of it is applied, as described by Abgrall, Bacigaloppi, and Tokareva [6]. In Figure 6.3, we show the parameters used in the scheme and the density plots for different mesh sizes $N = 64{,}256$. As we notice, even with few points the $\mathbb{B}^3$ solution is outperforming the other solutions, catching in a better way the edges of the discontinuities.

6.1.4. Euler equation—Woodward–Colella. We observe even more advantages of using a high order scheme in the following examples. First, we present the one proposed by Colella and Woodward [14]. It solves again Euler equation (6.1) with EOS (6.2) with $\gamma = 1.4$. The initial conditions are $\rho_0 = 1$, $v_0 = 0$, $p_0 = 10^3 \mathbb{1}_{[0,0.1]} + 10^{-2} \mathbb{1}_{[0.1,0.9]} + 10^2 \mathbb{1}_{[0.9,1]}$. We used again scheme (A.8) for this nonsmooth problem, with the parameters in Figure 6.4.

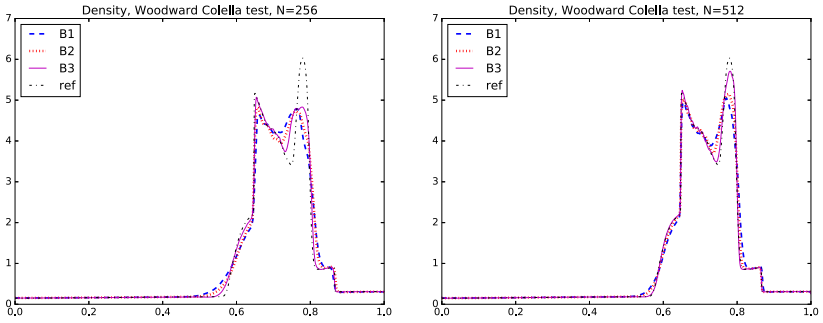
We observe that in this case, only $\mathbb{B}^3$ is able to catch the shape of the second peak (with 512 elements).



(a) $N = 64$ (b) $N = 256$

Ω	T	λ	ε	CFL	BC		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
$[0,1]$	0.16	2	10^{-9}	0.2	outflow	θ_1	1	1	2.5	θ_2	0	0.5	4

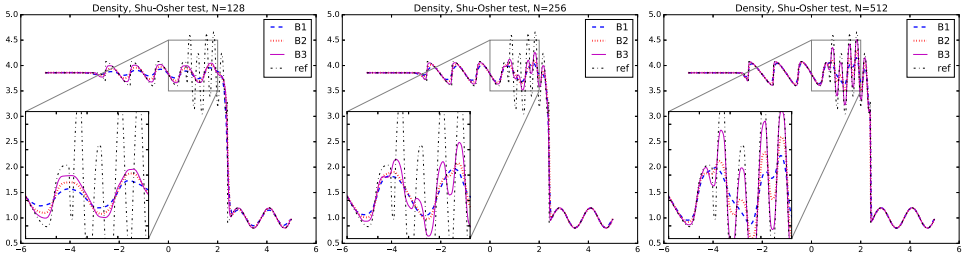
FIG. 6.3. *Density of* 1*D Sod test case.*



(a) $N = 256$ (b) $N = 512$

Ω	T	λ	ε	CFL	BC		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
$[0,1]$	0.038	20	10^{-9}	0.1	outflow	θ_1	0.5	0.8	5	θ_2	0	1	1

FIG. 6.4. *Density of Woodward–Colella test.*



(a) $N = 128$ (b) $N = 256$ (c) $N = 512$

Ω	T	λ	ε	CFL	BC		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
$[-5,5]$	1.8	3	10^{-9}	0.1	outflow	θ_1	0.5	0.8	3	θ_2	0	1	1

FIG. 6.5. *Density of Shu–Osher test.*

6.1.5. Euler equation—Shu–Osher test. The last test we performed in one dimension was proposed by Shu and Osher [25]. Again we have Euler equation (6.1) with EOS (6.2) with $\gamma = 1.4$. The initial conditions are

$$\begin{pmatrix}\rho_0\\ v_0\\ p_0\end{pmatrix} = \begin{pmatrix}3.857143\\ 2.629369\\ 10.333333\end{pmatrix} \text{ if } x \in [-5,-4], \quad \begin{pmatrix}\rho_0\\ v_0\\ p_0\end{pmatrix} = \begin{pmatrix}1+0.2\sin(5x)\\ 0\\ 1\end{pmatrix} \text{ if } x \in [-4,5].$$

As before, the scheme used is defined in (A.8). In Figure 6.5, we can see results for several Ns. Even here, the second and third order polynomials outperform the first order one. In particular, the oscillations are already captured with few points and the precision increases quickly if the order is high.

In all the tests performed, our method captures the correct behavior of the solutions. Moreover, it is convenient to choose high order approximations to get a faster convergence to the exact solution.

6.2. 2D numerical tests. Finally, we test the IMEX DeC scheme on some 2D tests. Again, we will present the macroscopic equations, but we will solve the kinetic model (2.2). The system of equations we are going to solve is the 2D Euler equations:

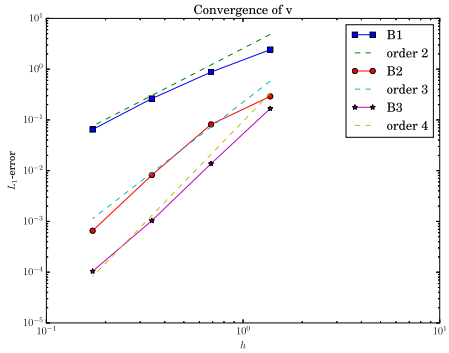
(6.3)
$$\partial_t U(\mathbf{x},t) + \partial_x A_1(U(\mathbf{x},t)) + \partial_y A_2(U(\mathbf{x},t)) = 0, \quad \mathbf{x} \in \Omega \subset \mathbb{R}^2, \quad U = \big(\rho, \rho u, \rho v, E\big),$$
$$A_1(U) = \big(\rho u, \rho u^2 + p, \rho uv, u(E+p)\big), \qquad A_2(U) = \big(\rho v, \rho uv, \rho v^2 + p, v(E+p)\big),$$

(6.4)
$$p = (\gamma - 1)\Big(E - 0.5\rho(u^2+v^2)\Big),$$

where ρ is the density, u the speed in x direction, v the speed in y direction, E the total energy, and p the pressure. A closure law is given by the EOS (6.4).

6.2.1. Euler equation—smooth vortex test case. To start, we want to study the convergence of the method also in two dimensions. To do so, we test our scheme with a steady vortex test case, so that we can compare the final solution with the initial one. The domain is a circle of radius 10 and center $(0,0)$. The exact conditions are imposed on the boundary.

To define the initial conditions, let us introduce the radius $r^2 := x^2 + y^2$, the coefficient $C(r) := e^{\frac{-r_0}{r_0^2 - r^2}} \mathbb{1}_{\{r<r0\}}$, where $r_0 := 5$ is the radius of the circle where



(a) 2D Euler convergence

Ω	T	λ	ε	CFL	BC		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
$\mathscr{B}_{10}$	1	3	10^{-9}	0.1	Dirichlet	θ_1	0.1	0.01	0.03

(b) Parameters steady vortex 2D

FIG. 6.6. *Convergence on Euler equations in two dimensions.*

the solution is not constant and $\beta := 5$. The modulus of the speed is defined as $|\underline{v}| := 2\beta C(r)\frac{r_0}{r_0^2 - r^2}$. The initial conditions and solutions for all times are

$$\left(\rho_0, u_0, v_0, p_0\right) = \left(\left(1 - \tfrac{\gamma-1}{\gamma}\beta^2 C(r)^2\right)^{\frac{1}{\gamma-1}}, (-y)|\underline{v}|, (x)|\underline{v}|, \rho_0^\gamma\right).$$

In our simulations $\gamma = 1.4$ for the EOS (6.4). The scheme used is (A.3) and the parameters chosen are in Figure 6.6(b). We use different refinements of the domain mesh. These are uniform triangular meshes and on the x-axis of Figure 6.6(a) one can see the maximum diameter of a cell of the mesh. As in 1D cases, in Figure 6.6(a) the convergence is reflecting the theoretical results running with the number of corrections $K = d + 1$ and subtimesteps.

6.2.2. Euler equation—Sod 2D test case. We tested the IMEX DeC method on the analogue of the Sod test in two dimensions. This test is again solving Euler equation (6.3) where $\gamma = 1.4$ in EOS (6.4). The domain Ω is a circle of radius $r = 1$ and center in $(0, 0)$. The initial conditions are $\left(\rho_0, u_0, v_0, p_0\right) = \left(1, 0, 0, 1\right)$ if $r < 0.5$ and $\left(\rho_0, u_0, v_0, p_0\right) = \left(0.125, 0, 0, 0.1\right)$ if $r \geq 0.5$.

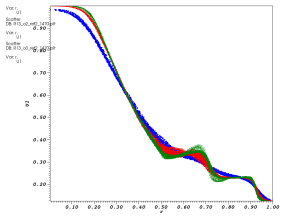
The parameters used for this test are in Figure 6.7(a). We use uniform triangular meshes and what is shown in Figures 6.7(d) to 6.7(f) is obtained with $N = 13548$ triangles on the domain. Figure 6.7(b) shows the scatter plot of the points of the density. The scheme used for this test case is given by the nodal residuals (A.8).

Comparing Figures 6.7(b) to 6.7(f), we observe that with higher order schemes we are able to better catch the sharpness of the shock moving on the domain. The mesh is chosen without particular attention to the geometry; nevertheless, in Figure 6.7(b), the points for the same values of the radius are not spread too much from one to another.

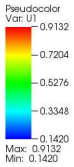
6.2.3. Euler equation—DMR 2D test case. In the end, we test our scheme on the double Mach reflection (DMR) problem presented in [17]. It consists of Euler equation (6.3) with $\gamma = 1.4$ in EOS (6.4). The domain is the rectangular shape $[-0.2, 3] \times [0, 2.2]$, cut on the bottom right part by an oblique edge passing through

T	0.25		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
λ	1.4	θ_1	0.1	0.1	0.01
ε	10^{-9}	θ_2	0	10^{-4}	10^{-4}
CFL	0.1	BC	outflow	Ω	$\mathscr{B}_1$

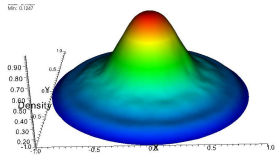
(a) parameters



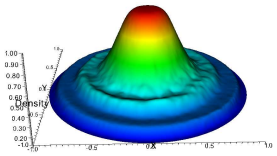
(b) Scatter of density of Sod test ($\mathbb{B}^1$ blue, $\mathbb{B}^2$ red and $\mathbb{B}^3$ green)



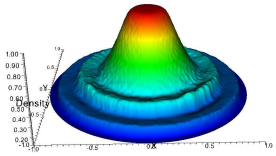
(c) scale



(d) $\mathbb{B}^1, N = 13548$



(e) $\mathbb{B}^2, N = 13548$



(f) $\mathbb{B}^3, N = 13548$

FIG. 6.7. *Density of Sod test.*

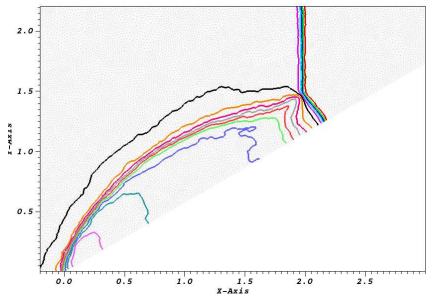
$(0,0)$ and $(3,1.7)$. We have wall boundary conditions on the bottom, on the top, and on the oblique edge of the mesh, inflow on the left edge, and outflow on the right one. The initial conditions have a discontinuity on $x = 0$. This shock has an initial speed in the right direction and, as time passes, the shock crosses the oblique surface and creates more internal shock surfaces. The initial conditions are $\left(\rho_0, u_0, v_0, p_0\right) = \left(8, 8.25, 0, 116.5\right)$ if $x \leq 0$ and $\left(\rho_0, u_0, v_0, p_0\right) = \left(1.4, 0, 0, 1\right)$ if $x > 0$.

The parameters used for scheme (A.8) are in Figure 6.8. The mesh we used is composed of $N = 19248$ triangular elements with a maximum diameter of 0.0369.

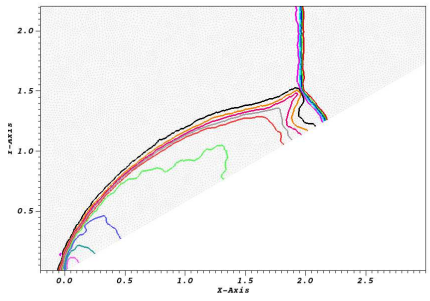
Again we can see in Figure 6.8 that the scheme catches the behavior of the shock and its reflection against the lower wall. Again, the sharpness of the shock is really well captured by the $\mathbb{B}^3$ scheme, while the others are less precise in defining the shock zone.

7. Conclusions and further investigations. We have presented a high order scheme for kinetic models of hyperbolic system of equations. The method proposed solves the stiffness of the relaxation term through an IMEX formulation (implicit for source term and explicit for advection term). Nevertheless, we were able to solve computationally explicitly the system, thanks to the structure of the model [9] and an auxiliary equation, which allows us not to recur to nonlinear solvers. The high order accuracy of the scheme is reached thanks to two ingredients. The first ingredient is the RD framework for spatial discretization [3], which is a FEM based method that is naturally high order because of the choice of different basis functions. The second is the high order time integration performed in the DeC method, which allows us to couple two operators, an IMEX easy-to-solve scheme and a high order time discretization RD scheme. The result is an iterative method able to reach high order accuracy and stability via few iterations. This is the first time, as far as we know, that the RD framework is used to solve hyperbolic systems with stiff source terms. Even if in this work we solved only one model, the extension to other models with similar properties should be straightforward and will be the focus of future research.

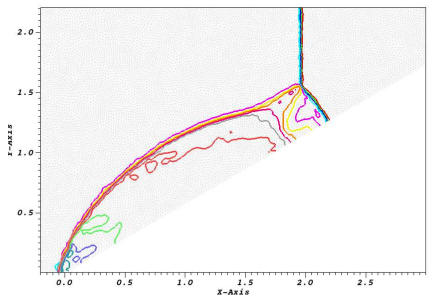
The results obtained both from a theoretical point of view and from the simulation side are satisfactory. Indeed, the theorems proved the AP property for our



(a) $\mathbb{B}^1$



(b) $\mathbb{B}^2$



(c) $\mathbb{B}^3$

T	0.2		$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
λ	15	θ_1	0.1	0.01	0.005
ε	10^{-9}	θ_2	0	10^{-4}	10^{-4}
CFL	0.1				

(d) Parameters

FIG. 6.8. *Density of DMR test* $\mathbb{B}^1, \mathbb{B}^2, \mathbb{B}^3$.

scheme and the rate of accuracy. In addition, the run simulations are reaching the expected accuracy in one and two dimensions, the correct behavior of the discontinuities of the solutions is well caught by the scheme, and, as the order increases, we see improvements in the prediction of the solutions.

Further investigations may be in the following directions. There are still some open questions over the complete automation of the scheme. For example, it is still not well known which relation occurs between parameters θ_1, θ_2, CFL, and the quality of the solution. This is a common problem with other works, such as [6]. There are

studies for 1D smooth solutions, where some relations between these quantities are shown, thanks to some von Neumann stability analysis [8, 26]. Nonetheless, these results are not easily extensible to nonlinear flux problems or 2D problems.

Finally, we are already working on some extensions of the scheme for multiphase flows equations and we believe that it can be applied also for a large variety of other problems, such as other BGK equations, viscoelasticity problems, or many other kinetic schemes.

Appendix A. Residual distribution schemes. The definition of an RD scheme (4.6) relies on stable and accurate definition of the nodal residuals. This should be done maintaining the conservation law (4.5). Many well-known schemes can be rewritten in this formulation [5]; for example, the SUPG scheme [19] is defined by

(A.1)
$$\phi_\sigma^K(f) = \int_K \varphi_\sigma \left(\sum_{d=1}^D \partial_{x_d} \Lambda_d f - S(f) \right) + h_K \int_K \left(\sum_{d=1}^D \partial_{x_d} \Lambda_d f \partial_{x_d} \varphi_\sigma \right) \tau \left(\sum_{d=1}^D \partial_{x_d} \Lambda_d f \cdot \partial_{x_d} f \right).$$

We use two types of scheme for the tests. One is suited for smooth solutions and it adds only a bit of artificial dissipation through some penalty terms. The second one is more robust and can deal with discontinuous solutions using a more elaborate limiter.

A.1. Smooth solutions residuals. When we are dealing with smooth tests and we know a priori that we do not need the extra diffusion to dump oscillations brought by discontinuities, we can use a pure Galerkin discretization with a stabilization term that penalizes the jump of the gradient (or higher derivatives) of the solution across cells edges [12, 4].

For the hyperbolic system (4.1), the scheme proceeds as follows $\forall \sigma \in \Sigma$:

$$\phi_\sigma^{\mathrm{E},1}(f) = \int_{\mathrm{E}} \varphi_\sigma \left(\sum_{d=1}^D \partial_{x_d} \Lambda_d f - S(f) \right) d\mathbf{x}, \tag{A.2}$$

$$\phi_\sigma^{\mathrm{E}}(f) = \phi_\sigma^{\mathrm{E},1}(f) + \sum_{z=1}^p \sum_{e|\text{edge of E}} \theta_z h_e^{2z} \int_e [\nabla^z f] \cdot [\nabla^z \varphi_\sigma] d\Gamma. \tag{A.3}$$

Here p is the degree of the polynomial of the basis functions we use, θ_z are positive coefficients, with the same physical dimension of a speed, and $[\cdot]$ is the jump across the edge e, namely, if e separates E and E^+, $[f] = f|_{\mathrm{E}} - f|_{\mathrm{E}^+}$. By the symbol ∇ in (A.3) we mean the derivative in the direction of the normal to the edge e, and h_e is the length of a 1D element of the mesh (the edge e in two dimensions, the size of a cell $|\mathrm{E}|$ in one dimension). The schemes just presented are naturally of order $p+1$. The parameters θ_p must be chosen carefully if we want the scheme to be stable. The stability analysis of this scheme in [26, 8] suggests some optimal values for these parameters in case of 1D linear fluxes, where the relations $\theta_1 \mathrm{CFL} \le C_1$ and $\theta_1 \ge C_2$ CFL must hold. The two coefficients C_1 and C_2 are hard to determine even for simple linear 1D scalar test cases. So, in our experiments we perform a hyperanalysis on these parameters for small times and we choose the one that better performs for a specific degree of polynomials.

A.2. Shock solutions residuals. If, a priori, we know that the solution of the test presents discontinuities, we use the following scheme. More details on the choice

of these schemes can be found in [6]. The procedure starts defining a local Galerkin Lax–Friedrichs type nodal residual on the steady part of original equation (4.1):

$$\phi_\sigma^{\mathrm{E},LxF}(f) := \int_{\mathrm{E}} \varphi_\sigma \left(\sum_{d=1}^{D} \partial_{x_d} \Lambda_d f - S(f) \right) d\mathbf{x} + \alpha_{\mathrm{E}}(f_\sigma - \overline{f}^{\mathrm{E}}), \tag{A.4}$$

$$\alpha_{\mathrm{E}} := \max_{\sigma \in \mathrm{E}} \max_{d} \left(\rho_S \left(\Lambda_d \right) \right) = \lambda, \tag{A.5}$$

where $\overline{f}^{\mathrm{E}}$ is the average of f over the cell E and α_{E} is the maximum eigenvalue of the Jacobian of the fluxes and ρ_S is the function returning the spectral radius of the input matrix. Then, to guarantee monotonicity of the solution near strong discontinuities, we proceed as follows:

$$\beta_\sigma^{\mathrm{E}}(f) := \max\left(\frac{\Phi_\sigma^{\mathrm{E},LxF}}{\Phi^{\mathrm{E}}}, 0 \right) \left(\sum_{j \in \mathrm{E}} \max\left(\frac{\Phi_j^{\mathrm{E},LxF}}{\Phi^{\mathrm{E}}}, 0 \right) \right)^{-1}, \quad \phi_\sigma^{*,\mathrm{E}} := \beta_\sigma^{\mathrm{E}} \phi^{\mathrm{E}}. \tag{A.6}$$

The divisions between vectors are meant componentwise. Then, we apply a convex combination between the new residual and the Lax–Friedrichs one, where the blending coefficient is Θ,

$$\Theta := \frac{|\Phi^{\mathrm{E}}|}{\sum_{j \in \mathrm{E}} |\Phi_j^{\mathrm{E},LxF}|}, \quad \phi_\sigma^{\cdot,\mathrm{E}} := (1-\Theta)\phi_\sigma^{*,\mathrm{E}} + \Theta \Phi_\sigma^{\mathrm{E},LxF}. \tag{A.7}$$

This scheme guarantees the monotonicity principle [3]. After that, to define the final scheme, we add to the scheme the jump stabilization terms

$$\phi_\sigma^{\mathrm{E}} := \phi_\sigma^{\cdot,\mathrm{E}} + \sum_{z=1}^{p} \sum_{e | \text{edge of E}} \theta_z h_e^{2z} \int_e [\nabla^z f] \cdot [\nabla^z \varphi_\sigma] d\Gamma. \tag{A.8}$$

Appendix B. Deferred correction properties.

B.1. Lipschitz continuity and coercivity. We now prove that the operators $\mathcal{L}^1$ (5.3) and $\mathcal{L}^2$ (4.8) verify all the hypothesis of Proposition 4.1.

PROPOSITION B.1. *$\mathcal{L}^1$ is coercive, i.e., $\exists \alpha_1 > 0$ s.t. $\forall \underline{f}, \underline{g} \in V_h^M$ and $m = 1, \ldots, M$, i.e.,*

$$\left\| \mathcal{L}_u^{1,m}(\underline{f}) - \mathcal{L}_u^{1,m}(\underline{g}) \right\| \geq \alpha_1 ||P\underline{f} - P\underline{g}||, \tag{B.1}$$

$$\left\| \mathcal{L}^{1,m}(\underline{f}) - \mathcal{L}^{1,m}(\underline{g}) \right\| \geq \alpha_1 ||\underline{f} - \underline{g}||. \tag{B.2}$$

Proof. We suppose that the initial states coincide for $\underline{f}$ and $\underline{g}$, i.e., $f^0 = g^0$, from the previous timestep. Then, (B.1) is trivial because

$$\mathcal{L}_{\sigma,u}^{1,m}(\underline{f}) - \mathcal{L}_{\sigma,u}^{1,m}(\underline{g}) = P(f_\sigma^m - g_\sigma^m), \tag{B.3}$$

which leads immediately to (B.1). For (B.2) we have to collect the implicit terms as done in (5.4b). Then, we can write

$$\mathcal{L}_\sigma^{1,m}(\underline{f}) - \mathcal{L}_\sigma^{1,m}(\underline{g}) = (f_\sigma^m - g_\sigma^m) - \frac{\Delta t}{\Delta t + \varepsilon}(\mathcal{M}(Pf_\sigma^m) - \mathcal{M}(Pg_\sigma^m)) = f_\sigma^m - g_\sigma^m. \tag{B.4}$$

The last step is possible since the Maxwellians in our scheme are computed from the *auxiliary* equation and they are actually explicitly computed, so they must coincide, since $f^0 = g^0$. If we write the operator explicitly for both Pf and f, we can see that the coercivity constant is $\alpha_1 = 1$, given any norm. ∎

Before proving the Lipschitz continuity, we define the norm $||\cdot||$ for a function $f \in V_h$, which is consistent with the $\mathcal{L}^2$ norm, and the norm $|||\cdot|||$ of all the subtimesteps defined as

$$||f||^2 := \sum_{\sigma \in D_h} |\mathrm{E}_\sigma| f_\sigma^2, \quad |||\underline{f}|||^2 = ||| \left(f^0, \dots, f^M\right) |||^2 := \frac{1}{M} \sum_{m=0}^{M} ||f^m||^2. \tag{B.5}$$

Moreover, we will need the definition of the following seminorms:

$$|f|_{1,x}^2 := \sum_{\sigma \in D_h} |\mathrm{E}_\sigma| \left(\max_{\mathrm{E}|\sigma \in \mathrm{E}} \max_{x \in \mathrm{E}} \frac{f_\sigma - f(x)}{d(\mathrm{E})} \right)^2, \tag{B.6}$$

$$|\underline{f}|_{1,t}^2 := \sum_{\sigma \in D_h} |\mathrm{E}_\sigma| \left(\max_{m=1,\dots,M} \frac{f^m - f^{m-1}}{\Delta t^m} \right)^2, \tag{B.7}$$

where $d(\mathrm{E})$ is the diameter of the cell E and it is bounded by $\max_{\mathrm{E}} d(\mathrm{E}) = h$. In particular, we note that $|f|_{1,x} \le |f|_1 = ||\nabla f||_{L^2}$ for every discretization mesh.

Proposition B.2. *Assume some regularity on the solutions, more precisely,*

$$|f|_{1,x} \le C_1 ||f||, \tag{B.8}$$

$$|\underline{f}|_{1,t} \le C_2 |||\underline{f}|||, \tag{B.9}$$

where C_1 and C_2 do not depend on the mesh size h and timestep Δt. Moreover, we require that nodal residuals verify

$$\sum_{\sigma \in D_h} \frac{1}{|\mathrm{E}_\sigma|} \left(\sum_{\mathrm{E}|\sigma \in \mathrm{E}} \phi_\sigma^{\mathrm{E}}(f) - \phi_\sigma^{\mathrm{E}}(g) \right)^2 \le C_3 \sum_{\sigma \in D_h} |\mathrm{E}_\sigma| (f_\sigma - g_\sigma)^2 = C_3 ||f - g||^2; \tag{B.10}$$

then, $\mathcal{L}^1 - \mathcal{L}^2$ is Lipschitz continuous, i.e., $\exists \alpha_2 > 0$ s.t. $\forall \underline{f}, \underline{g} \in V_h^M$

$$||| \left(\mathcal{L}_u^1(\underline{f}) - \mathcal{L}_u^1(\underline{g})\right) - \left(\mathcal{L}_u^2(\underline{f}) - \mathcal{L}_u^2(\underline{g})\right) ||| \le \alpha_2 \Delta |||P\underline{f} - P\underline{g}|||, \tag{B.11}$$

$$||| \left(\mathcal{L}^1(\underline{f}) - \mathcal{L}^1(\underline{g})\right) - \left(\mathcal{L}^2(\underline{f}) - \mathcal{L}^2(\underline{g})\right) ||| \le \alpha_2 \Delta |||\underline{f} - \underline{g}|||. \tag{B.12}$$

Remark B.3 (regularity of the solution). The extra hypotheses added are related to the regularity of the solution. Of course, when they are not satisfied, for example, when there are shocks in the solution, (B.8) does not hold. Anyway, we see numerically a big improvement in higher order solutions. Equation (B.10), in our case, is given by the consistency of the nodal residuals, the Lipschitz continuity of the flux F, and the regularity of the solutions f, g as stated in (B.8).

Proof. The estimation of (B.11) is a simplification of the case of (B.12), so we will skip its proof.

For simplicity, we introduce the differences $\delta f := f - g$, $\delta\phi_\sigma^K(f) := \phi_\sigma^K(f) - \phi_\sigma^K(g)$, $\delta\mathcal{M}(Pf) := \mathcal{M}(Pf) - \mathcal{M}(Pg)$, $\delta\mathcal{L} := \mathcal{L}^1 - \mathcal{L}^2$, and $\delta\mathcal{I}(\underline{f}) := \mathcal{I}_0(\underline{f}) - \mathcal{I}_{\mathcal{M}}(\underline{f})$.

Now, we split the operators into two parts. The first one is composed of the term related to the time derivative and the source term $\mathcal{L}_{ts}$, and the second one concerns the advection part $\mathcal{L}_{ad}$. If we write explicitly the source and time part, we get

(B.13)
$$\delta\mathcal{L}^m_{ts,\sigma}(\underline{f}) - \delta\mathcal{L}^m_{ts,\sigma}(\underline{g}) = \sum_{\mathrm{E}|\sigma\in\mathrm{E}} \frac{1}{|\mathrm{E}_\sigma|}\frac{\varepsilon}{\varepsilon+\Delta t^m}\Bigg[\int_{\mathrm{E}}\varphi_\sigma\left(\delta f^m_\sigma - \delta f^m\right)$$
$$-\frac{\Delta t^m}{\varepsilon}\int_{\mathrm{E}}\varphi_\sigma\left(\delta\mathcal{M}(Pf^m_\sigma)-\delta f^m_\sigma\right)+\frac{1}{\varepsilon+\Delta t^m}\int_{t^0}^{t^m}\mathcal{I}_M\left(\delta\phi^{\mathrm{E}}_{s,\sigma}(f^0),\dots,\delta\phi^{\mathrm{E}}_{s,\sigma}(f^M),s\right)ds\Bigg].$$

Supposing that the residuals are a consistent discretization of fluxes and source terms, we can use the Galerkin discretization instead of any other one. Moreover, we add and subtract the residual in timestep $t^{n,m}$, i.e., $\phi_{ts,\sigma}(\delta f^m)$. So, we can write, neglecting $\mathcal{O}(\Delta^2|||\underline{f}-\underline{g}|||)$,

(B.14a) $$\mathcal{L}^{1,m}_{ts,\sigma}(\underline{f}) - \mathcal{L}^{1,m}_{ts,\sigma}(\underline{g}) - \mathcal{L}^{2,m}_{ts,\sigma}(\underline{f}) + \mathcal{L}^{2,m}_{ts,\sigma}(\underline{g}) + \mathcal{O}(\Delta^2|||\underline{f}-\underline{g}|||)$$

(B.14b)
$$= \frac{1}{|\mathrm{E}_\sigma|}\int_\Omega\varphi_\sigma\left(\delta f^m_\sigma-\delta f^m\right) - \frac{1}{|\mathrm{E}_\sigma|}\frac{\Delta t^m}{(\varepsilon+\Delta t^m)}\int_\Omega\varphi_\sigma\left(\delta\mathcal{M}(Pf^m_\sigma)-\delta\mathcal{M}(Pf^m)\right)$$
$$+\frac{1}{\varepsilon+\Delta t^m}\int_{t^0}^{t^m}\mathcal{I}_M(\delta\phi_{s,\sigma}(f^0)-\delta\phi_{s,\sigma}(f^m),\dots,\delta\phi_{s,\sigma}(f^M)-\delta\phi_{s,\sigma}(f^m),s)ds.$$

Now, we sum over the degrees of freedom and we square the previous quantity. We use Lemma A.1 of [4] to pass from coefficients v_σ to pointwise evaluation $v(\sigma)$, with abuse of notation. It states that $\sum_{\sigma\in\mathrm{E}}|v_\sigma - v_{\sigma'}| \le C_{\mathrm{E}}\sum_{\sigma\in\mathrm{E}}|v(\sigma)-v(\sigma')|$ where C_{E} is the norm of the inverse of the matrix $(\varphi_\sigma(\sigma'))_{\sigma,\sigma'}$ and it depends on E only via the aspect ratio of the element E.

(B.15a)
$$\sum_{\sigma\in D_h}|\mathrm{E}_\sigma|\left(\mathcal{L}^{1,m}_{ts,\sigma}(\underline{f}) - \mathcal{L}^{1,m}_{ts,\sigma}(\underline{g}) - \mathcal{L}^{2,m}_{ts,\sigma}(\underline{f}) + \mathcal{L}^{2,m}_{ts,\sigma}(\underline{g})\right)^2$$
$$\le C_a h^2\sum_{\sigma\in D_h}\frac{1}{|\mathrm{E}_\sigma|}\left(\int_\Omega\varphi_\sigma\left(\frac{\delta f^m_\sigma-\delta f^m(x)}{d(\mathrm{E})}\right)\right)^2$$

(B.15b)
$$+C_b h^2\frac{\Delta t^m}{(\varepsilon+\Delta t^m)}\sum_{\sigma\in D_h}\frac{1}{|\mathrm{E}_\sigma|}\left(\int_\Omega\varphi_\sigma\frac{\delta\mathcal{M}(Pf^m)(\sigma)-\delta\mathcal{M}(Pf^m)}{d(\mathrm{E})}\right)^2$$
$$+C_c\frac{\Delta t^m}{\varepsilon+\Delta t^m}\sum_{\sigma\in D_h}|C_\sigma|\max_r\left(\delta\phi_{s,\sigma}(f^r)-\delta\phi_{s,\sigma}(f^m)\right)^2$$

(B.15c) $$\le C_d h^2(|\delta f^m|^2_{1,x} + |\delta\mathcal{M}(Pf^m)|^2_{1,x} + \max_r||\delta f^r-\delta f^m||^2)$$

(B.15d) $$\le C_e h^2(||\delta f^m||^2 + ||\delta\mathcal{M}(Pf^m)||^2 + \Delta t^2|\delta\underline{f}|^2_{1,t})$$

(B.15e) $$\le C_f h^2|||\delta\underline{f}|||^2 + \mathcal{O}(h^4) \le C_4 h^2|||\underline{f}-\underline{g}|||^2.$$

In (B.15b) we explicitly bring the scale h outside the first two sums, while in the third term we just bound the interpolant polynomial with the maximum of the interpolant values times a constant; in (B.15c) we use the definition of the seminorm (B.6), the Lipschitz continuity of residuals (B.10), the product rule for integrals, and the bound $\Delta t^m \le \Delta t^m + \varepsilon$. In (B.15d) we use the inequality (B.8) and the definition of the seminorm (B.7). In (B.15e) we use the fact that the Maxwellians $\mathcal{M}$ and the

projections P are Lipschitz continuous, the inequality (B.9), and the fact that $\Delta t \sim h$. The constant C_4 does not depend on $h, \Delta t$, or ε, but it depends on the size of the domain, on the Lipschitz continuity of the Maxwellians, on the regularity of the mesh, and on basis functions.

For the advection term a similar computation is carried out, but in this case the error is an $\mathcal{O}(\Delta t)$. Using the notation of $\phi_\sigma := \sum_{K|\sigma\in K} \phi_\sigma^K$, we write

$$||\mathcal{S}_x||^2 := \sum_{\sigma\in D_h} |\mathrm{E}_\sigma| \left(\delta\mathcal{L}^{1,m}_{ad,\sigma}(\underline{f}) - \delta\mathcal{L}^{1,m}_{ad,\sigma}(\underline{g})\right)^2 \tag{B.16a}$$

$$= \sum_{\sigma\in D_h} \frac{1}{|\mathrm{E}_\sigma|} \left(\frac{\varepsilon}{\varepsilon + \Delta t^m} \int_{t^{n,0}}^{t^{n,m}} \delta\mathcal{I}\left(\delta\phi_{ad,\sigma}(f^0), \dots, \delta\phi_{ad,\sigma}(f^M), s\right) ds\right)^2 \tag{B.16b}$$

$$\leq C_l \sum_{\sigma\in D_h} \frac{\Delta t^2}{|\mathrm{E}_\sigma|} \left(\sum_{\mathrm{E}|\sigma\in\mathrm{E}} \max_{m=1,\dots,M} \frac{|\delta\phi^{\mathrm{E}}_{ad,\sigma}(f^m) - \delta\phi^{\mathrm{E}}_{ad,\sigma}(f^{m-1})|}{\Delta t^m}\right)^2. \tag{B.16c}$$

In (B.16c) we use the bound $\varepsilon \leq \varepsilon + \Delta t^m$ and the fact that $\mathcal{I}_0$ is a zero order approximation of $\mathcal{I}_M$, so, adding the integration in time, we get the error estimation above.

$$||\mathcal{S}_x||^2 \leq C_q \sum_{\sigma\in D_h} \Delta t^2 |\mathrm{E}_\sigma| \left(\max_{m=1,\dots,M} \frac{|\delta f^m - \delta f^{m-1}|}{\Delta t^m}\right)^2 \tag{B.16d}$$

$$\leq C_p \Delta t^2 \sum_{m=1}^{M} |f^m - g^m|^2_{1,t} \leq C_5 \Delta t^2 |||\underline{f} - \underline{g}|||^2. \tag{B.16e}$$

In (B.16d) we use the Lipschitz continuity and consistency hypothesis on the residuals (B.10). Finally, in (B.16e) we use the definition of seminorm (B.7) and we apply the bound in (B.9). C_5 does not depend on Δt, h, or ε, but only on fluxes, geometry, and basis functions.

Summing up the inequalities (B.15e) and (B.16e), we prove the thesis of the proposition. ∎

Acknowledgments. We acknowledge Paola Bacigaluppi and Svetlana Tokareva for their contributions in coding and discussing the residual distribution formulation.

REFERENCES

[1] R. Abgrall, *Toward the ultimate conservative scheme: Following the quest*, J. Comput. Phys., 167 (2001), pp. 277–315, https://doi.org/10.1006/jcph.2000.6672.

[2] R. Abgrall, *Essentially non-oscillatory residual distribution schemes for hyperbolic problems*, J. Comput. Phys., 214 (2006), pp. 773–808, https://doi.org/10.1016/j.jcp.2005.10.034.

[3] R. Abgrall, *Residual distribution schemes: Current status and future trends*, Comput. Fluids, 35 (2006), pp. 641–669, https://doi.org/10.1016/j.compfluid.2005.01.007.

[4] R. Abgrall, *High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices*, J. Sci. Comput., 73 (2017), pp. 461–494, https://doi.org/10.1007/s10915-017-0498-4.

[5] R. Abgrall, *Some remarks about conservation for residual distribution schemes*, Comput. Methods Appl. Math., 18 (2018), pp. 327–351.

[6] R. Abgrall, P. Bacigaluppi, and S. Tokareva, *High-order residual distribution scheme for the time-dependent euler equations of fluid dynamics*, Comput. Math. Appl., 78 (2019), pp. 274–297, https://doi.org/10.1016/j.camwa.2018.05.009.

[7] R. Abgrall, A. Larat, and M. Ricchiuto, *Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid unstructured meshes*, J. Comput. Phys., 230 (2011), pp. 4103–4136, https://doi.org/10.1016/j.jcp.2010.07.035.

[8] R. Abgrall and D. Torlo, *Some Preliminary Results on a Kinetic Scheme that Has a Lattice Boltzmann Method Flavour*, arXiv:1904.12928, 2019.

[9] D. Aregba-Driollet and R. Natalini, *Discrete kinetic schemes for systems of conservation laws*, in Hyperbolic Problems: Theory, Numerics, Applications, Internat. Ser. Numer. Math. 129, Springer, New York, 1999, pp. 1–10, https://doi.org/10.1007/978-3-0348-8720-5_1.

[10] D. Aregba-Driollet and R. Natalini, *Discrete kinetic schemes for multidimensional systems of conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 1973–2004, https://doi.org/10.1137/S0036142998343075.

[11] S. Boscarino, J. Qiu, and G. Russo, *Implicit-explicit integral deferred correction methods for stiff problems*, SIAM J. Sci. Comput., 40 (2017), pp. A787–A816.

[12] E. Burman and P. Hansbo, *Edge stabilization for galerkin approximations of convection–diffusion–reaction problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1437–1453, https://doi.org/10.1016/j.cma.2003.12.032

[13] A. Christlieb, B. Ong, and J. Qiu, *Integral deferred correction methods constructed with high order runge-kutta integrators*, Math. of Comp., 79 (2010), pp. 761–783.

[14] P. Colella and P. R. Woodward, *The piecewise parabolic method (PPM) for gas-dynamical simulations*, J. Comput. Phys., 54 (1984), pp. 174–201, https://doi.org/10.1016/0021-9991(84)90143-8.

[15] H. Deconinck and M. Ricchiuto, *Residual Distribution Schemes: Foundations and Analysis*, John Wiley & Sons, New York, 2004, https://doi.org/10.1002/0470091355.ecm054.

[16] A. Dutt, L. Greengard, and V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT, 40 (2000), pp. 241–266, https://doi.org/10.1023/A:1022338906936.

[17] H. Glaz, P. Colella, I. I. Glass, and L. R. Deschambault, *A numerical study of oblique shock-wave reflections with experimental comparisons*, Proc. A, 398 (1985), pp. 117–140.

[18] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu, *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*, World Scientific, River Edge, NJ, 2011.

[19] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert, *A new finite element formulation for computational fluid dynamics:* VIII. *The Galerkin/least-squares method for advective-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 173–189, https://doi.org/10.1016/0045-7825(89)90111-4.

[20] S. Jin and P. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.

[21] M. L. Minion, *Semi-implicit spectral deferred correction methods for ordinary differential equations*, Commun. Math. Sci., 1 (2003), pp. 471–500, https://projecteuclid.org:443/euclid.cms/1250880097.

[22] P. Öffner and D. Torlo, *Arbitrary High-Order, Conservative and Positive Preserving Patankar-Type Deferred Correction Schemes*, arXiv:1905.09237, 2019.

[23] L. Pareschi and G. Russo, *Implicit–explicit runge–kutta schemes and applications to hyperbolic systems with relaxation*, J. Sci. Comput., 25 (2005), pp. 129–155, https://doi.org/10.1007/s10915-004-4636-4.

[24] M. Ricchiuto and R. Abgrall, *Explicit Runge-Kutta residual distribution schemes for time dependent problems: Second order case*, J. Comput. Phys., 229 (2010), pp. 5653–5691, https://doi.org/10.1016/j.jcp.2010.04.002.

[25] C. W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471, https://doi.org/10.1016/0021-9991(88)90177-5.

[26] D. Torlo, *Hyperbolic Problems: High Order Methods and Model Order Reduction*, Ph.D. Dissertation, University of Zurich, 2020.