

A New Asymptotic-Preserving Dual Formulation Finite-Volume Method for the Compressible Euler Equations

Alina Chertock*, Smadar Karni†, Alexander Kurganov‡ and Lorenzo Micalizzi§

Abstract

The paper focuses on the development of numerical methods for the compressible Euler equations. It is well-known that if the Mach number is small, the system becomes stiff and hence explicit schemes suffer from severe time-step restrictions, making them inefficient or even impractical. Our objective is to develop an asymptotic preserving (AP) scheme that remains uniformly accurate and stable across all Mach numbers.

Instead of the conservative hyperbolic flux splitting approach, which is widely used to design AP schemes, we consider a primitive (nonconservative) formulation and introduce a nonconservative hyperbolic splitting. The resulting system is discretized using a semi-implicit approach: the stiff part is handled semi-implicitly using second-order central differences, while the nonstiff part is treated explicitly using a second-order path-conservative central-upwind discretization. A key feature of our method is that the pressure at each time level is computed by solving a well-posed Poisson-type elliptic equation, thereby enforcing the AP property. Simultaneously, we evolve the conservative form of the system using a semi-discrete central-upwind (CU) scheme. At the end of each stage of the time discretization, we perform a special post-processing that selects the appropriate numerical solution depending on the Mach number. This guarantees that in low-Mach-number regimes, the solution is obtained by the AP nonconservative scheme, while in higher-Mach-number regimes, a sharp and physically relevant solution is computed by the conservative CU scheme.

Numerical experiments confirm that the proposed AP scheme achieves the expected second order of accuracy and that the time-step constraint is independent of the Mach number, making it a robust and efficient alternative to conventional explicit methods.

Key words: Compressible Euler equations; low Mach number; asymptotic preserving (AP) schemes; hyperbolic splitting; semi-implicit methods; deferred correction.

AMS subject classification: 76M12, 65M08, 65L04, 35B40, 35L60.

*Department of Mathematics, North Carolina State University, Raleigh, NC 27695, USA; chertock@math.ncsu.edu

†Department of Mathematics, University of Michigan, 48109, USA; karni@umich.edu

‡Department of Mathematics and Shenzhen International Center for Mathematics, Southern University of Science and Technology, Shenzhen, 518055, China; alexander@sustech.edu.cn

§Department of Mathematics, North Carolina State University, Raleigh, NC 27695, USA; lmicali@ncsu.edu

1 Introduction

The paper focuses on the compressible Euler equations, which, like any other hyperbolic system of PDEs, are characterized by a finite speed of propagation. This plays a crucial role in the development of explicit numerical methods, for which a major stability requirement is to keep the time steps inversely proportional to the maximum wave speed over the entire computational domain.

It is well-known that low-Mach-number flows pose several major challenges for numerical simulations. A distinctive feature of such regimes is the appearance of both slow material waves, which transport quantities like entropy and vorticity, and fast acoustic waves, whose speeds scale inversely with the Mach number. As the Mach number decreases, the resulting stiffness imposes severe time-step restrictions on explicit methods and leads to excessive numerical diffusion, making such schemes inefficient or even impractical for real applications. Fully-implicit methods can address the stiffness, but have their own drawbacks: they tend to oversmear material waves (see, e.g., [10]), require the solution of large nonlinear systems, and may fail to capture the correct solution in the zero-Mach-number limit.

To overcome these difficulties, a widely adopted strategy is to use either implicit-explicit (IMEX) or semi-implicit (SI) methods based on conservative hyperbolic flux splitting. This approach decomposes the hyperbolic flux into stiff and nonstiff components in a manner that preserves the conservative structure of the original system. The fast (stiff) part, associated with acoustic waves, is treated (semi-)implicitly to relax time-step limitations, while the slow (nonstiff) part is handled explicitly to accurately capture the evolution of material waves without excessive numerical diffusion.

It is also known that, as the Mach number tends to zero, the compressible Euler equations reduce to the incompressible Euler equations. It is essential to ensure that numerical schemes also exhibit the same limiting behavior at the discrete level and provide a consistent discretization of the incompressible Euler equations as the Mach number tends to zero. Schemes that maintain this property are called asymptotic-preserving (AP). They were originally introduced to capture steady-state solutions for neutron transport in the diffusive regime [33, 34], but the specific definition was introduced in [22, 28, 30] in the context of stiff kinetic equations. In recent years, AP schemes have been extensively studied and applied for simulating low-Mach-number flows; see, e.g., [7, 9–11, 13, 16–20, 26, 31, 41, 45, 47, 49] for a non-exhaustive list of references.

All of the aforementioned AP schemes, which were designed for either the isentropic or full compressible Euler system, are based on different flux splitting strategies. A very simple and robust flux splitting, which was proposed in [26] for the isentropic Euler equations and later extended to the rotating shallow water equations in [36], seems to be rather optimal in the sense that it very accurately identifies and separates a linear stiff pressure term, which is then discretized implicitly. However, extending this flux splitting to the full Euler equations presents significant challenges.

In this paper, we propose an alternative way of accurately identifying and separating a stiff part of the full Euler system: we first rewrite the studied system in a nonconservative form and then introduce a nonconservative hyperbolic splitting, which may be naturally viewed as an extension of the flux splitting from [26]. We integrate the nonconservative system using a SI method implemented as follows. The stiff part is approximated semi-implicitly using a simple second-order accurate central-differencing, and the nonstiff part is handled explicitly using a second-order **path-conservative central-upwind (PCCU)** discretization. The resulting SI approach is then realized in such a way that the pressure update consists of solving a Poisson-type elliptic equation: this is

used to enforce the AP property.

However, the resulting SI method can only be applied to low-Mach-number regimes, where the magnitude of discontinuous waves is small. For large Mach numbers, solving nonconservative formulations of the Euler equations in the presence of discontinuities typically leads to nonphysical computed solutions, as was demonstrated in [4, 27]. We therefore apply a dual formulation (DF) approach, which has been recently introduced in [14] (for other recent works on DF methods, we refer the reader to [2, 3, 5, 43]), and solve the nonconservative and conservative formulations simultaneously. The latter one is discretized in a fully-explicit manner using the second-order semi-discrete central-upwind (CU) discretization from [32]. This way, at each stage of a multi-stage SI time discretization (we have used the deferred correction (DeC) time discretization from [42]), two copies of the computed solution are evolved: one is AP, but nonconservative, while the second one is conservative, but non-AP. Hence, upon the completion of each stage of the time evolution, we post-process the obtained solutions to automatically ensure that in low-Mach-number regimes, the overall numerical solution is obtained by the AP nonconservative scheme, while in large (intermediate)-Mach-number regimes, the solution reduces to the sharp and conservative solution obtained by the CU scheme.

The rest of the paper is organized as follows. In §2, we give the necessary preliminaries: we introduce the governing equations, namely, non-dimensional conservative and nonconservative (primitive) formulations of the full Euler equations, discuss their zero-Mach-number limit, and briefly review the considered DF framework. In §3, we introduce the novel AP DF finite-volume (DF-FV) scheme for compressible Euler equations, providing a rigorous proof of its AP character. In §4, we demonstrate the performance of the proposed scheme on a number of challenging numerical examples. Finally, concluding remarks can be found in §5.

2 Preliminaries

The main goal of this section is to provide the background needed for presenting the proposed AP scheme. Specifically, we will describe:

- a nonconservative reformulation of the Euler equations in terms of primitive variables, which allows for a natural decomposition of the terms, which are stiff and nonstiff in the low-Mach-number regime;
- a formal asymptotic analysis of the studied equations in low-Mach-number regimes, providing the incompressible system that the AP scheme must accurately approximate in the zero-Mach-number limit;
- a DF framework, in which both conservative and nonconservative formulations of the studied system are numerically solved simultaneously exploiting the advantages of each of them in the corresponding Mach-number regimes.

2.1 Conservative and Primitive Formulations

After suitable non-dimensionalization and rescaling, the two-dimensional (2-D) compressible Euler equations can be written in the conservative form as

$$\rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \quad (2.1)$$

$$(\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \frac{1}{\varepsilon^2} \nabla p = \mathbf{0}, \quad (2.2)$$

$$E_t + \nabla \cdot ((E + p) \mathbf{u}) = 0. \quad (2.3)$$

Here, ρ , $\mathbf{u} := (u, v)^\top$, and E denote the density, velocity, and total energy, respectively, p is the pressure, ε is the reference Mach number, and the system is closed by the equation of state, which, in the case of an ideal gas, reads as

$$E = \frac{p}{\gamma - 1} + \frac{\varepsilon^2}{2} \rho (u^2 + v^2), \quad (2.4)$$

with γ being the specific heat ratio. This system is hyperbolic and features acoustic waves traveling with (maximum) speed $|\mathbf{u}| + c$, where c is the speed of sound given by $c := \frac{1}{\varepsilon} \sqrt{\gamma p / \rho}$. Notice that in low-Mach-number regimes, the acoustic waves travel at a high (maximum) speed proportional to $1/\varepsilon$.

For the purpose of deriving our AP scheme, we also consider an equivalent nonconservative formulation of the system (2.1)–(2.4) in terms of the primitive variables ρ , $\mathbf{u}$, and p :

$$\rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \quad (2.5)$$

$$\mathbf{u}_t + (\mathbf{u} \cdot \nabla) \mathbf{u} + \frac{1}{\varepsilon^2 \rho} \nabla p = \mathbf{0}, \quad (2.6)$$

$$p_t + \mathbf{u} \cdot \nabla p + \gamma p \nabla \cdot \mathbf{u} = 0. \quad (2.7)$$

We emphasize that this formulation is equivalent to the conservative system (2.1)–(2.4) only for smooth solutions, and numerical approximations of (2.5)–(2.7) typically converge to nonphysical solutions when discontinuities are present; see [4, 27] for a detailed discussion.

2.2 Zero-Mach-Number Limit

It is well-known (see, e.g., a formal analysis in [6, 37]) that in the zero-Mach-number limit the compressible Euler equations reduce to the incompressible ones. To illustrate this, we examine the formal behavior of the primitive system (2.5)–(2.7) as $\varepsilon \rightarrow 0$. We substitute the formal expansions

$$\rho = \rho^{(0)} + \varepsilon \rho^{(1)} + \varepsilon^2 \rho^{(2)} + \dots, \quad \mathbf{u} = \mathbf{u}^{(0)} + \varepsilon \mathbf{u}^{(1)} + \varepsilon^2 \mathbf{u}^{(2)} + \dots, \quad p = p^{(0)} + \varepsilon p^{(1)} + \varepsilon^2 p^{(2)} + \dots$$

into (2.5)–(2.7) and collect terms by powers of ε . This yields

$$\mathcal{O}(\varepsilon^{-2}) : \quad \nabla p^{(0)} = \mathbf{0}, \quad (2.8)$$

$$\mathcal{O}(\varepsilon^{-1}) : \quad \nabla p^{(1)} = \mathbf{0}, \quad (2.9)$$

$$\mathcal{O}(1) : \quad \rho_t^{(0)} + \nabla \cdot (\rho^{(0)} \mathbf{u}^{(0)}) = 0, \quad (2.10)$$

$$\mathbf{u}_t^{(0)} + (\mathbf{u}^{(0)} \cdot \nabla) \mathbf{u}^{(0)} + \frac{1}{\rho^{(0)}} \nabla p^{(2)} = \mathbf{0}, \quad (2.11)$$

$$p_t^{(0)} + \gamma p^{(0)} \nabla \cdot \mathbf{u}^{(0)} = 0, \quad (2.12)$$

$$\mathcal{O}(\varepsilon) : \quad p_t^{(1)} + \gamma (p^{(1)} \nabla \cdot \mathbf{u}^{(0)} + p^{(0)} \nabla \cdot \mathbf{u}^{(1)}) = 0. \quad (2.13)$$

It follows from (2.8)–(2.9) that $p^{(0)}(x, y, t) = p^{(0)}(t)$ and $p^{(1)}(x, y, t) = p^{(1)}(t)$ are spatially uniform. One can also show that both $p^{(0)}$ and $p^{(1)}$ are independent of time, provided the following Dirichlet boundary condition holds:

$$p = \bar{p}_0 + \varepsilon \bar{p}_1 + \varepsilon^2 p_2 + \dots, \quad \forall (x, y) \in \partial\Omega, \quad (2.14)$$

where $\bar{p}_0 > 0$ and $\bar{p}_1$ are constants, $p_k(x, y, t)$, $k \geq 2$ are bounded functions, and Ω is the spatial domain with boundary $\partial\Omega$. In fact, such a boundary condition implies $p^{(0)} \equiv \bar{p}_0$ and $p^{(1)} \equiv \bar{p}_1$. Using this in (2.12), one concludes that $\nabla \cdot \mathbf{u}^{(0)} = 0$, which further implies $\nabla \cdot \mathbf{u}^{(1)} = 0$ thanks to (2.13). Hence, the zero-Mach-number limiting equations are

$$\rho_t^{(0)} + \nabla \cdot (\rho^{(0)} \mathbf{u}^{(0)}) = 0, \quad (2.15)$$

$$\mathbf{u}_t^{(0)} + (\mathbf{u}^{(0)} \cdot \nabla) \mathbf{u}^{(0)} + \frac{1}{\rho^{(0)}} \nabla p^{(2)} = \mathbf{0}, \quad (2.16)$$

$$\nabla \cdot \mathbf{u}^{(0)} = 0, \quad p^{(0)} \equiv \bar{p}_0, \quad (2.17)$$

and the correct low-Mach-number scaling for the $\mathcal{O}(\varepsilon)$ terms is

$$\nabla \cdot \mathbf{u}^{(1)} = 0, \quad p^{(1)} \equiv \bar{p}_1. \quad (2.18)$$

2.3 Dual Formulation (DF) Framework

As outlined above, a new AP scheme for the compressible Euler equations will be constructed within the DF framework, in which both conservative and primitive formulations are numerically solved simultaneously. The conservative form ensures proper handling of discontinuities, while the primitive form is used to achieve the AP property in the zero-Mach-number limit. While the DF methodology is not specifically developed to handle multiscale features, it lays the groundwork for the AP scheme developed in subsequent sections by enabling an efficient and accurate treatment of the studied Euler equations in both compressible and nearly incompressible flow regimes.

We consider a general 2-D hyperbolic system of conservation laws,

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x + \mathbf{G}(\mathbf{U})_y = \mathbf{0}, \quad (2.19)$$

where $\mathbf{U}$ is the vector of conservative variables and $\mathbf{F}$ and $\mathbf{G}$ are fluxes, and rewrite it in an equivalent nonconservative form

$$\mathbf{V}_t + \tilde{\mathbf{F}}(\mathbf{V})_x + \tilde{\mathbf{G}}(\mathbf{V})_y = B(\mathbf{V})\mathbf{V}_x + C(\mathbf{V})\mathbf{V}_y, \quad (2.20)$$

where $\mathbf{V}$ is the vector of nonconservative variables, $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{G}}$ are the corresponding fluxes, and $B(\mathbf{V})\mathbf{V}_x$ and $C(\mathbf{V})\mathbf{V}_y$ are the nonconservative product terms.

The key idea of the DF approach is to evolve the solutions of (2.19) and (2.20) simultaneously. A crucial step in DF-based methods is a post-processing, in which the evolved values of the nonconservative solution $\mathbf{V}$ are replaced with a more reliable approximation after the update. This step is necessary because long-term evolutions of $\mathbf{V}$ by directly solving the nonconservative system (2.20) may lead to nonphysical solutions in the presence of discontinuities, which typically appear when the studied Euler system is considered in the compressible (large/medium-Mach-number) regime.

The post-processing can be described as follows. After advancing the solutions of (2.19) and (2.20) from a certain time level t to the next time level $t + \Delta t$, the evolved values of $\mathbf{V}(t + \Delta t)$ are replaced with

$$r(\mathbf{V}(\mathbf{U}(t + \Delta t)), \mathbf{V}(t + \Delta t)), \quad (2.21)$$

where r is a suitable replacement function and $\mathbf{V}(\mathbf{U})$ is a conservative-primitive variable transformation. In the simplest nonstiff case, one can set

$$r(\mathbf{V}(\mathbf{U}(t + \Delta t)), \mathbf{V}(t + \Delta t)) = \mathbf{V}(\mathbf{U}(t + \Delta t)). \quad (2.22)$$

However, in the development of the AP scheme below, we will modify the post-processing (2.22) by taking an appropriate function r in (2.21) to ensure that in the nearly incompressible (low-Mach-number) regime the AP $\mathbf{V}$ -solution is not overwritten by the non-AP conservative one.

3 Novel AP Scheme for Compressible Euler Equations

Building on the DF framework described in §2.3, we now present a novel AP scheme designed for the compressible Euler equations across all Mach-number regimes, from fully compressible to nearly incompressible flows. The proposed method couples the conservative (2.1)–(2.4) and primitive (2.5)–(2.7) formulations of the system, ensuring stability, accuracy, and consistency with the analytical asymptotic behavior as $\varepsilon \rightarrow 0$.

In this section, we provide a complete description of the proposed space-time discretization, starting with the primitive system (2.5)–(2.7). In §3.1, we outline its temporal integration, which is based on a new hyperbolic splitting and an SI approach. In §3.2, we present a fully discrete second-order AP scheme for the primitive system, and in §3.3, we describe the semi-discrete CU scheme employed for the conservative system. §3.4 is devoted to the clarification of important implementation details. Finally, in §3.5, we present the Mach-number dependent post-processing strategy used to reconcile primitive and conservative variables.

3.1 Novel AP Time Discretization of the Primitive System

We begin by providing a precise definition of an asymptotic-preserving (AP) time discretization in the context of the zero-Mach-number limit.

Definition 3.1 (AP time discretization) *Assume that the semi-discrete (continuous in space) solution at time t^n , $(\rho^n(x, y), \mathbf{u}^n(x, y), p^n(x, y))$, can be expanded as*

$$\begin{aligned} \rho^n &= \rho^{(0),n} + \varepsilon \rho^{(1),n} + \varepsilon^2 \rho^{(2),n} + \dots, & \mathbf{u}^n &= \mathbf{u}^{(0),n} + \varepsilon \mathbf{u}^{(1),n} + \varepsilon^2 \mathbf{u}^{(2),n} + \dots, \\ p^n &= p^{(0),n} + \varepsilon p^{(1),n} + \varepsilon^2 p^{(2),n} + \dots, \end{aligned} \quad (3.1)$$

which is compatible with the asymptotic limits (2.17) and (2.18), that is,

$$p^{(0),n} \equiv \bar{p}_0, \quad p^{(1),n} \equiv \bar{p}_1, \quad (3.2)$$

$$\nabla \cdot \mathbf{u}^{(0),n} = 0, \quad \nabla \cdot \mathbf{u}^{(1),n} = 0, \quad (3.3)$$

and assume that the Dirichlet boundary condition (2.14) holds, namely,

$$p \equiv \bar{p}_0 + \varepsilon \bar{p}_1 + \varepsilon^2 p_2 + \dots, \quad \forall (x, y) \in \partial\Omega. \quad (3.4)$$

Let us consider a one-step time discretization that produces at time $t^{n+1} = t^n + \Delta t$ an approximation $(\rho^{n+1}, \mathbf{u}^{n+1}, p^{n+1})$. We say that such time discretization is AP if it admits an asymptotic expansion of the type (3.1) and yields a consistent discretization of (2.15)–(2.18) as $\varepsilon \rightarrow 0$.

To construct an AP time discretization, we first perform a hyperbolic splitting of the primitive system (2.5)–(2.7), separating the stiff pressure-driven terms from the nonstiff convective terms. This splitting enables the use of an SI integration strategy in which the stiff and nonstiff terms are treated semi-implicitly and explicitly, respectively, ensuring uniform stability and asymptotic consistency as $\varepsilon \rightarrow 0$.

3.1.1 A New Hyperbolic Splitting

We follow the idea from [26, 36] and split the nonconservative system into two parts corresponding to the slow and fast dynamics as follows. We first define the time-dependent variables

$$\rho_{\max}(t) := \max_{(x,y) \in \Omega} \rho(x, y, t), \quad p_{\min}(t) := \min_{(x,y) \in \Omega} p(x, y, t), \quad (3.5)$$

and then add and subtract $\frac{1}{\varepsilon^2 \rho_{\max}} \nabla p$ and $\gamma p_{\min} \nabla \cdot \mathbf{u}$ from (2.6) and (2.7), respectively, to rewrite system (2.5)–(2.7) as follows:

$$\rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \quad (3.6)$$

$$\mathbf{u}_t + (\mathbf{u} \cdot \nabla) \mathbf{u} + \frac{\rho_{\max} - \rho}{\varepsilon^2 \rho \rho_{\max}} \nabla p = -\frac{1}{\varepsilon^2 \rho_{\max}} \nabla p, \quad (3.7)$$

$$p_t + \mathbf{u} \cdot \nabla p + \gamma(p - p_{\min}) \nabla \cdot \mathbf{u} = -\gamma p_{\min} \nabla \cdot \mathbf{u}. \quad (3.8)$$

This system can be put in the following vector form:

$$\mathbf{V}_t + \tilde{\mathbf{F}}(\mathbf{V})_x + \tilde{\mathbf{G}}(\mathbf{V})_y = \tilde{B}(\mathbf{V})\mathbf{V}_x + \tilde{C}(\mathbf{V})\mathbf{V}_y + \hat{B}(\mathbf{V})\mathbf{V}_x + \hat{C}(\mathbf{V})\mathbf{V}_y, \quad (3.9)$$

where $\mathbf{V} := (\rho, u, v, p)^\top$, the nonlinear nonstiff (slow dynamics) part consists of the fluxes

$$\tilde{\mathbf{F}}(\mathbf{V}) = \left(\rho u, \frac{u^2}{2}, 0, 0 \right)^\top \quad \text{and} \quad \tilde{\mathbf{G}}(\mathbf{V}) = \left(\rho v, 0, \frac{v^2}{2}, 0 \right)^\top,$$

and the nonstiff nonconservative terms $\tilde{B}(\mathbf{V})\mathbf{V}_x + \tilde{C}(\mathbf{V})\mathbf{V}_y$ with matrices

$$\tilde{B} = - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\rho_{\max} - \rho}{\varepsilon^2 \rho \rho_{\max}} \\ 0 & v & 0 & 0 \\ 0 & \gamma(p - p_{\min}) & 0 & u \end{pmatrix} \quad \text{and} \quad \tilde{C} = - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & u & 0 \\ 0 & 0 & 0 & \frac{\rho_{\max} - \rho}{\varepsilon^2 \rho \rho_{\max}} \\ 0 & 0 & \gamma(p - p_{\min}) & v \end{pmatrix},$$

while, the linear stiff (fast dynamics) part consists of the stiff nonconservative terms $\hat{B}(\mathbf{V})\mathbf{V}_x + \hat{C}(\mathbf{V})\mathbf{V}_y$ with matrices

$$\hat{B} = - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\varepsilon^2 \rho_{\max}} \\ 0 & 0 & 0 & 0 \\ 0 & \gamma p_{\min} & 0 & 0 \end{pmatrix} \quad \text{and} \quad \hat{C} = - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\varepsilon^2 \rho_{\max}} \\ 0 & 0 & \gamma p_{\min} & 0 \end{pmatrix}.$$

We stress that the subsystem $\mathbf{V}_t + \tilde{\mathbf{F}}(\mathbf{V})_x + \tilde{\mathbf{G}}(\mathbf{V})_y = \tilde{B}(\mathbf{V})\mathbf{V}_x + \tilde{C}(\mathbf{V})\mathbf{V}_y$ is indeed nonstiff as the eigenvalues of the matrices $\frac{\partial \tilde{\mathbf{F}}}{\partial \mathbf{V}}(\mathbf{V}) - \tilde{B}(\mathbf{V})$ and $\frac{\partial \tilde{\mathbf{G}}}{\partial \mathbf{V}}(\mathbf{V}) - \tilde{C}(\mathbf{V})$, are $\{u \pm \tilde{c}, u, u\}$ and $\{v \pm \tilde{c}, v, v\}$, respectively, with

$$\tilde{c} := \frac{1}{\varepsilon} \sqrt{\gamma \frac{(\rho_{\max} - \rho)(p - p_{\min})}{\rho \rho_{\max}}}, \quad (3.10)$$

which are real and of size $\mathcal{O}(1)$ thanks to the definitions of $\rho_{\max}$ and $p_{\min}$ in (3.5) and to the asymptotic analysis in §2.2, which ensure that

$$0 \leq \rho_{\max} - \rho = \mathcal{O}(1) \quad \text{and} \quad 0 \leq p - p_{\min} = \mathcal{O}(\varepsilon^2). \quad (3.11)$$

In the next subsection, we will utilize this splitting and design an AP time discretization based on an explicit approximation of the nonstiff subsystem and an SI discretization of the stiff terms on the right-hand sides (RHSs) of (3.7) and (3.8).

Remark 3.1 *It should be pointed out that extensions of the proposed splitting to more general equations of state would require a re-examination of the primitive-variable formulation.*

3.1.2 First-Order AP SI Time Discretization

The simplest first-order AP SI time discretization of the system (3.9) reads as

$$\begin{aligned} \frac{\rho^{n+1} - \rho^n}{\Delta t} + \nabla \cdot (\rho^n \mathbf{u}^n) &= 0, \\ \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + (\mathbf{u}^n \cdot \nabla) \mathbf{u}^n + \frac{\rho_{\max}^n - \rho^n}{\varepsilon^2 \rho^n \rho_{\max}^n} \nabla p^n + \frac{1}{\varepsilon^2 \rho_{\max}^n} \nabla p^{n+1} &= \mathbf{0}, \\ \frac{p^{n+1} - p^n}{\Delta t} + \mathbf{u}^n \cdot \nabla p^n + \gamma(p^n - p_{\min}^n) \nabla \cdot \mathbf{u}^n + \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{n+1} &= 0, \end{aligned} \quad (3.12)$$

which can also be written in the following vector form:

$$\frac{\mathbf{V}^{n+1} - \mathbf{V}^n}{\Delta t} + \mathcal{R}^n + \mathcal{L}^{n,n+1} = \mathbf{0},$$

where $\mathbf{V}^n := (\rho^n, \mathbf{u}^n, p^n)^\top \approx \mathbf{V}(t^n)$, $\rho_{\max}^n := \rho_{\max}(t^n)$, $p_{\min}^n := p_{\min}(t^n)$, and

$$\begin{aligned} \mathcal{R}^n &:= \tilde{\mathbf{F}}(\mathbf{V}^n)_x + \tilde{\mathbf{G}}(\mathbf{V}^n)_y - \tilde{B}(\mathbf{V}^n)\mathbf{V}_x^n - \tilde{C}(\mathbf{V}^n)\mathbf{V}_y^n \\ &= \begin{pmatrix} (\mathcal{R}^\rho)^n \\ (\mathcal{R}^u)^n \\ (\mathcal{R}^p)^n \end{pmatrix} = \begin{pmatrix} \nabla \cdot (\rho^n \mathbf{u}^n) \\ (\mathbf{u}^n \cdot \nabla) \mathbf{u}^n + \frac{\rho_{\max}^n - \rho^n}{\varepsilon^2 \rho^n \rho_{\max}^n} \nabla p^n \\ \mathbf{u}^n \cdot \nabla p^n + \gamma(p^n - p_{\min}^n) \nabla \cdot \mathbf{u}^n \end{pmatrix}, \end{aligned} \quad (3.13)$$

$$\mathcal{L}^{n,n+1} := -\hat{B}(\mathbf{V}^n)\mathbf{V}_x^{n+1} - \hat{C}(\mathbf{V}^n)\mathbf{V}_y^{n+1} = \begin{pmatrix} (\mathcal{L}^\rho)^{n,n+1} \\ (\mathcal{L}^u)^{n,n+1} \\ (\mathcal{L}^p)^{n,n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{\nabla p^{n+1}}{\varepsilon^2 \rho_{\max}^n} \\ \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{n+1} \end{pmatrix}. \quad (3.14)$$

Notice that in (3.14), $\mathcal{L}^{n,n+1}$ is discretized in an SI (rather than fully implicit) manner, where both $\widehat{B}$ and $\widehat{C}$ are evaluated at $\mathbf{V}^n$ (and not at $\mathbf{V}^{n+1}$), which prevents from numerically solving complicated systems of nonlinear algebraic equations.

We shall now prove that the time discretization (3.12) is indeed AP, provided that the time step is computed based on the wave speeds of the nonstiff subsystem, that is, provided that

$$\Delta t = K_{\text{CFL}} \min \left\{ \frac{\Delta x}{\max_{(x,y) \in \Omega} (|u| + \tilde{c})}, \frac{\Delta y}{\max_{(x,y) \in \Omega} (|v| + \tilde{c})} \right\}, \quad (3.15)$$

where K_{CFL} is a CFL number and Δx and Δy are mesh sizes used in the spatial discretization. Notice that selecting the time step Δt according to (3.15) makes it asymptotically independent of ε as, according to (3.10)–(3.11), $\tilde{c} = \mathcal{O}(1)$.

Theorem 3.1 *The first-order SI time discretization (3.12) is AP according to Definition 3.1, provided that Δt is computed as in (3.15).*

Proof: We begin by formally showing that the computed solution ρ^{n+1} , $\mathbf{u}^{n+1}$, p^{n+1} admits an expansion of the type (3.1) satisfying (3.2) and (3.3) in the limit as $\varepsilon \rightarrow 0$. We substitute the corresponding expansion of the numerical solution at time t^n into the scheme (3.12) and use (3.2)–(3.3) to obtain

$$\begin{aligned} \rho^{n+1} &= \rho^{(0),n} + \varepsilon \rho^{(1),n} + \varepsilon^2 \rho^{(2),n} - \Delta t \nabla \cdot (\rho^{(0),n} \mathbf{u}^{(0),n}) \\ &\quad - \varepsilon \Delta t [\nabla \cdot (\rho^{(1),n} \mathbf{u}^{(0),n}) + \nabla \cdot (\rho^{(0),n} \mathbf{u}^{(1),n})] \\ &\quad - \varepsilon^2 \Delta t [\nabla \cdot (\rho^{(2),n} \mathbf{u}^{(0),n}) + \nabla \cdot (\rho^{(1),n} \mathbf{u}^{(1),n}) + \nabla \cdot (\rho^{(0),n} \mathbf{u}^{(2),n})] + \mathcal{O}(\varepsilon^3), \end{aligned} \quad (3.16)$$

$$\begin{aligned} \mathbf{u}^{n+1} &= \mathbf{u}^{(0),n} + \varepsilon \mathbf{u}^{(1),n} + \varepsilon^2 \mathbf{u}^{(2),n} - \Delta t \left[(\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(0),n} + \frac{\rho_{\max}^n - \rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} \right] \\ &\quad - \varepsilon \Delta t \left[(\mathbf{u}^{(1),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(1),n} - \frac{\rho^{(1),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} - \frac{\rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(3),n} \right] \\ &\quad - \varepsilon^2 \Delta t \left[(\mathbf{u}^{(2),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(1),n} \cdot \nabla) \mathbf{u}^{(1),n} + (\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(2),n} \right. \\ &\quad \left. - \frac{\rho^{(2),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} - \frac{\rho^{(1),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(3),n} - \frac{\rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(4),n} \right] - \frac{\Delta t}{\varepsilon^2 \rho_{\max}^n} \nabla p^{n+1} + \mathcal{O}(\varepsilon^3), \end{aligned} \quad (3.17)$$

$$p^{n+1} = \bar{p}_0 + \varepsilon \bar{p}_1 + \varepsilon^2 p^{(2),n} - \varepsilon^2 \Delta t \mathbf{u}^{(0),n} \cdot \nabla p^{(2),n} - \Delta t \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{n+1} + \mathcal{O}(\varepsilon^3). \quad (3.18)$$

Thanks to the explicit nature of the density update in (3.16), we conclude that indeed ρ^{n+1} admits the required asymptotic expansion $\rho^{n+1} = \rho^{(0),n+1} + \varepsilon \rho^{(1),n+1} + \varepsilon \rho^{(2),n+1} + \dots$, where the different terms of the expansion are obtained by collecting corresponding powers of ε :

$$\begin{aligned} \rho^{(0),n+1} &= \rho^{(0),n} - \Delta t \nabla \cdot (\rho^{(0),n} \mathbf{u}^{(0),n}), \\ \rho^{(1),n+1} &= \rho^{(1),n} - \Delta t [\nabla \cdot (\rho^{(1),n} \mathbf{u}^{(0),n}) + \nabla \cdot (\rho^{(0),n} \mathbf{u}^{(1),n})], \\ \rho^{(2),n+1} &= \rho^{(2),n} - \Delta t [\nabla \cdot (\rho^{(2),n} \mathbf{u}^{(0),n}) + \nabla \cdot (\rho^{(1),n} \mathbf{u}^{(1),n}) + \nabla \cdot (\rho^{(0),n} \mathbf{u}^{(2),n})]. \end{aligned} \quad (3.19)$$

In order to show that also p^{n+1} admits an expansion of the same type, we take the divergence of the velocity equation (3.17), substitute $\nabla \cdot \mathbf{u}^{n+1}$ into the pressure equation (3.18), and use the

divergence-free assumption (3.3) to obtain

$$\begin{aligned}
p^{n+1} = & \bar{p}_0 + \varepsilon \bar{p}_1 + \varepsilon^2 p^{(2),n} - \varepsilon^2 \Delta t \mathbf{u}^{(0),n} \cdot \nabla p^{(2),n} \\
& - \Delta t \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{(0),n} - \varepsilon \Delta t \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{(1),n} - \varepsilon^2 \Delta t \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{(2),n} \\
& + (\Delta t)^2 \gamma p_{\min}^n \nabla \cdot \left[(\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(0),n} + \frac{\rho_{\max}^n - \rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} \right] \\
& + \varepsilon (\Delta t)^2 \gamma p_{\min}^n \nabla \cdot \left[(\mathbf{u}^{(1),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(1),n} - \frac{\rho^{(1),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} - \frac{\rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(3),n} \right] \\
& + \varepsilon^2 (\Delta t)^2 \gamma p_{\min}^n \nabla \cdot \left[(\mathbf{u}^{(2),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(1),n} \cdot \nabla) \mathbf{u}^{(1),n} + (\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(2),n} \right. \\
& \quad \left. - \frac{\rho^{(2),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} - \frac{\rho^{(1),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(3),n} - \frac{\rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(4),n} \right] + \frac{(\Delta t)^2 \gamma p_{\min}^n}{\varepsilon^2 \rho_{\max}^n} \Delta p^{n+1} + \mathcal{O}(\varepsilon^3),
\end{aligned}$$

which implies that p^{n+1} is the solution of the elliptic equation

$$-\Delta p^{n+1} + \frac{\varepsilon^2 \rho_{\max}^n}{(\Delta t)^2 \gamma p_{\min}^n} p^{n+1} = \mathcal{O}(\varepsilon^2),$$

subject to the boundary condition (3.4). According to the theory of perturbed linear operators [29], one can conclude that

$$p^{n+1} = p^{(0),n+1} + \varepsilon p^{(1),n+1} + \varepsilon^2 p^{(2),n+1} + \dots, \quad (3.20)$$

$$p^{(0),n+1} = p^{(0),n} \equiv \bar{p}_0, \quad p^{(1),n+1} = p^{(1),n} \equiv \bar{p}_1, \quad (3.21)$$

and hence

$$\nabla p^{n+1} = \varepsilon^2 \nabla p^{(2),n+1} + \varepsilon^3 \nabla p^{(3),n+1} + \varepsilon^4 \nabla p^{(4),n+1} + \dots,$$

which we substitute into (3.17) to obtain the velocity expansion

$$\mathbf{u}^{n+1} = \mathbf{u}^{(0),n+1} + \varepsilon \mathbf{u}^{(1),n+1} + \varepsilon^2 \mathbf{u}^{(2),n+1} + \dots$$

with

$$\begin{aligned}
\mathbf{u}^{(0),n+1} = & \mathbf{u}^{(0),n} - \Delta t \left[(\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(0),n} + \frac{\rho_{\max}^n - \rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} \right] - \frac{\Delta t}{\rho_{\max}^n} \nabla p^{(2),n+1}, \\
\mathbf{u}^{(1),n+1} = & \mathbf{u}^{(1),n} - \Delta t \left[(\mathbf{u}^{(1),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(1),n} \right. \\
& \quad \left. - \frac{\rho^{(1),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} - \frac{\rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(3),n} \right] - \frac{\Delta t}{\rho_{\max}^n} \nabla p^{(3),n+1}, \\
\mathbf{u}^{(2),n+1} = & \mathbf{u}^{(2),n} - \Delta t \left[(\mathbf{u}^{(2),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(1),n} \cdot \nabla) \mathbf{u}^{(1),n} + (\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(2),n} \right. \\
& \quad \left. - \frac{\rho^{(2),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} - \frac{\rho^{(1),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(3),n} - \frac{\rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(4),n} \right] - \frac{\Delta t}{\rho_{\max}^n} \nabla p^{(4),n+1}.
\end{aligned} \quad (3.22)$$

We now need to show that (3.2) and (3.3) hold for the updated solution, along with the consistency of the scheme (3.12) with (2.15) and (2.16) as $\varepsilon \rightarrow 0$. We have already shown that

(3.2) holds; see (3.21). The divergence-free conditions (3.3) can be deduced from the pressure update (3.18), which in view of the obtained results yields

$$\begin{aligned}\varepsilon^2 p^{(2),n+1} &= \varepsilon^2 p^{(2),n} - \varepsilon^2 \Delta t \mathbf{u}^{(0),n} \cdot \nabla p^{(2),n} - \Delta t \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{(0),n+1} \\ &\quad - \varepsilon \Delta t \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{(1),n+1} - \varepsilon^2 \Delta t \gamma p_{\min}^n \nabla \cdot \mathbf{u}^{(2),n+1} + \mathcal{O}(\varepsilon^3).\end{aligned}$$

Collecting the power-like terms of ε , we deduce $\nabla \cdot \mathbf{u}^{(0),n+1} = \nabla \cdot \mathbf{u}^{(1),n+1} = 0$.

The consistency with (2.15) immediately follows from the first equation in (3.19). To show the consistency with (2.16), we rewrite the first equation in (3.22) as

$$\mathbf{u}^{(0),n+1} = \mathbf{u}^{(0),n} - \Delta t \left[(\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(0),n} + \frac{\nabla p^{(2),n}}{\rho^{(0),n}} \right] - \frac{\Delta t}{\rho_{\max}^n} (\nabla p^{(2),n+1} - \nabla p^{(2),n}), \quad (3.23)$$

which is a consistent discretization of (2.16).

We remark that since the zeroth and first modes of the pressure are constant, the evolution of the pressure in the zero-Mach-number limit essentially consists of the evolution of the second mode. Thus, $\nabla p^{(2),n+1} - \nabla p^{(2),n} \approx \mathcal{O}(\Delta t)$ and the last term in (3.23), in fact, represents a temporal diffusion term, which is proportional to $\mathcal{O}((\Delta t)^2)$. We also remark that according to (3.15), the time step Δt is asymptotically independent of ε . $\blacksquare$

3.1.3 Second-Order AP SI Time Discretization

We now introduce a second-order AP SI time discretization, which is based on the DeC approach, which was originally introduced in [21]. Our second-order AP SI-DeC time discretization is directly related to the IMEX-DeC methods presented in [42] and based on the DeC formulation introduced in [1]; see also [38, 39].

According to the second-order AP SI-DeC time discretization, the solution of (3.9) is evolved from $t = t^n$ to $t = t^{n+1}$ through the following two stages:

$$\begin{aligned}V^* &= V^n - \Delta t \mathcal{R}^n - \Delta t \mathcal{L}^{n,*}, \\ V^{n+1} &= V^n - \frac{\Delta t}{2} [\mathcal{R}^n + \mathcal{R}^*] - \frac{\Delta t}{2} [\mathcal{L}^{n,n} - \mathcal{L}^{*,*}] - \Delta t \mathcal{L}^{*,n+1},\end{aligned} \quad (3.24)$$

where the upper index $*$ is associated with the intermediate solution V^* , and the definitions of the operators $\mathcal{R}^*$ and $\mathcal{L}^{n,*}$, $\mathcal{L}^{*,*}$, and $\mathcal{L}^{*,n+1}$ are analogous to those given in (3.13) and (3.14), respectively.

The scheme (3.24) can be equivalently written as

$$\begin{aligned}\rho^* &= \rho^n - \Delta t \nabla \cdot (\rho^n \mathbf{u}^n), \\ \mathbf{u}^* &= \mathbf{u}^n - \Delta t \left[(\mathbf{u}^n \cdot \nabla) \mathbf{u}^n + \frac{\rho_{\max}^n - \rho^n}{\varepsilon^2 \rho^n \rho_{\max}^n} \nabla p^n \right] - \frac{\Delta t}{\varepsilon^2 \rho_{\max}^n} \nabla p^*, \\ p^* &= p^n - \Delta t [\mathbf{u}^n \cdot \nabla p^n + \gamma (p^n - p_{\min}^n) \nabla \cdot \mathbf{u}^n] - \Delta t \gamma p_{\min}^n \nabla \cdot \mathbf{u}^*,\end{aligned} \quad (3.25)$$

and

$$\begin{aligned}
\rho^{n+1} &= \rho^n - \frac{\Delta t}{2} [\nabla \cdot (\rho^n \mathbf{u}^n) + \nabla \cdot (\rho^* \mathbf{u}^*)], \\
\mathbf{u}^{n+1} &= \mathbf{u}^n - \frac{\Delta t}{2} \left[(\mathbf{u}^n \cdot \nabla) \mathbf{u}^n + (\mathbf{u}^* \cdot \nabla) \mathbf{u}^* + \frac{\rho_{\max}^n - \rho^n}{\varepsilon^2 \rho^n \rho_{\max}^n} \nabla p^n + \frac{\rho_{\max}^* - \rho^*}{\varepsilon^2 \rho^* \rho_{\max}^*} \nabla p^* \right] \\
&\quad - \frac{\Delta t}{2\varepsilon^2} \left(\frac{\nabla p^n}{\rho_{\max}^n} - \frac{\nabla p^*}{\rho_{\max}^*} \right) - \frac{\Delta t}{\varepsilon^2 \rho_{\max}^*} \nabla p^{n+1}, \\
p^{n+1} &= p^n - \frac{\Delta t}{2} [\mathbf{u}^n \cdot \nabla p^n + \mathbf{u}^* \cdot \nabla p^* + \gamma(p^n - p_{\min}^n) \nabla \cdot \mathbf{u}^n + \gamma(p^* - p_{\min}^*) \nabla \cdot \mathbf{u}^*] \\
&\quad - \frac{\Delta t}{2} \gamma(p_{\min}^n \nabla \cdot \mathbf{u}^n - p_{\min}^* \nabla \cdot \mathbf{u}^*) - \Delta t \gamma p_{\min}^* \nabla \cdot \mathbf{u}^{n+1}.
\end{aligned} \tag{3.26}$$

This second-order time discretization is indeed AP as shown in the next theorem.

Theorem 3.2 *The second-order SI-DeC discretization (3.25)–(3.26) is AP according to Definition 3.1, provided that Δt is computed according to (3.15).*

Proof: The proof proceeds along the same lines and uses the same arguments as in the proof of Theorem 3.1.

We begin by observing that the first stage of the second-order SI-DeC discretization coincides with the first-order AP SI time discretization studied before. Therefore, according to Theorem 3.1, the intermediate solution ρ^* , $\mathbf{u}^*$, p^* , obtained by (3.25) admits an expansion of the type (3.1), that is,

$$\begin{aligned}
\rho^* &= \rho^{(0),*} + \varepsilon \rho^{(1),*} + \varepsilon^2 \rho^{(2),*} + \dots, \quad \mathbf{u}^* = \mathbf{u}^{(0),*} + \varepsilon \mathbf{u}^{(1),*} + \varepsilon^2 \mathbf{u}^{(2),*} + \dots, \\
p^* &= p^{(0),*} + \varepsilon p^{(1),*} + \varepsilon^2 p^{(2),*} + \dots,
\end{aligned} \tag{3.27}$$

with

$$p^{(0),*} \equiv \bar{p}_0, \quad p^{(1),*} \equiv \bar{p}_1, \quad \nabla \cdot \mathbf{u}^{(0),*} = 0, \quad \nabla \cdot \mathbf{u}^{(1),*} = 0. \tag{3.28}$$

We then substitute the expansions (3.1) and (3.27) into (3.26) and use the conditions (3.2)–(3.3) and (3.28) to obtain

$$\begin{aligned}
\rho^{n+1} &= \rho^{(0),n} + \varepsilon \rho^{(1),n} + \varepsilon^2 \rho^{(2),n} - \frac{\Delta t}{2} \left[\nabla \cdot (\rho^{(0),n} \mathbf{u}^{(0),n}) + \nabla \cdot (\rho^{(0),*} \mathbf{u}^{(0),*}) \right] \\
&\quad - \frac{\varepsilon \Delta t}{2} \left[\nabla \cdot (\rho^{(1),n} \mathbf{u}^{(0),n}) + \nabla \cdot (\rho^{(0),n} \mathbf{u}^{(1),n}) + \nabla \cdot (\rho^{(1),*} \mathbf{u}^{(0),*}) + \nabla \cdot (\rho^{(0),*} \mathbf{u}^{(1),*}) \right] \\
&\quad - \frac{\varepsilon^2 \Delta t}{2} \left[\nabla \cdot (\rho^{(2),n} \mathbf{u}^{(0),n}) + \nabla \cdot (\rho^{(1),n} \mathbf{u}^{(1),n}) + \nabla \cdot (\rho^{(0),n} \mathbf{u}^{(2),n}) \right. \\
&\quad \left. + \nabla \cdot (\rho^{(2),*} \mathbf{u}^{(0),*}) + \nabla \cdot (\rho^{(1),*} \mathbf{u}^{(1),*}) + \nabla \cdot (\rho^{(0),*} \mathbf{u}^{(2),*}) \right] + \mathcal{O}(\varepsilon^3),
\end{aligned} \tag{3.29}$$

$$\begin{aligned}
\mathbf{u}^{n+1} = & \mathbf{u}^{(0),n} + \varepsilon \mathbf{u}^{(1),n} + \varepsilon^2 \mathbf{u}^{(2),n} \\
& - \frac{\Delta t}{2} \left[(\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(0),*} \cdot \nabla) \mathbf{u}^{(0),*} + \frac{\rho_{\max}^n - \rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} + \frac{\rho_{\max}^* - \rho^{(0),*}}{\rho^{(0),*} \rho_{\max}^*} \nabla p^{(2),*} \right] \\
& - \frac{\varepsilon \Delta t}{2} \left[(\mathbf{u}^{(1),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(1),n} + (\mathbf{u}^{(1),*} \cdot \nabla) \mathbf{u}^{(0),*} + (\mathbf{u}^{(0),*} \cdot \nabla) \mathbf{u}^{(1),*} \right. \\
& \quad \left. - \frac{\rho^{(1),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} - \frac{\rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(3),n} - \frac{\rho^{(1),*}}{\rho^{(0),*} \rho_{\max}^n} \nabla p^{(2),*} - \frac{\rho^{(0),*}}{\rho^{(0),*} \rho_{\max}^*} \nabla p^{(3),*} \right] \\
& - \frac{\varepsilon^2 \Delta t}{2} \left[(\mathbf{u}^{(2),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(1),n} \cdot \nabla) \mathbf{u}^{(1),n} + (\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(2),n} \right. \\
& \quad \left. + (\mathbf{u}^{(2),*} \cdot \nabla) \mathbf{u}^{(0),*} + (\mathbf{u}^{(1),*} \cdot \nabla) \mathbf{u}^{(1),*} + (\mathbf{u}^{(0),*} \cdot \nabla) \mathbf{u}^{(2),*} \right. \\
& \quad \left. - \frac{\rho^{(2),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(2),n} - \frac{\rho^{(1),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(3),n} - \frac{\rho^{(0),n}}{\rho^{(0),n} \rho_{\max}^n} \nabla p^{(4),n} \right. \\
& \quad \left. - \frac{\rho^{(2),*}}{\rho^{(0),*} \rho_{\max}^*} \nabla p^{(2),*} - \frac{\rho^{(1),*}}{\rho^{(0),*} \rho_{\max}^*} \nabla p^{(3),*} - \frac{\rho^{(0),*}}{\rho^{(0),*} \rho_{\max}^*} \nabla p^{(4),*} \right] \\
& - \frac{\Delta t}{2} \left[\frac{\nabla p^{(2),n}}{\rho_{\max}^n} - \frac{\nabla p^{(2),*}}{\rho_{\max}^*} \right] - \frac{\varepsilon \Delta t}{2} \left[\frac{\nabla p^{(3),n}}{\rho_{\max}^n} - \frac{\nabla p^{(3),*}}{\rho_{\max}^*} \right] - \frac{\varepsilon^2 \Delta t}{2} \left[\frac{\nabla p^{(4),n}}{\rho_{\max}^n} - \frac{\nabla p^{(4),*}}{\rho_{\max}^*} \right] \\
& - \frac{\Delta t}{\varepsilon^2 \rho_{\max}^*} \nabla p^{n+1} + \mathcal{O}(\varepsilon^3),
\end{aligned} \tag{3.30}$$

$$\begin{aligned}
p^{n+1} = & \bar{p}_0 + \varepsilon \bar{p}_1 + \varepsilon^2 p^{(2),n} - \frac{\varepsilon^2 \Delta t}{2} \left[\mathbf{u}^{(0),n} \cdot \nabla p^{(2),n} + \mathbf{u}^{(0),*} \cdot \nabla p^{(2),*} \right] \\
& - \frac{\varepsilon^2 \Delta t \gamma}{2} \left[p_{\min}^n \nabla \cdot \mathbf{u}^{(2),n} - p_{\min}^* \nabla \cdot \mathbf{u}^{(2),*} \right] - \Delta t \gamma p_{\min}^* \nabla \cdot \mathbf{u}^{n+1} + \mathcal{O}(\varepsilon^3).
\end{aligned} \tag{3.31}$$

The explicit nature of the density update (3.29) implies that ρ^{n+1} admits the required expansion $\rho^{n+1} = \rho^{(0),n+1} + \varepsilon \rho^{(1),n+1} + \varepsilon^2 \rho^{(2),n+1} + \dots$ with $\rho^{(0),n+1}$ satisfying

$$\rho^{(0),n+1} = \rho^{(0),n} - \frac{\Delta t}{2} \left[\nabla \cdot (\rho^{(0),n} \mathbf{u}^{(0),n}) + \nabla \cdot (\rho^{(0),*} \mathbf{u}^{(0),*}) \right], \tag{3.32}$$

and other coefficients satisfying the equations, which can be easily obtained by grouping the corresponding powers of ε .

As in the proof of Theorem 3.1, we show that p^{n+1} admits the asymptotic expansion by proving that it satisfies a well-posed elliptic problem with suitable boundary conditions. Taking the divergence of the velocity equation (3.30) and substituting $\nabla \cdot \mathbf{u}^{n+1}$ into the pressure equation (3.31) yields

$$-\Delta p^{n+1} + \frac{\varepsilon^2 \rho_{\max}^*}{(\Delta t)^2 \gamma p_{\min}^*} p^{n+1} = \mathcal{O}(\varepsilon^2).$$

This together with the boundary conditions (2.14), results in the same expansion for p^{n+1} , which we have established in (3.20)–(3.21) for the first-order SI method, leading to

$$\nabla p^{n+1} = \varepsilon^2 \left[\nabla p^{(2),n+1} + \varepsilon \nabla p^{(3),n+1} + \varepsilon^2 \nabla p^{(4),n+1} + \mathcal{O}(\varepsilon^3) \right]. \tag{3.33}$$

Next, we substitute (3.33) into the the velocity equation (3.30) and a straightforward grouping of the power-like terms of ε gives the equations for the coefficients of the velocity expansion

$\mathbf{u}^{n+1} = \mathbf{u}^{(0),n+1} + \varepsilon \mathbf{u}^{(1),n+1} + \varepsilon^2 \mathbf{u}^{(2),n+1} + \dots$. The equation for $\mathbf{u}^{(0),n+1}$ is

$$\begin{aligned} \mathbf{u}^{(0),n+1} = & \mathbf{u}^{(0),n} - \frac{\Delta t}{2} \left[(\mathbf{u}^{(0),n} \cdot \nabla) \mathbf{u}^{(0),n} + (\mathbf{u}^{(0),*} \cdot \nabla) \mathbf{u}^{(0),*} + \frac{\nabla p^{(2),n}}{\rho^{(0),n}} + \frac{\nabla p^{(2),*}}{\rho^{(0),*}} \right] \\ & - \frac{\Delta t}{\rho_{\max}^*} (\nabla p^{(2),n+1} - \nabla p^{(2),*}), \end{aligned} \quad (3.34)$$

and the other equations can be obtained similarly.

Let us now show the consistency with the asymptotic limit. The required conditions (3.21) on the pressure modes have been already shown. The divergence-free conditions for the velocity modes are then established from the pressure update (3.31), which becomes

$$\begin{aligned} \varepsilon^2 p^{(2),n+1} = & \varepsilon^2 p^{(2),n} - \frac{\varepsilon^2 \Delta t}{2} \left[\mathbf{u}^{(0),n} \cdot \nabla p^{(2),n} + \mathbf{u}^{(0),*} \cdot \nabla p^{(2),*} \right] - \frac{\varepsilon^2 \Delta t \gamma}{2} \left[p_{\min}^n \nabla \cdot \mathbf{u}^{(2),n} - p_{\min}^* \nabla \cdot \mathbf{u}^{(2),*} \right] \\ & - \Delta t \gamma p_{\min}^* \nabla \cdot \mathbf{u}^{(0),n+1} - \varepsilon \Delta t \gamma p_{\min}^* \nabla \cdot \mathbf{u}^{(1),n+1} - \varepsilon^2 \Delta t \gamma p_{\min}^* \nabla \cdot \mathbf{u}^{(2),n+1} + \mathcal{O}(\varepsilon^3). \end{aligned}$$

It is clear that the $\mathcal{O}(1)$ and $\mathcal{O}(\varepsilon)$ terms here vanish, that is, $\nabla \cdot \mathbf{u}^{(0),n+1} = \nabla \cdot \mathbf{u}^{(1),n+1} = 0$.

Finally, we notice that (3.32) and (3.34) are consistent discretizations of (2.15) and (2.16) with the last term in (3.34) representing a temporal diffusion, which is consistent with the order of accuracy of the scheme and thus proportional to $\mathcal{O}((\Delta t)^3)$. $\blacksquare$

Remark 3.2 *The described AP SI-DeC time discretization can be extended to arbitrarily high order in a straightforward way within the DeC framework. For the sake of brevity, we restrict our consideration to the second order of accuracy, which matches the accuracy that will be used in the spatial discretization discussed in §3.2.*

3.2 Fully Discrete Second-Order AP Scheme for the Primitive System

In this section, we construct a fully discrete scheme based on the second-order AP SI time discretization presented in §3.1.3. To this end, we first introduce uniform Cartesian cells $I_{j,k} := [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \times [y_{k-\frac{1}{2}}, y_{k+\frac{1}{2}}]$ with $x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}} \equiv \Delta x$ and $y_{k+\frac{1}{2}} - y_{k-\frac{1}{2}} \equiv \Delta y$, centered at (x_j, y_k) with $x_j = (x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})/2$ and $y_k = (y_{k-\frac{1}{2}} + y_{k+\frac{1}{2}})/2$, and assume that the cell averages $\bar{\mathbf{V}}_{j,k}^n := \frac{1}{\Delta x \Delta y} \iint_{I_{j,k}} \mathbf{V}(x, y, t^n) dx dy$ are available at time t^n .

The fully discrete FV version of the second-order AP scheme (3.24) reads as

$$\bar{\mathbf{V}}_{j,k}^* = \bar{\mathbf{V}}_{j,k}^n - \Delta t \mathcal{R}_{j,k}^n - \Delta t \mathcal{L}_{j,k}^{n,*}, \quad (3.35)$$

$$\bar{\mathbf{V}}_{j,k}^{n+1} = \bar{\mathbf{V}}_{j,k}^n - \frac{\Delta t}{2} [\mathcal{R}_{j,k}^n + \mathcal{R}_{j,k}^*] - \frac{\Delta t}{2} [\mathcal{L}_{j,k}^{n,n} - \mathcal{L}_{j,k}^{*,*}] - \Delta t \mathcal{L}_{j,k}^{*,n+1}, \quad (3.36)$$

where $\mathcal{R}_{j,k}^n$ and $\mathcal{R}_{j,k}^*$ are obtained using the PCCU discretization from [3, 15], which is a low-dissipation generalization of the PCCU discretization from [12], while $\mathcal{L}_{j,k}^{n,*}$, $\mathcal{L}_{j,k}^{n,n}$, $\mathcal{L}_{j,k}^{*,*}$, and $\mathcal{L}_{j,k}^{*,n+1}$ are obtained using central differences. In what follows, for the sake of brevity, we provide details on $\mathcal{R}_{j,k}^n$ and $\mathcal{L}_{j,k}^{n,*}$ only, whereas the remaining discretizations are obtained in a similar manner.

We begin with

$$\begin{aligned} \mathcal{R}_{j,k}^n := & \frac{1}{\Delta x} \left[\tilde{\mathcal{F}}_{j+\frac{1}{2},k}^n - \tilde{\mathcal{F}}_{j-\frac{1}{2},k}^n - \tilde{B}_{j,k}^n - \frac{\tilde{a}_{j-\frac{1}{2},k}^{+,n} \tilde{B}_{\Psi,j-\frac{1}{2},k}^n}{\tilde{a}_{j-\frac{1}{2},k}^{+,n} - \tilde{a}_{j-\frac{1}{2},k}^{-,n}} + \frac{\tilde{a}_{j+\frac{1}{2},k}^{-,n} \tilde{B}_{\Psi,j+\frac{1}{2},k}^n}{\tilde{a}_{j+\frac{1}{2},k}^{+,n} - \tilde{a}_{j+\frac{1}{2},k}^{-,n}} \right] \\ & + \frac{1}{\Delta y} \left[\tilde{\mathcal{G}}_{j,k+\frac{1}{2}}^n - \tilde{\mathcal{G}}_{j,k-\frac{1}{2}}^n - \tilde{C}_{j,k}^n - \frac{\tilde{b}_{j,k-\frac{1}{2}}^{+,n} \tilde{C}_{\Psi,j,k-\frac{1}{2}}^n}{\tilde{b}_{j,k-\frac{1}{2}}^{+,n} - \tilde{b}_{j,k-\frac{1}{2}}^{-,n}} + \frac{\tilde{b}_{j,k+\frac{1}{2}}^{-,n} \tilde{C}_{\Psi,j,k+\frac{1}{2}}^n}{\tilde{b}_{j,k+\frac{1}{2}}^{+,n} - \tilde{b}_{j,k+\frac{1}{2}}^{-,n}} \right], \end{aligned} \quad (3.37)$$

where $\tilde{\mathcal{F}}_{j+\frac{1}{2},k}^n$ and $\tilde{\mathcal{G}}_{j,k+\frac{1}{2}}^n$ are the CU numerical fluxes

$$\begin{aligned} \tilde{\mathcal{F}}_{j+\frac{1}{2},k}^n &:= \frac{\tilde{a}_{j+\frac{1}{2},k}^{+,n} \tilde{F}(\mathbf{V}_{j+\frac{1}{2},k}^{-,n}) - \tilde{a}_{j+\frac{1}{2},k}^{-,n} \tilde{F}(\mathbf{V}_{j+\frac{1}{2},k}^{+,n})}{\tilde{a}_{j+\frac{1}{2},k}^{+,n} - \tilde{a}_{j+\frac{1}{2},k}^{-,n}} + \frac{\tilde{a}_{j+\frac{1}{2},k}^{+,n} \tilde{a}_{j+\frac{1}{2},k}^{-,n}}{\tilde{a}_{j+\frac{1}{2},k}^{+,n} - \tilde{a}_{j+\frac{1}{2},k}^{-,n}} \left(\mathbf{V}_{j+\frac{1}{2},k}^{+,n} - \mathbf{V}_{j+\frac{1}{2},k}^{-,n} - \delta \mathbf{V}_{j+\frac{1}{2},k}^n \right), \\ \tilde{\mathcal{G}}_{j,k+\frac{1}{2}}^n &:= \frac{\tilde{b}_{j,k+\frac{1}{2}}^{+,n} \tilde{G}(\mathbf{V}_{j,k+\frac{1}{2}}^{-,n}) - \tilde{b}_{j,k+\frac{1}{2}}^{-,n} \tilde{G}(\mathbf{V}_{j,k+\frac{1}{2}}^{+,n})}{\tilde{b}_{j,k+\frac{1}{2}}^{+,n} - \tilde{b}_{j,k+\frac{1}{2}}^{-,n}} + \frac{\tilde{b}_{j,k+\frac{1}{2}}^{+,n} \tilde{b}_{j,k+\frac{1}{2}}^{-,n}}{\tilde{b}_{j,k+\frac{1}{2}}^{+,n} - \tilde{b}_{j,k+\frac{1}{2}}^{-,n}} \left(\mathbf{V}_{j,k+\frac{1}{2}}^{+,n} - \mathbf{V}_{j,k+\frac{1}{2}}^{-,n} - \delta \mathbf{V}_{j,k+\frac{1}{2}}^n \right), \end{aligned}$$

and $\mathbf{V}_{j+\frac{1}{2},k}^{\pm,n}$ and $\mathbf{V}_{j,k+\frac{1}{2}}^{\pm,n}$ are reconstructed values of $\mathbf{V}$ at the midpoints of the cell interfaces, $\delta \mathbf{V}_{j+\frac{1}{2},k}^n$ and $\delta \mathbf{V}_{j,k+\frac{1}{2}}^n$ are “built-in” anti-diffusion terms, and $\tilde{a}_{j+\frac{1}{2},k}^{\pm,n}$ and $\tilde{a}_{j,k+\frac{1}{2}}^{\pm,n}$ denote the one-sided local propagation speeds of the nonstiff subsystem in the x - and y -direction, respectively.

The point values

$$\begin{aligned} \mathbf{V}_{j+\frac{1}{2},k}^{-,n} &:= \bar{\mathbf{V}}_{j,k}^n + \frac{\Delta x}{2} (\mathbf{V}_x)_{j,k}^n, & \mathbf{V}_{j+\frac{1}{2},k}^{+,n} &:= \bar{\mathbf{V}}_{j+1,k}^n - \frac{\Delta x}{2} (\mathbf{V}_x)_{j+1,k}^n, \\ \mathbf{V}_{j,k+\frac{1}{2}}^{-,n} &:= \bar{\mathbf{V}}_{j,k}^n + \frac{\Delta y}{2} (\mathbf{V}_y)_{j,k}^n, & \mathbf{V}_{j,k+\frac{1}{2}}^{+,n} &:= \bar{\mathbf{V}}_{j,k+1}^n - \frac{\Delta y}{2} (\mathbf{V}_y)_{j,k+1}^n, \end{aligned} \quad (3.38)$$

are computed using the piecewise linear reconstruction

$$\bar{\mathbf{V}}_{j,k}^n + (\mathbf{V}_x)_{j,k}^n (x - x_j) + (\mathbf{V}_y)_{j,k}^n (y - y_k), \quad (x, y) \in I_{j,k},$$

in which the slopes $(\mathbf{V}_x)_{j,k}^n$ and $(\mathbf{V}_y)_{j,k}^n$ are approximated using the generalized minmod limiter (see, e.g., [35, 40, 44]):

$$\begin{aligned} (\mathbf{V}_x)_{j,k}^n &:= \text{minmod} \left(\theta \frac{\bar{\mathbf{V}}_{j,k}^n - \bar{\mathbf{V}}_{j-1,k}^n}{\Delta x}, \frac{\bar{\mathbf{V}}_{j+1,k}^n - \bar{\mathbf{V}}_{j-1,k}^n}{2\Delta x}, \theta \frac{\bar{\mathbf{V}}_{j+1,k}^n - \bar{\mathbf{V}}_{j,k}^n}{\Delta x} \right), \\ (\mathbf{V}_y)_{j,k}^n &:= \text{minmod} \left(\theta \frac{\bar{\mathbf{V}}_{j,k}^n - \bar{\mathbf{V}}_{j,k-1}^n}{\Delta y}, \frac{\bar{\mathbf{V}}_{j,k+1}^n - \bar{\mathbf{V}}_{j,k-1}^n}{2\Delta y}, \theta \frac{\bar{\mathbf{V}}_{j,k+1}^n - \bar{\mathbf{V}}_{j,k}^n}{\Delta y} \right), \end{aligned} \quad (3.39)$$

where the minmod function, defined by

$$\text{minmod}(z_1, z_2, \dots) := \begin{cases} \min(z_1, z_2, \dots) & \text{if } z_i > 0, \forall i, \\ \max(z_1, z_2, \dots) & \text{if } z_i < 0, \forall i, \\ 0 & \text{otherwise,} \end{cases}$$

is applied in a componentwise manner. The parameter $\theta \in [1, 2]$ in (3.39) is to be chosen to adjust the amount of numerical dissipation present in the resulting scheme, with larger values of θ leading to sharper but, in general, more oscillatory solutions.

The one-sided local speeds of propagation are estimated using the smallest and largest eigenvalues of the matrices $\frac{\partial \tilde{\mathbf{F}}}{\partial \mathbf{V}}(\mathbf{V}) - \tilde{\mathbf{B}}(\mathbf{V})$ and $\frac{\partial \tilde{\mathbf{G}}}{\partial \mathbf{V}}(\mathbf{V}) - \tilde{\mathbf{C}}(\mathbf{V})$ as follows:

$$\begin{aligned} \tilde{a}_{j+\frac{1}{2},k}^{-,n} &:= \min \left\{ u_{j+\frac{1}{2},k}^{-,n} - \tilde{c}_{j+\frac{1}{2},k}^{-,n}, u_{j+\frac{1}{2},k}^{+,n} - \tilde{c}_{j+\frac{1}{2},k}^{+,n}, -\delta \right\}, \\ \tilde{a}_{j+\frac{1}{2},k}^{+,n} &:= \max \left\{ u_{j+\frac{1}{2},k}^{-,n} + \tilde{c}_{j+\frac{1}{2},k}^{-,n}, u_{j+\frac{1}{2},k}^{+,n} + \tilde{c}_{j+\frac{1}{2},k}^{+,n}, \delta \right\}, \\ \tilde{b}_{j,k+\frac{1}{2}}^{-,n} &:= \min \left\{ v_{j,k+\frac{1}{2}}^{-,n} - \tilde{c}_{j,k+\frac{1}{2}}^{-,n}, v_{j,k+\frac{1}{2}}^{+,n} - \tilde{c}_{j,k+\frac{1}{2}}^{+,n}, -\delta \right\}, \\ \tilde{b}_{j,k+\frac{1}{2}}^{+,n} &:= \max \left\{ v_{j,k+\frac{1}{2}}^{-,n} + \tilde{c}_{j,k+\frac{1}{2}}^{-,n}, v_{j,k+\frac{1}{2}}^{+,n} + \tilde{c}_{j,k+\frac{1}{2}}^{+,n}, \delta \right\}, \end{aligned} \quad (3.40)$$

where the sound speeds

$$\tilde{c}_{j+\frac{1}{2},k}^{\pm,n} := \frac{1}{\varepsilon} \sqrt{\gamma \frac{(\rho_{\max}^n - \rho_{j+\frac{1}{2},k}^{\pm,n})(p_{j+\frac{1}{2},k}^{\pm,n} - p_{\min}^n)}{\rho_{j+\frac{1}{2},k}^{\pm,n} \rho_{\max}^n}}, \quad \tilde{c}_{j,k+\frac{1}{2}}^{\pm,n} := \frac{1}{\varepsilon} \sqrt{\gamma \frac{(\rho_{\max}^n - \rho_{j,k+\frac{1}{2}}^{\pm,n})(p_{j,k+\frac{1}{2}}^{\pm,n} - p_{\min}^n)}{\rho_{j,k+\frac{1}{2}}^{\pm,n} \rho_{\max}^n}}$$

are computed using the following discrete versions of $\rho_{\max}^n$ and $p_{\min}^n$

$$\rho_{\max}^n := \max_{j,k} \bar{\rho}_{j,k}^n, \quad p_{\min}^n := \min_{j,k} \bar{p}_{j,k}^n, \quad (3.41)$$

and δ is a small positive parameter introduced to prevent divisions by 0 (we have taken $\delta := 10^{-15}$ in the numerical experiments reported in §4).

The “built-in” anti-diffusion terms are

$$\begin{aligned} \delta \mathbf{V}_{j+\frac{1}{2},k}^n &:= \text{minmod} \left(\mathbf{V}_{j+\frac{1}{2},k}^{\text{int},n} - \mathbf{V}_{j+\frac{1}{2},k}^{-,n}, \mathbf{V}_{j+\frac{1}{2},k}^{+,n} - \mathbf{V}_{j+\frac{1}{2},k}^{\text{int},n} \right), \\ \delta \mathbf{V}_{j,k+\frac{1}{2}}^n &:= \text{minmod} \left(\mathbf{V}_{j,k+\frac{1}{2}}^{\text{int},n} - \mathbf{V}_{j,k+\frac{1}{2}}^{-,n}, \mathbf{V}_{j,k+\frac{1}{2}}^{+,n} - \mathbf{V}_{j,k+\frac{1}{2}}^{\text{int},n} \right), \end{aligned}$$

where

$$\begin{aligned} \mathbf{V}_{j+\frac{1}{2},k}^{\text{int},n} &:= \frac{\tilde{a}_{j+\frac{1}{2},k}^{+,n} \mathbf{V}_{j+\frac{1}{2},k}^{+,n} - \tilde{a}_{j+\frac{1}{2},k}^{-,n} \mathbf{V}_{j+\frac{1}{2},k}^{-,n} - \tilde{\mathbf{F}}(\mathbf{V}_{j+\frac{1}{2},k}^{+,n}) + \tilde{\mathbf{F}}(\mathbf{V}_{j+\frac{1}{2},k}^{-,n})}{\tilde{a}_{j+\frac{1}{2},k}^{+,n} - \tilde{a}_{j+\frac{1}{2},k}^{-,n}}, \\ \mathbf{V}_{j,k+\frac{1}{2}}^{\text{int},n} &:= \frac{\tilde{b}_{j,k+\frac{1}{2}}^{+,n} \mathbf{V}_{j,k+\frac{1}{2}}^{+,n} - \tilde{b}_{j,k+\frac{1}{2}}^{-,n} \mathbf{V}_{j,k+\frac{1}{2}}^{-,n} - \tilde{\mathbf{G}}(\mathbf{V}_{j,k+\frac{1}{2}}^{+,n}) + \tilde{\mathbf{G}}(\mathbf{V}_{j,k+\frac{1}{2}}^{-,n})}{\tilde{b}_{j,k+\frac{1}{2}}^{+,n} - \tilde{b}_{j,k+\frac{1}{2}}^{-,n}}. \end{aligned}$$

Finally,

$$\mathcal{L}_{j,k}^{n,*} := \left(0, \frac{\bar{p}_{j+1,k}^* - \bar{p}_{j-1,k}^*}{2\Delta x \varepsilon^2 \rho_{\max}^n}, \frac{\bar{p}_{j,k+1}^* - \bar{p}_{j,k-1}^*}{2\Delta y \varepsilon^2 \rho_{\max}^n}, \gamma p_{\min}^n \nabla \cdot \bar{\mathbf{u}}_{j,k}^* \right)^\top, \quad (3.42)$$

where $\nabla \cdot \bar{\mathbf{u}}_{j,k}$ denotes the discrete divergence operator computed using second-order central differences:

$$\nabla \cdot \bar{\mathbf{u}}_{j,k} := \frac{\bar{u}_{j+1,k} - \bar{u}_{j-1,k}}{2\Delta x} + \frac{\bar{v}_{j,k+1} - \bar{v}_{j,k-1}}{2\Delta y}. \quad (3.43)$$

3.3 Semi-Discrete CU Scheme for the Conservative System

We now consider the conservative formulation (2.1)–(2.3), which can be put into the following vector form:

$$\begin{aligned} U_t + \mathbf{F}(U)_x + \mathbf{G}(U)_y &= \mathbf{0}, \quad \mathbf{U} := (\rho, \rho u, \rho v, E)^\top, \\ \mathbf{F}(U) &:= \left(\rho u, \rho u^2 + \frac{p}{\varepsilon^2}, \rho uv, u(E + p) \right)^\top, \quad \mathbf{G}(U) := \left(\rho v, \rho uv, \rho v^2 + \frac{p}{\varepsilon^2}, v(E + p) \right)^\top. \end{aligned} \quad (3.44)$$

In the semi-discrete CU scheme, the cell averages $\bar{U}_{j,k}(t) \approx \frac{1}{\Delta x \Delta y} \iint_{I_{j,k}} U(x, y, t) dx dy$ are evolved in time by numerically solving the following system of ODEs:

$$\frac{d}{dt} \bar{U}_{j,k} = -\frac{\mathcal{F}_{j+\frac{1}{2},k} - \mathcal{F}_{j-\frac{1}{2},k}}{\Delta x} - \frac{\mathcal{G}_{j,k+\frac{1}{2}} - \mathcal{G}_{j,k-\frac{1}{2}}}{\Delta y}, \quad (3.45)$$

where $\mathcal{F}_{j+\frac{1}{2},k}$ and $\mathcal{G}_{j,k+\frac{1}{2}}$ are the CU numerical fluxes from [32] defined as

$$\begin{aligned} \mathcal{F}_{j+\frac{1}{2},k} &:= \frac{a_{j+\frac{1}{2},k}^+ \mathbf{F}(U_{j+\frac{1}{2},k}^-) - a_{j+\frac{1}{2},k}^- \mathbf{F}(U_{j+\frac{1}{2},k}^+)}{a_{j+\frac{1}{2},k}^+ - a_{j+\frac{1}{2},k}^-} \\ &\quad + \frac{a_{j+\frac{1}{2},k}^+ a_{j+\frac{1}{2},k}^-}{a_{j+\frac{1}{2},k}^+ - a_{j+\frac{1}{2},k}^-} \left(U_{j+\frac{1}{2},k}^+ - U_{j+\frac{1}{2},k}^- - \delta U_{j+\frac{1}{2},k} \right), \\ \mathcal{G}_{j,k+\frac{1}{2}} &:= \frac{b_{j,k+\frac{1}{2}}^+ \mathbf{G}(U_{j,k+\frac{1}{2}}^-) - b_{j,k+\frac{1}{2}}^- \mathbf{G}(U_{j,k+\frac{1}{2}}^+)}{b_{j,k+\frac{1}{2}}^+ - b_{j,k+\frac{1}{2}}^-} \\ &\quad + \frac{b_{j,k+\frac{1}{2}}^+ b_{j,k+\frac{1}{2}}^-}{b_{j,k+\frac{1}{2}}^+ - b_{j,k+\frac{1}{2}}^-} \left(U_{j,k+\frac{1}{2}}^+ - U_{j,k+\frac{1}{2}}^- - \delta U_{j,k+\frac{1}{2}} \right). \end{aligned} \quad (3.46)$$

Here, the interface values $U_{j+\frac{1}{2},k}^\pm := U(V_{j+\frac{1}{2},k}^\pm)$ and $U_{j,k+\frac{1}{2}}^\pm := U(V_{j,k+\frac{1}{2}}^\pm)$ are computed from the reconstructed primitive variables $V_{j+\frac{1}{2},k}^\pm$ and $V_{j,k+\frac{1}{2}}^\pm$ (see §3.2) at the corresponding time level via a straightforward transformation $U(V)$ from V to U . The quantities $a_{j+\frac{1}{2},k}^\pm$ and $b_{j,k+\frac{1}{2}}^\pm$ are the one-sided local speeds of propagation for the conservative system (3.44) in the x - and y -direction, respectively. They are estimated using the largest and smallest eigenvalues of the corresponding flux Jacobians as follows:

$$\begin{aligned} a_{j+\frac{1}{2},k}^- &:= \min \left\{ u_{j+\frac{1}{2},k}^- - c_{j+\frac{1}{2},k}^-, u_{j+\frac{1}{2},k}^+ - c_{j+\frac{1}{2},k}^+, -\delta \right\}, & c_{j+\frac{1}{2},k}^\pm &:= \frac{1}{\varepsilon} \sqrt{\frac{\gamma p_{j+\frac{1}{2},k}^\pm}{\rho_{j+\frac{1}{2},k}^\pm}}, \\ a_{j+\frac{1}{2},k}^+ &:= \max \left\{ u_{j+\frac{1}{2},k}^- + c_{j+\frac{1}{2},k}^-, u_{j+\frac{1}{2},k}^+ + c_{j+\frac{1}{2},k}^+, \delta \right\}, \\ b_{j,k+\frac{1}{2}}^- &:= \min \left\{ v_{j,k+\frac{1}{2}}^- - c_{j,k+\frac{1}{2}}^-, v_{j,k+\frac{1}{2}}^+ - c_{j,k+\frac{1}{2}}^+, -\delta \right\}, & c_{j,k+\frac{1}{2}}^\pm &:= \frac{1}{\varepsilon} \sqrt{\frac{\gamma p_{j,k+\frac{1}{2}}^\pm}{\rho_{j,k+\frac{1}{2}}^\pm}}, \\ b_{j,k+\frac{1}{2}}^+ &:= \max \left\{ v_{j,k+\frac{1}{2}}^- + c_{j,k+\frac{1}{2}}^-, v_{j,k+\frac{1}{2}}^+ + c_{j,k+\frac{1}{2}}^+, \delta \right\}, \end{aligned} \quad (3.47)$$

where $\delta := 10^{-15}$ is used to avoid divisions by 0.

The “built-in” anti-diffusion terms are

$$\begin{aligned}\delta U_{j+\frac{1}{2},k} &:= \text{minmod}\left(U_{j+\frac{1}{2},k}^{\text{int}} - U_{j+\frac{1}{2},k}^-, U_{j+\frac{1}{2},k}^+ - U_{j+\frac{1}{2},k}^{\text{int}}\right), \\ \delta U_{j,k+\frac{1}{2}} &:= \text{minmod}\left(U_{j,k+\frac{1}{2}}^{\text{int}} - U_{j,k+\frac{1}{2}}^-, U_{j,k+\frac{1}{2}}^+ - U_{j,k+\frac{1}{2}}^{\text{int}}\right),\end{aligned}\tag{3.48}$$

with

$$\begin{aligned}U_{j+\frac{1}{2},k}^{\text{int}} &:= \frac{a_{j+\frac{1}{2},k}^+ U_{j+\frac{1}{2},k}^+ - a_{j+\frac{1}{2},k}^- U_{j+\frac{1}{2},k}^- - \mathbf{F}(U_{j+\frac{1}{2},k}^+) + \mathbf{F}(U_{j+\frac{1}{2},k}^-)}{a_{j+\frac{1}{2},k}^+ - a_{j+\frac{1}{2},k}^-}, \\ U_{j,k+\frac{1}{2}}^{\text{int}} &:= \frac{b_{j,k+\frac{1}{2}}^+ U_{j,k+\frac{1}{2}}^+ - b_{j,k+\frac{1}{2}}^- U_{j,k+\frac{1}{2}}^- - \mathbf{G}(U_{j,k+\frac{1}{2}}^+) + \mathbf{G}(U_{j,k+\frac{1}{2}}^-)}{b_{j,k+\frac{1}{2}}^+ - b_{j,k+\frac{1}{2}}^-}.\end{aligned}\tag{3.49}$$

Note that most of the indexed quantities in the semi-discrete setting above are time-dependent, but we have omitted this dependence to ease the notation.

Finally, the system of ODEs (3.45) has to be integrated in time using an appropriate ODE solver. Its solution is performed simultaneously with one of the primitive systems using the explicit counterpart of the SI-DeC scheme, and a post-processing is performed at each stage, as explained in §3.4.

3.4 Implementation Details

In our DF-FV approach, the solutions of the primitive and conservative systems are evolved simultaneously according to the following algorithm.

- **Step 1 (Compute $\bar{\rho}_{j,k}^*$).** We use the ρ -equation in (3.35) to obtain

$$\bar{\rho}_{j,k}^* = \bar{\rho}_{j,k}^n - \Delta t (\mathcal{R}^\rho)_{j,k}^n.$$

- **Step 2 (Solve the linear elliptic equation for $\bar{p}_{j,k}^*$).** We apply the discrete divergence operator (3.43) to the $\mathbf{u}$ -equations in (3.35) and substitute them into the p -equation in (3.35) to obtain the following linear system of algebraic equations for $\bar{p}_{j,k}^*$, which is a discretization of the linear elliptic equation for p^* :

$$\bar{p}_{j,k}^* - \frac{(\Delta t)^2 \gamma p_{\min}^n}{\varepsilon^2 \rho_{\max}^n} \Delta \bar{p}_{j,k}^* = \bar{p}_{j,k}^n - \Delta t (\mathcal{R}^p)_{j,k}^n - \Delta t \gamma p_{\min}^n \nabla \cdot \bar{\mathbf{u}}_{j,k}^n + (\Delta t)^2 \gamma p_{\min}^n \nabla \cdot (\mathcal{R}^{\mathbf{u}})_{j,k}^n,$$

where the discrete Laplacian $\Delta \bar{p}_{j,k}$ is defined as

$$\Delta \bar{p}_{j,k} := \frac{\bar{p}_{j-1,k} - 2\bar{p}_{j,k} + \bar{p}_{j+1,k}}{(\Delta x)^2} + \frac{\bar{p}_{j,k-1} - 2\bar{p}_{j,k} + \bar{p}_{j,k+1}}{(\Delta y)^2}.$$

- **Step 3 (Compute $\bar{\mathbf{u}}_{j,k}^*$).** Once $\bar{p}_{j,k}^*$ is available, we use the $\mathbf{u}$ -equations in (3.35) to obtain

$$\bar{\mathbf{u}}_{j,k}^* = \bar{\mathbf{u}}_{j,k}^n - \Delta t (\mathcal{R}^{\mathbf{u}})_{j,k}^n - \Delta t (\mathcal{L}^{\mathbf{u}})_{j,k}^{n,*}.$$

- **Step 4 (Compute $\bar{\mathbf{U}}_{j,k}^*$).** We perform the conservative update with the explicit counterpart of the SI-DeC scheme to obtain the solution $\bar{\mathbf{U}}_{j,k}^*$ at the intermediate stage

$$\bar{\mathbf{U}}_{j,k}^* = \bar{\mathbf{U}}_{j,k}^n - \Delta t \left[\frac{\mathcal{F}_{j+\frac{1}{2},k}^n - \mathcal{F}_{j-\frac{1}{2},k}^n}{\Delta x} + \frac{\mathcal{G}_{j,k+\frac{1}{2}}^n - \mathcal{G}_{j,k-\frac{1}{2}}^n}{\Delta y} \right],\tag{3.50}$$

and then post-process the primitive solution by replacing $\bar{\mathbf{V}}_{j,k}^*$ with $r(\mathbf{V}(\bar{\mathbf{U}}_{j,k}^*), \bar{\mathbf{V}}_{j,k}^*)$; see §3.5.

- **Step 5 (Compute $\bar{\rho}_{j,k}^{n+1}$).** We solve the ρ -equation in (3.36) to obtain

$$\bar{\rho}_{j,k}^{n+1} = \bar{\rho}_{j,k}^n - \frac{\Delta t}{2} [(\mathcal{R}^\rho)_{j,k}^n + (\mathcal{R}^\rho)_{j,k}^*].$$

- **Step 6 (Solve the linear elliptic equation for $\bar{p}_{j,k}^{n+1}$).** We apply the discrete divergence operator (3.43) to the $\mathbf{u}$ -equations in (3.36) and substitute them into the p -equation in (3.36) to obtain the following linear system of algebraic equations for $\bar{p}_{j,k}^{n+1}$, which is a discretization of the linear elliptic equation for p^{n+1} :

$$\begin{aligned} \bar{p}_{j,k}^{n+1} - \frac{(\Delta t)^2 \gamma p_{\min}^*}{\varepsilon^2 \rho_{\max}^*} \Delta \bar{p}_{j,k}^{n+1} &= \bar{p}_{j,k}^n - \frac{\Delta t}{2} [(\mathcal{R}^p)_{j,k}^n + (\mathcal{R}^p)_{j,k}^*] - \frac{\Delta t}{2} [(\mathcal{L}^p)_{j,k}^{n,n} - (\mathcal{L}^p)_{j,k}^{*,*}] \\ &\quad - \Delta t \gamma p_{\min}^* \nabla \cdot \bar{\mathbf{u}}_{j,k}^n + \frac{(\Delta t)^2 \gamma p_{\min}^*}{2} \nabla \cdot [(\mathcal{R}^u)_{j,k}^n + (\mathcal{R}^u)_{j,k}^*] - \frac{(\Delta t)^2 \gamma p_{\min}^*}{2} \nabla \cdot [(\mathcal{L}^u)_{j,k}^{n,n} - (\mathcal{L}^u)_{j,k}^{*,*}]. \end{aligned}$$

- **Step 7 (Compute $\bar{\mathbf{u}}_{j,k}^{n+1}$).** Once $\bar{p}_{j,k}^{n+1}$ is available, we compute

$$\bar{\mathbf{u}}_{j,k}^{n+1} = \bar{\mathbf{u}}_{j,k}^n - \frac{\Delta t}{2} [(\mathcal{R}^u)_{j,k}^n + (\mathcal{R}^u)_{j,k}^*] - \frac{\Delta t}{2} [(\mathcal{L}^u)_{j,k}^{n,n} - (\mathcal{L}^u)_{j,k}^{*,*}] - \Delta t (\mathcal{L}^u)_{j,k}^{*,n+1}.$$

- **Step 8 (Compute $\bar{\mathbf{U}}_{j,k}^{n+1}$).** Finally, we use the explicit part of the SI-DeC scheme to evaluate

$$\begin{aligned} \bar{\mathbf{U}}_{j,k}^{n+1} &= \bar{\mathbf{U}}_{j,k}^n - \frac{\Delta t}{2} \left[\frac{\mathcal{F}_{j+\frac{1}{2},k}^n - \mathcal{F}_{j-\frac{1}{2},k}^n}{\Delta x} + \frac{\mathcal{F}_{j+\frac{1}{2},k}^* - \mathcal{F}_{j-\frac{1}{2},k}^*}{\Delta x} \right] \\ &\quad - \frac{\Delta t}{2} \left[\frac{\mathcal{G}_{j,k+\frac{1}{2}}^n - \mathcal{G}_{j,k-\frac{1}{2}}^n}{\Delta y} + \frac{\mathcal{G}_{j,k+\frac{1}{2}}^* - \mathcal{G}_{j,k-\frac{1}{2}}^*}{\Delta y} \right]. \end{aligned} \tag{3.51}$$

and then post-process the primitive solution by replacing $\bar{\mathbf{V}}_{j,k}^{n+1}$ with $r(\mathbf{V}(\bar{\mathbf{U}}_{j,k}^{n+1}), \bar{\mathbf{V}}_{j,k}^{n+1})$; see §3.5.

We recall that the interface values of $\mathbf{U}$ needed for the computation of the numerical fluxes in (3.50) and (3.51) are obtained from the reconstructed primitive variables $\mathbf{V}$ at the corresponding time levels. Note that the conservative updates are, as a matter of fact, explicit, since they are performed using the explicit part of the SI-DeC scheme.

Remark 3.3 *It should be observed that some parts of the reported algorithm can be parallelized, for example, Steps 1 and 2 or Steps 5 and 6. Furthermore, Steps 1–3 can be performed in parallel with Step 4 before applying the post-processing, as well as Steps 5–7 with Step 8.*

Remark 3.4 *It should be observed that working with two sets of variables induces a computational overhead compared to approaches based on a single formulation. However, the computational cost is not doubled. For example, the same reconstructed cell interface values are shared by both the $\mathbf{V}$ - and $\mathbf{U}$ -solutions, and therefore no additional reconstruction procedure is required. Moreover, the evolution of the $\mathbf{U}$ -solution is performed explicitly and is therefore computationally less demanding than the SI evolution of the $\mathbf{V}$ -solution.*

3.5 Post-Processing

As mentioned in §3.4, upon completion of Steps 4 and 8, we replace $\bar{\mathbf{V}}_{j,k}^*$ with $r(\mathbf{V}(\bar{\mathbf{U}}_{j,k}^*), \bar{\mathbf{V}}_{j,k}^*)$ and $\bar{\mathbf{V}}_{j,k}^{n+1}$ with $r(\mathbf{V}(\bar{\mathbf{U}}_{j,k}^{n+1}), \bar{\mathbf{V}}_{j,k}^{n+1})$, respectively. The function r is selected based on the following considerations. Since the $\mathbf{V}$ -solution is AP but nonconservative, and the $\mathbf{U}$ -solution is conservative but non-AP, we use their convex combination with coefficients dependent on ε , leveraging the AP SI method in the low-Mach-number regime and the sharp conservative CU scheme in the moderate- and high-Mach-number regimes—thus ensuring accuracy, stability, and physical consistency across all flow regimes. Specifically, we select the following replacement function r :

$$r(\mathbf{V}(\bar{\mathbf{U}}_{j,k}), \bar{\mathbf{V}}_{j,k}) = (1 - s(\varepsilon)) \mathbf{V}(\bar{\mathbf{U}}_{j,k}) + s(\varepsilon) \bar{\mathbf{V}}_{j,k}, \quad \forall j, k, \quad (3.52)$$

where s is a suitable switching function, which is supposed to be increasing, continuous, and satisfy $s(1) = 0$ and $s(0) = 1$. Moreover, in the high-Mach-number regime, s should be ~ 0 so that the primitive variables $\mathbf{V}$ are almost completely overwritten by $\mathbf{V}(\mathbf{U})$, while, in the low-Mach-number regime, s should be ~ 1 so that the primitive variables $\mathbf{V}$ stay almost unchanged. For intermediate values of ε , a smooth transition between 1 and 0 is expected.

4 Numerical Examples

In this section, we verify the accuracy and robustness of the proposed AP scheme on a variety of numerical examples across different values of ε . In all of the numerical examples, we:

- Take the minmod parameter $\theta = 1.3$;
- Adaptively select time steps based on the time-step restriction (3.15) for the nonstiff part of the primitive system;
- Set $\gamma = 1.4$ (except for Example 1, in which $\gamma = 2$);
- Modify (3.5) to

$$\rho_{\max} = \max_{(x,y) \in \Omega} \rho + \varepsilon^4, \quad p_{\min} = \min_{(x,y) \in \Omega} p - \varepsilon^4. \quad (4.1)$$

Notice that this modification has almost no impact in the low-Mach-number regime, but it aims at adding more upwinding and thus improving the stability property of the resulting AP scheme when ε is large;

- Choose the following switching function:

$$s(\varepsilon) = \begin{cases} 1 - \varepsilon^\alpha, & 0 < \varepsilon \leq \varepsilon_0, \\ \exp\left(1 - \frac{1}{1 - \left(\frac{\varepsilon - \varepsilon_0}{\varepsilon_1 - \varepsilon_0}\right)^2}\right) [(1 - \varepsilon_0^\alpha) - (1 - \varepsilon_1)^\alpha] + (1 - \varepsilon_1)^\alpha, & \varepsilon_0 < \varepsilon < \varepsilon_1, \\ (1 - \varepsilon)^\alpha, & \varepsilon_1 \leq \varepsilon \leq 1, \end{cases}$$

where ε_0 , ε_1 , and α are positive constants taken to be $\varepsilon_0 = 0.15$, $\varepsilon_1 = 0.4$, and $\alpha = 14$ in all of the numerical examples below. This switching function is plotted in Figure 4.1.

Remark 4.1 *Modification (4.1) of the wave speeds is used to improve the handling of discontinuous solutions for large (intermediate)-Mach-number regimes. However, we would like to emphasize that it vanishes as $\varepsilon \rightarrow 0$ and therefore does not affect either the consistency of the method or the AP analysis presented in the paper.*

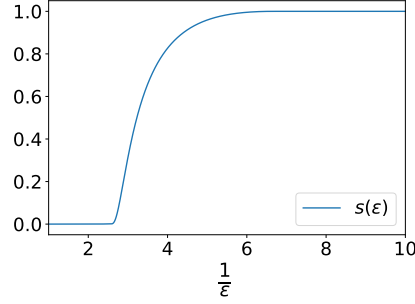


Figure 4.1: Switching function $s(\varepsilon)$ plotted with respect to $1/\varepsilon$.

Remark 4.2 *We stress that the same switching function, $s(\varepsilon)$, together with the same values of the associated parameters, has been used in all of the numerical examples reported below. The obtained numerical results indicate that this choice is sufficiently robust across the considered Mach-number regimes. Further studies on the optimal design of the switching function and on the selection of its parameters may constitute interesting directions for future research.*

Example 1—Accuracy Test for Low-Mach-Number Smooth Vortex

In this example taken from [49], we consider a smooth, unsteady Mach dependent vortex over the computational domain $[-10, 10] \times [-10, 10]$ subject to the periodic boundary conditions. The analytical solution is given, modulo the periodicity, by

$$\begin{aligned} \rho(\mathbf{x}_r) &= 1 - \frac{\varepsilon^2}{16\pi^2} e^{1-\|\mathbf{x}_r\|_2^2}, \quad u(\mathbf{x}_r) = 1 - \frac{\varepsilon y_r}{2\pi} e^{\frac{1-\|\mathbf{x}_r\|_2^2}{2}}, \quad v(\mathbf{x}_r) = 1 + \frac{\varepsilon x_r}{2\pi} e^{\frac{1-\|\mathbf{x}_r\|_2^2}{2}}, \\ E(\mathbf{x}_r) &= 1 + \varepsilon^2 \left[\rho^2(\mathbf{x}_r) + \frac{\rho(\mathbf{x}_r)}{2} (u^2(\mathbf{x}_r) + v^2(\mathbf{x}_r)) \right], \end{aligned}$$

where $\mathbf{x}_r(x, y, t) = (x_r, y_r)^\top := (x - t, y - t)^\top$.

We take the CFL number $K_{\text{CFL}} = 0.475$ and compute the numerical solution until the final time $t = 0.1$ on a series of uniform $N \times N$ meshes with $N = 64, 128, 256$, and 512 for $\varepsilon = 1, 0.1, 0.01$, and 0.001 . We study the convergence in terms of the L^1 -errors, defined for each scalar component V of the $\mathbf{V}$ -solution at the final time as

$$\Delta x \Delta y \sum_{j,k} |V_{j,k}(t = 0.1) - V(x_j, y_k, t = 0.1)|,$$

where $V(x, y, t)$ denotes the corresponding component of the exact solution. The obtained errors are reported in Figure 4.2, showing that the expected second-order convergence rate has been achieved in all variables for all considered ε . One can also observe that, for fixed mesh refinement, the error decreases for decreasing ε as a result of the convergence of the analytical solution to the incompressible limit $(\rho, u, v, p)^\top = (1, 1, 1, 1)^\top$ and of the AP character of the proposed AP DF-FV scheme.

Example 2—Gresho Vortex

This example was introduced in [25] and, since then, it has been widely used as a common benchmark to numerically validate the AP property. We consider a steady vortex over the computational

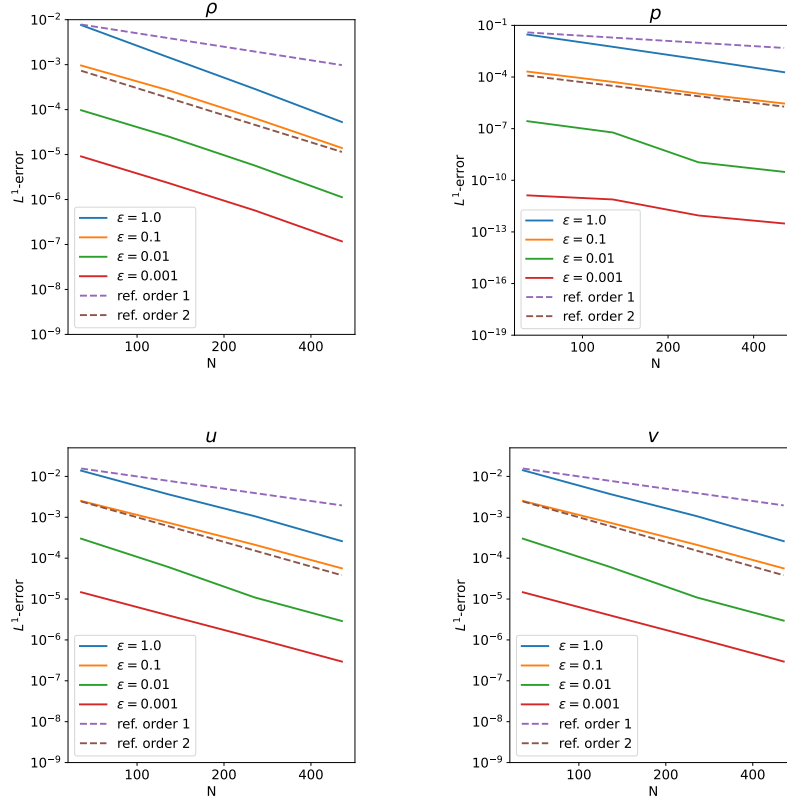


Figure 4.2: Example 1: Convergence analysis.

domain $[0, 1] \times [0, 1]$ subject to the periodic boundary conditions. At any time t , the shape of the vortex is given by

$$\rho(r) \equiv 1, \quad u(r) = -\frac{y_r}{r}\psi(r), \quad v(r) = \frac{x_r}{r}\psi(r),$$

$$p(r) = \begin{cases} 1 + 12.5\varepsilon^2 r^2, & r < 0.2, \\ 1 + \varepsilon^2 (4 \ln(5r) + 4 - 20r + 12.5r^2), & 0.2 \leq r < 0.4, \\ 1 + \varepsilon^2 (4 \ln 2 - 2), & r \geq 0.4, \end{cases}$$

where

$$x_r := x - 0.5, \quad y_r := y - 0.5, \quad r := \sqrt{x_r^2 + y_r^2}, \quad \psi(r) := \begin{cases} 5r, & r < 0.2, \\ 2 - 5r, & 0.2 \leq r < 0.4, \\ 0, & r \geq 0.4. \end{cases}$$

We take the CFL number $K_{\text{CFL}} = 0.475$ and compute the numerical solution until the final time $t = 1$ on a uniform 128×128 mesh for $\varepsilon = 10^{-\alpha}$ with $\alpha = 1, \dots, 6$, and report the obtained local Mach number, defined as $\|\mathbf{u}\|_2 / \sqrt{\gamma}$, in Figure 4.3 along with its initial distribution. According to what is expected due to the AP feature of the scheme, the shape of the vortex is preserved and no evident dependency on ε can be observed.

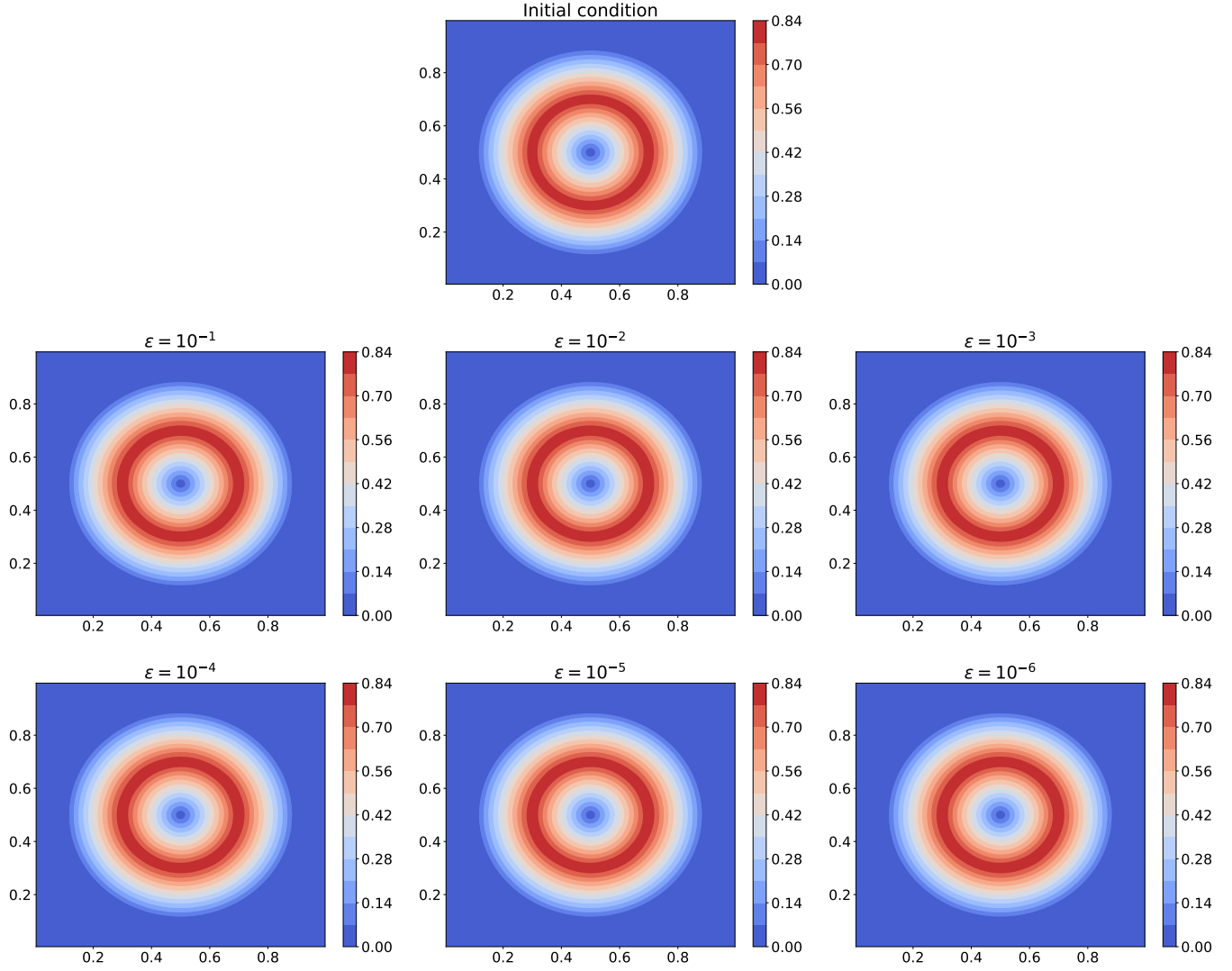


Figure 4.3: Example 2: Initial local Mach number independently of ε (top) and local Mach number at $t = 1$ for different values of ε .

Example 3—Baroclinic Vorticity Generation

In this example taken from [41], we consider a low-Mach-number flow with $\varepsilon = 0.05$ involving an acoustic wave, which moves within two density layers in the computational domain $[-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}] \times [0, \frac{2}{5\varepsilon}]$ subject to the periodic boundary conditions. The initial conditions are

$$\rho(x, y, 0) = 1 + \frac{\varepsilon}{2000}[1 + \cos(\varepsilon\pi x)] + 4.5\varepsilon y - \begin{cases} 0, & 0 \leq y \leq \frac{1}{5\varepsilon}, \\ 1.8, & \text{otherwise,} \end{cases}$$

$$u(x, y, 0) = \frac{\sqrt{\gamma}}{2}[1 + \cos(\varepsilon\pi x)], \quad v(x, y, 0) \equiv 0, \quad p(x, y, 0) = 1 + \frac{\varepsilon\gamma}{2}[1 + \cos(\varepsilon\pi x)].$$

It should be observed that the initial density discontinuity is not accompanied by a corresponding pressure discontinuity. Therefore, the initial data are not expected to generate any strong propagating compressible waves. On the other hands, the acoustic wave induces different accelerations

in the two density layers, which results in rotational excitation and in the formation of a long-wavelength sinusoidal shear layer. Due to the interaction with the acoustic wave, such a shear layer becomes unstable, and several Kelvin-Helmholtz-type unstable structures originate from it.

The numerical solution is computed with the CFL number $K_{\text{CFL}} = 0.475$ until the final time $t = 20$ on a 800×160 uniform mesh. The density at times $t = 0, 10$, and 20 is plotted in Figure 4.4. Since the solution develops instabilities, no strong convergence is expected in this example; see [49]. One can, however, observe that the underlying physics is correctly captured.

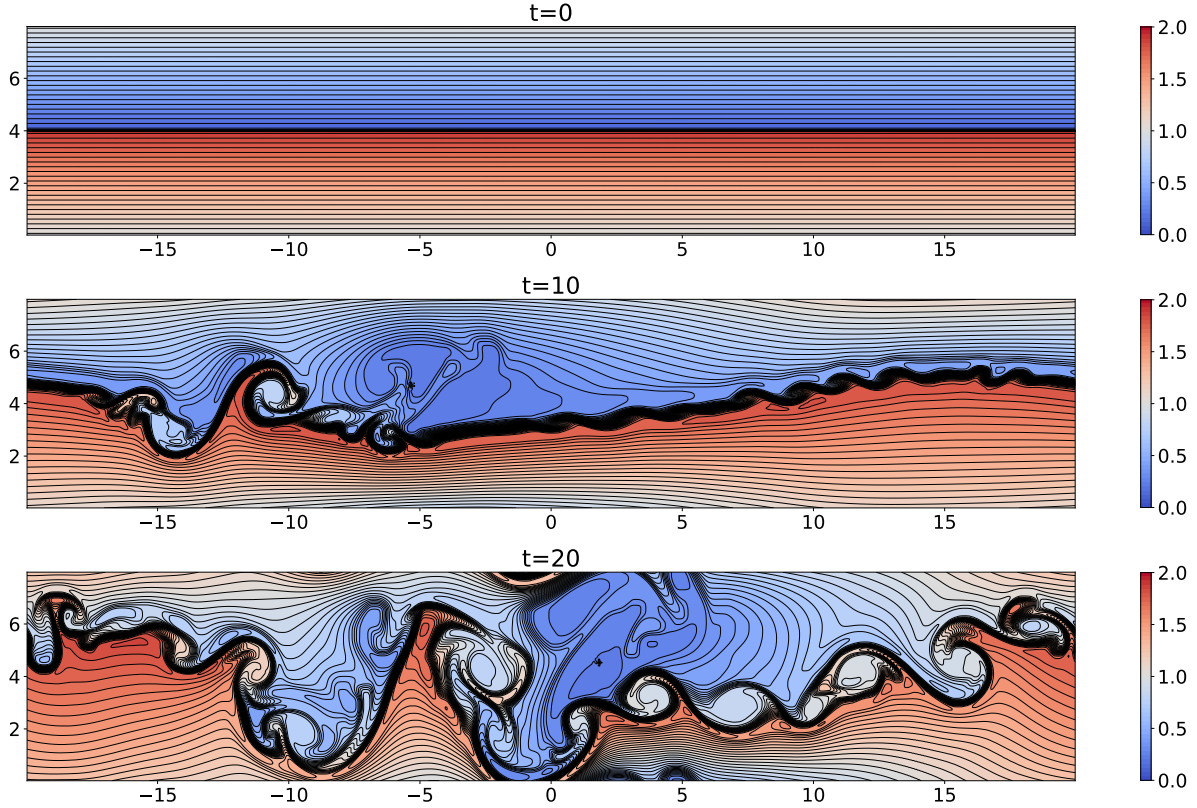


Figure 4.4: Example 3: Density at different times.

Example 4—Double Shear Layer Problem

In the following test case, originally introduced in [8] for the incompressible Navier–Stokes equations and subsequently adopted in, e.g., [10, 48, 49] in the context of compressible Euler equations in the low-Mach-number regime, a shear layer develops, and the AP property of the proposed scheme can be assessed. In particular, we would like to check whether the scheme maintains its consistency for small values of ε , that is, in the almost incompressible regime.

The initial conditions,

$$\rho(x, y, 0) \equiv \frac{\pi}{15}, \quad u(x, y, 0) = \begin{cases} \tanh\left[15\left(\frac{y}{\pi} - \frac{1}{2}\right)\right], & y \leq \pi, \\ \tanh\left[15\left(\frac{3}{2} - \frac{y}{\pi}\right)\right], & \text{otherwise,} \end{cases} \quad v(x, y, 0) = 0.05 \sin x,$$

$$p(x, y, 0) \equiv \frac{1}{\gamma},$$

are prescribed in the computational domain $[0, 2\pi] \times [0, 2\pi]$ subject to the periodic boundary conditions. The initial vorticity $\omega := v_x - u_y$, where the derivatives are approximated using second-order central differences, is plotted in Figure 4.5.

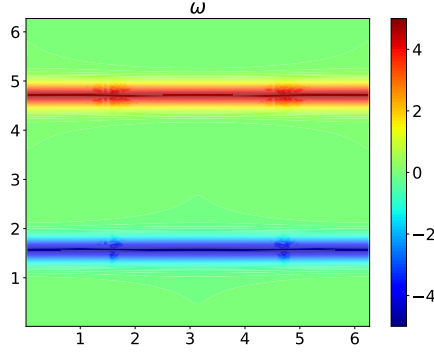


Figure 4.5: Example 4: Initial vorticity.

We compute the numerical solutions for $\varepsilon = 10^{-\alpha}$ with $\alpha = 1, \dots, 6$ until the final time $t = 10$ on a 256×256 uniform mesh using $K_{\text{CFL}} = 0.1$. Figures 4.6 and 4.7 display the vorticity at times $t = 6$ and $t = 10$, respectively, for different ε . The obtained results are consistent with those reported in [10, 49]. Moreover, no macroscopic dependence on ε is observed, providing further evidence of the AP property of the proposed DF-FV scheme.

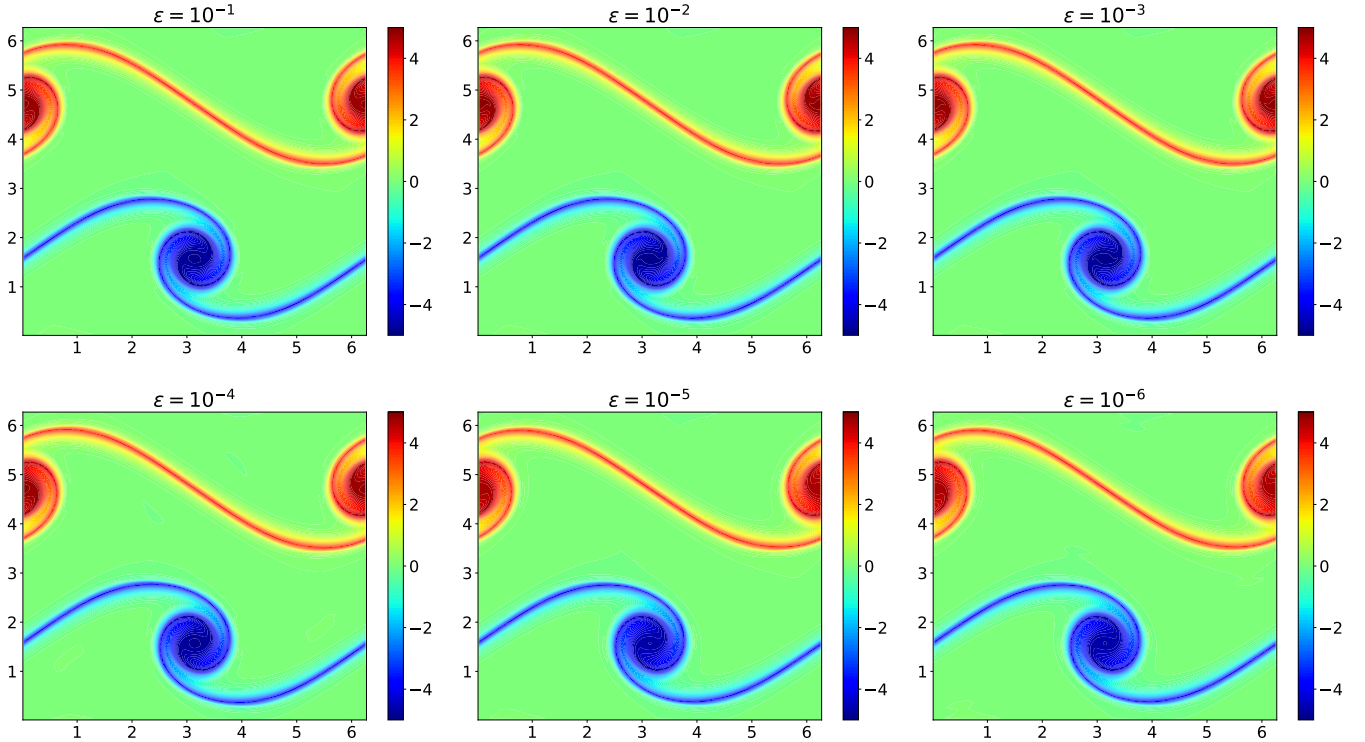
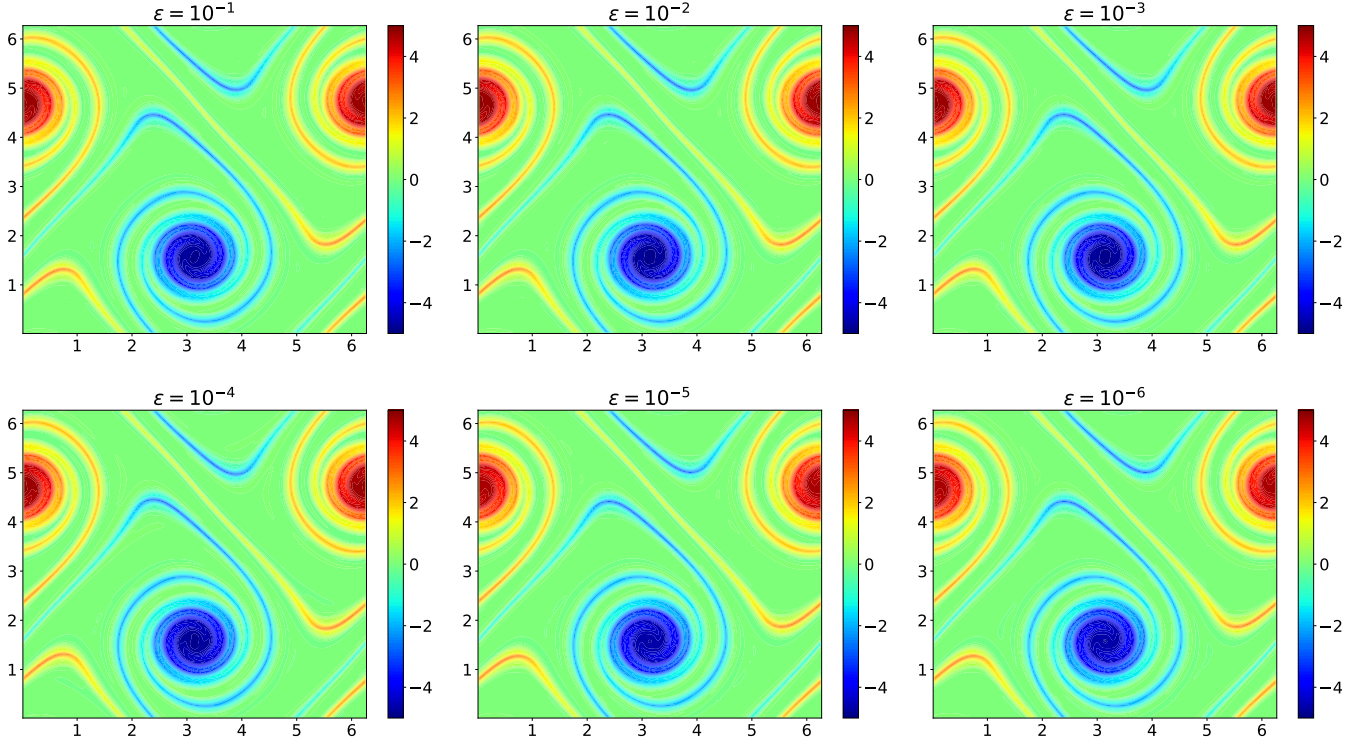
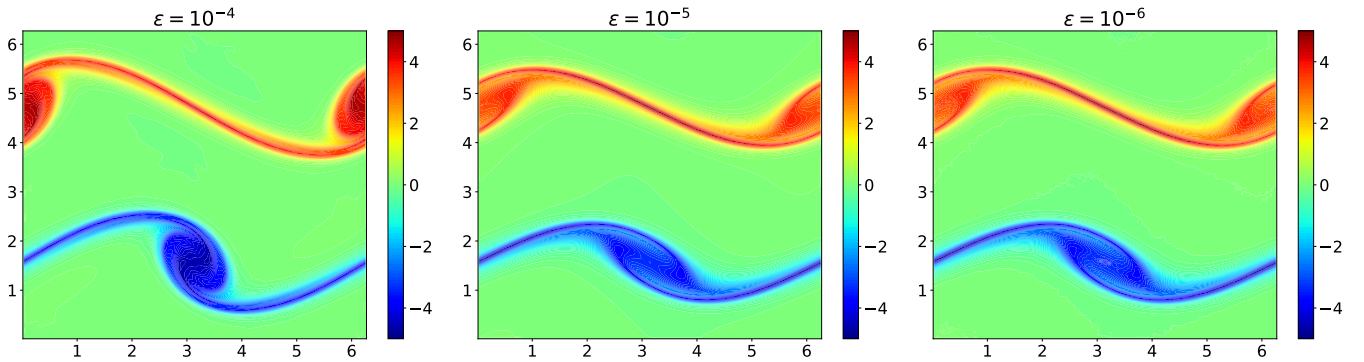


Figure 4.6: Example 4: Vorticity at $t = 6$ for different values of ε . $K_{\text{CFL}} = 0.1$.

Figure 4.7: Example 4: The same as in Figure 4.6, but at $t = 10$.

We remark that the simulations remain stable for larger CFL numbers. However, the use of larger K_{CFL} may lead to a noticeable increase in the amount of the numerical diffusion for very small values of $\varepsilon \lesssim 10^{-3}$. To illustrate this, we recompute the solution with $K_{\text{CFL}} = 0.475$ for $\varepsilon = 10^{-4}$, 10^{-5} , and 10^{-6} and plot the obtained results (for $t = 6$) in Figure 4.8. As one can clearly see, the numerical solution is now substantially more diffusive compared with those reported in the bottom 'row of Figure 4.6.

Figure 4.8: Example 4: The same as in Figure 4.6, but for $K_{\text{CFL}} = 0.475$ (left).

Example 5—Explosion Problem

In the last numerical example, we consider an explosion problem taken from [46]. The initial data,

$$(\rho, u, v, p)(x, y, 0) = \begin{cases} (1, 0, 0, 1), & \sqrt{x^2 + y^2} < 0.4, \\ (0.125, 0, 0, 0.1), & \text{otherwise,} \end{cases}$$

are prescribed in the computational domain $[-1, 1] \times [-1, 1]$ subject to the free boundary conditions.

The main objective of this test is to verify that the proposed AP DF-FV scheme remains accurate and stable, also in the high-Mach-number regime, in which strong shocks and contact discontinuities may be present. To this end, we perform simulations for several values of ε . For $\varepsilon = 0.9, 0.6$, and 0.3 the final times are $t = 0.2, 0.15$, and 0.08 , respectively. The surface plots of the density ρ computed on a uniform mesh with 400×400 cells using $K_{\text{CFL}} = 0.475$ are reported in Figure 4.9, where one can see that the obtained solutions are oscillation-free and their nonsmooth features are accurately resolved for all values of ε .

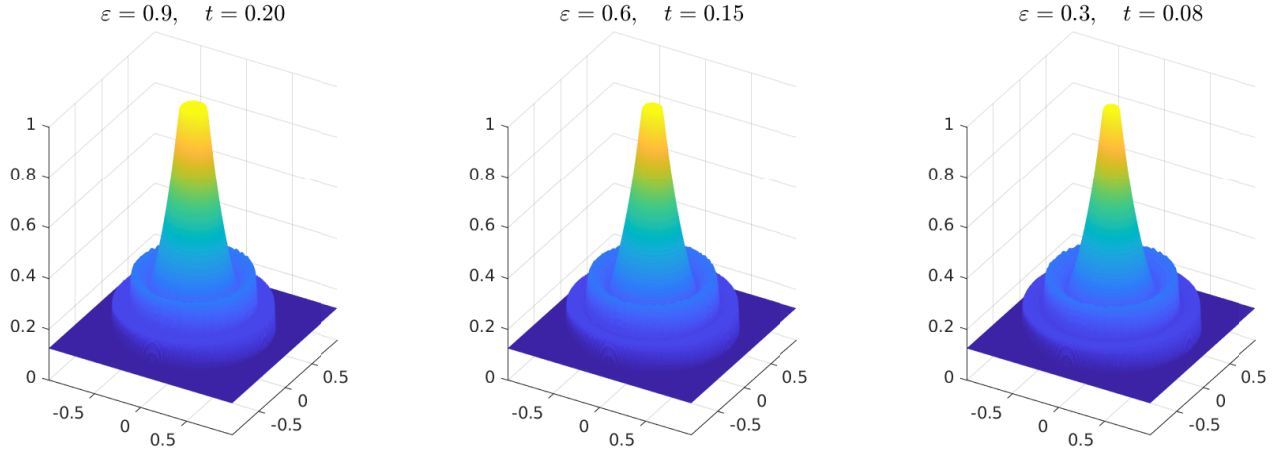


Figure 4.9: Example 5: Surface plot of density for different values of ε and corresponding times.

To further assess the correctness of the computed solutions, we plot their one-dimensional (1-D) slices along the diagonal $y = x$ in Figure 4.10 together with the corresponding slices of the reference solution, which was obtained using the second-order semi-discrete CU scheme from [32] on a much finer mesh with 2000×2000 cells using the CFL number 0.2 and the three-stage third-order strong stability preserving (SSP) Runge-Kutta method [23, 24]. As one can clearly see, the computed solutions show a perfect agreement with the reference ones, and the discontinuities locations are correctly captured.

We stress that for the large (intermediate) Mach numbers considered in this example, the $\mathbf{U}$ -solutions are reliable and plays a crucial role in ensuring a correct handling of the discontinuities through the post-processing. We omit the plots of the $\mathbf{U}$ -solutions because they are visually indistinguishable from the ones reported. In this situation, one can see how crucial is the role of the post-processing. Without it, the nonconservative evolution of the primitive variables would lead to incorrect solutions. While in the case when $\varepsilon = 0.9$ or 0.6 , the weight $s(\varepsilon)$ in (3.52) is very close to 1 and thus a post-processed solution is basically the conservative one, $\varepsilon = 0.3$ corresponds to a truly intermediate-Mach-number regime and it is instructive to look at the unreliable $\mathbf{V}$ -solution

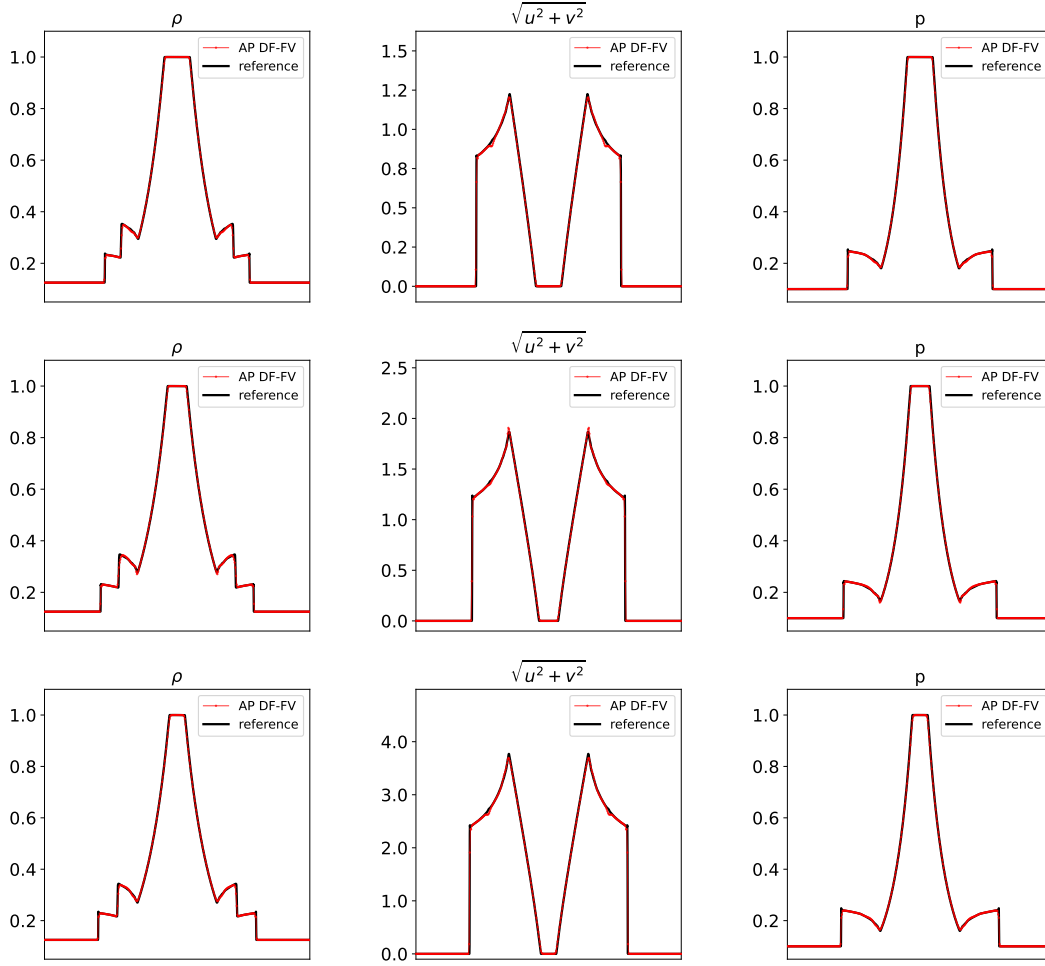


Figure 4.10: Example 5: 1-D slices of the computed solutions along $y = x$ for different values of ε and at different times: $\varepsilon = 0.9$, $t = 0.2$ (top row), $\varepsilon = 0.6$, $t = 0.15$ (middle row), and $\varepsilon = 0.3$, $t = 0.08$ (bottom row).

obtained by the same scheme but without the post-processing in this case. This $\mathbf{V}$ -solution is reported in Figure 4.11, where one can see a slight difference in the location of the discontinuities as well as other small numerical artifacts. Furthermore, we emphasize that in order to run the “solely” $\mathbf{V}$ -simulation, it was necessary to lower K_{CFL} down to 0.01.

Remark 4.3 We stress that in this example, both terms in the numerator in (3.10) will vanish if $\rho_{\max}$ and $p_{\min}$ are computed using (3.5). While the modification (4.1) ensures positivity of $\tilde{c}$, the resulting time steps might still be too big to guarantee stability of the AP DF-FV method. Therefore, we set $\Delta t = 10^{-4}$ for the first 10 time steps for the simulations involving $\varepsilon = 0.6$ and 0.3.

Remark 4.4 Let us remark that discontinuities are unlikely to occur in low-Mach-number flows. Consequently, the above tests with $\varepsilon = 0.6$ and 0.3 should be regarded as “academic” and are primarily intended to demonstrate that the proposed AP DF-FV scheme is capable of handling discontinuities even in the low-Mach-number regime.

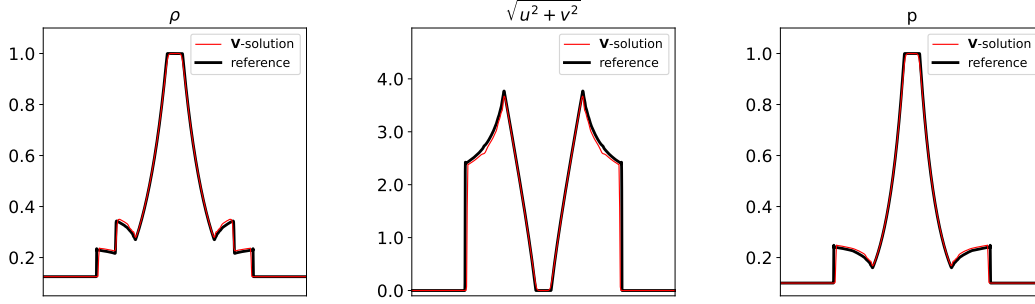


Figure 4.11: Example 5: 1-D slices of the “solely” V -solution along $y = x$ for $\varepsilon = 0.3$, computed at $t = 0.08$ without post-processing.

5 Conclusion

We have presented a novel asymptotic-preserving (AP) numerical method for the compressible Euler equations that is effective across all Mach-number regimes, including the low-Mach-number one, where standard explicit schemes become inefficient. The key idea is a new hyperbolic splitting, inspired by the flux-splitting approach introduced in [26]. The new splitting is applied to a primitive (nonconservative) formulation of the Euler equations, which enables one to design an efficient semi-implicit (SI) time discretization. Our splitting isolates stiff linear terms, which are discretized semi-implicitly: this leads to a well-posed linear elliptic problem, which ensures the AP property of the resulting scheme.

To overcome the well-known difficulties associated with the use of nonconservative formulations in the presence of discontinuities, we implement the proposed AP scheme within the recently introduced dual formulation framework [3, 14]. In this approach, the conservative and primitive systems are solved simultaneously, and their resulting solutions are post-processed to ensure the correct capturing of discontinuities while retaining the AP property of the primitive-based SI approach.

The proposed AP dual formulation finite-volume (DF-FV) method has been thoroughly validated on several benchmarks ranging from the fully compressible to the nearly incompressible regime, demonstrating both high accuracy and robustness of the method. Future work will focus on extending the AP DF-FV framework to more complex systems and on developing higher-order spatial and temporal discretizations.

Acknowledgment: The work of A. Chertock was supported in part by NSF grant DMS-2208438. The work of A. Kurganov was supported in part by NSFC grant W2431004. The work of L. Micalizzi was supported in part by the LeRoy B. Martin, Jr. Distinguished Professorship Foundation.

References

- [1] R. ABGRALL, High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices, *J. Sci. Comput.*, 73 (2017), pp. 461–494.
- [2] R. ABGRALL, A combination of residual distribution and the active flux formulations or a new class of schemes that can combine several writings of the same hyperbolic problem: application to the 1D Euler equations, *Commun. Appl. Math. Comput.*, 5 (2023), pp. 370–402.

- [3] R. ABGRALL, A. CHERTOCK, A. KURGANOV, AND L. MICALIZZI, Dual formulation finite-volume methods on overlapping meshes for hyperbolic conservation laws, *Comput. & Fluids*, 307 (2026). Paper No. 106952.
- [4] R. ABGRALL AND S. KARNI, A comment on the computation of non-conservative products, *J. Comput. Phys.*, 229 (2010), pp. 2759–2763.
- [5] R. ABGRALL AND Y. LIU, A new approach for designing well-balanced schemes for the shallow water equations: a combination of conservative and primitive formulations, *SIAM J. Sci. Comput.*, 46 (2024), pp. A3375–A3400.
- [6] T. ALAZARD, Incompressible limit of the nonisentropic Euler equations with the solid wall boundary conditions, *Adv. Differential Equations*, 10 (2005), pp. 19–44.
- [7] P. ALLEGRINI AND M.-H. VIGNAL, Study of a new low-oscillating second-order all-Mach number IMEX finite volume scheme for the full Euler equations, *SIAM J. Sci. Comput.*, 47 (2025), pp. A268–A299.
- [8] J. B. BELL, P. COLELLA, AND H. M. GLAZ, A second-order projection method for the incompressible Navier-Stokes equations, *J. Comput. Phys.*, 85 (1989), pp. 257–283.
- [9] S. BOSCARINO, J.-M. QIU, G. RUSSO, AND T. XIONG, A high order semi-implicit IMEX WENO scheme for the all-Mach isentropic Euler system, *J. Comput. Phys.*, 392 (2019), pp. 594–618.
- [10] S. BOSCARINO, G. RUSSO, AND L. SCANDURRA, All Mach number second order semi-implicit scheme for the Euler equations of gas dynamics, *J. Sci. Comput.*, 77 (2018), pp. 850–884.
- [11] W. BOSCHERI, G. DIMARCO, R. LOUBÈRE, M. TAVELLI, AND M.-H. VIGNAL, A second order all Mach number IMEX finite volume solver for the three dimensional Euler equations, *J. Comput. Phys.*, 415 (2020). Paper No. 109486.
- [12] M. J. CASTRO DÍAZ, A. KURGANOV, AND T. MORALES DE LUNA, Path-conservative central-upwind schemes for nonconservative hyperbolic systems, *ESAIM Math. Model. Numer. Anal.*, 53 (2019), pp. 959–985.
- [13] C. CHALONS, M. GIRARDIN, AND S. KOKH, An all-regime Lagrange-projection like scheme for the gas dynamics equations on unstructured meshes, *Commun. Comput. Phys.*, 20 (2016), pp. 188–233.
- [14] A. CHERTOCK, Q. FU, A. KURGANOV, AND L. MICALIZZI, New adaptive numerical methods based on dual formulation of hyperbolic conservation laws. Submitted; arXiv:2601.20000.
- [15] S. CHU, A. KURGANOV, AND M. NA, Fifth-order A-WENO schemes based on the path-conservative central-upwind method, *J. Comput. Phys.*, 469 (2022). Paper No. 111508.
- [16] F. CORDIER, P. DEGOND, AND A. KUMBARO, An asymptotic-preserving all-speed scheme for the Euler and Navier-Stokes equations, *J. Comput. Phys.*, 231 (2012), pp. 5685–5704.

- [17] P. DEGOND, S. JIN, AND J.-G. LIU, Mach-number uniform asymptotic-preserving gauge schemes for compressible flows, *Bull. Inst. Math. Acad. Sin. (N.S.)*, 2 (2007), pp. 851–892.
- [18] P. DEGOND AND M. TANG, All speed scheme for the low Mach number limit of the isentropic Euler equations, *Commun. Comput. Phys.*, 10 (2011), pp. 1–31.
- [19] G. DIMARCO, R. LOUBÈRE, V. MICHEL-DANSAC, AND M.-H. VIGNAL, Second-order implicit-explicit total variation diminishing schemes for the Euler system in the low Mach regime, *J. Comput. Phys.*, 372 (2018), pp. 178–201.
- [20] G. DIMARCO, R. LOUBÈRE, AND M.-H. VIGNAL, Study of a new asymptotic preserving scheme for the Euler system in the low Mach number limit, *SIAM J. Sci. Comput.*, 39 (2017), pp. A2099–A2128.
- [21] L. FOX AND E. T. GOODWIN, Some new methods for the numerical integration of ordinary differential equations, *Proc. Cambridge Philos. Soc.*, 45 (1949), pp. 373–388.
- [22] F. GOLSE, S. JIN, AND C. D. LEVERMORE, The convergence of numerical transfer schemes in diffusive regimes. I. Discrete-ordinate method, *SIAM J. Numer. Anal.*, 36 (1999), pp. 1333–1369.
- [23] S. GOTTLIEB, D. KETCHESON, AND C.-W. SHU, Strong stability preserving Runge-Kutta and multistep time discretizations, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2011.
- [24] S. GOTTLIEB, C.-W. SHU, AND E. TADMOR, Strong stability-preserving high-order time discretization methods, *SIAM Rev.*, 43 (2001), pp. 89–112.
- [25] P. M. GRESHO AND S. T. CHAN, On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via a finite element method that also introduces a nearly consistent mass matrix. II. Implementation, *Internat. J. Numer. Methods Fluids*, 11 (1990), pp. 621–659. *Computational methods in flow analysis* (Okayama, 1988).
- [26] J. HAACK, S. JIN, AND J.-G. LIU, An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations, *Commun. Comput. Phys.*, 12 (2012), pp. 955–980.
- [27] T. Y. HOU AND P. G. LEFLOCH, Why nonconservative schemes converge to wrong solutions: error analysis, *Math. Comp.*, 62 (1994), pp. 497–530.
- [28] S. JIN, Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations, *SIAM J. Sci. Comput.*, 21 (1999), pp. 441–454 (electronic).
- [29] T. KATO, Perturbation theory for linear operators, *Classics in Mathematics*, Springer-Verlag, Berlin, 1980 ed., 1995.
- [30] A. KLAR, An asymptotic preserving numerical scheme for kinetic equations in the low Mach number limit, *SIAM J. Numer. Anal.*, 36 (1999), pp. 1507–1527.
- [31] R. KLEIN, Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics, I: One-dimensional flow, *J. Comput. Phys.*, 121 (1995), pp. 213–237.

- [32] A. KURGANOV AND C.-T. LIN, On the reduction of numerical dissipation in central-upwind schemes, *Commun. Comput. Phys.*, 2 (2007), pp. 141–163.
- [33] E. W. LARSEN AND J. E. MOREL, Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. II, *J. Comput. Phys.*, 83 (1989), pp. 212–236.
- [34] E. W. LARSEN, J. E. MOREL, AND W. F. MILLER, JR., Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes, *J. Comput. Phys.*, 69 (1987), pp. 283–324.
- [35] K.-A. LIE AND S. NOELLE, An improved quadrature rule for the flux-computation in staggered central difference schemes in multidimensions, *J. Sci. Comput.*, 63 (2003), pp. 1539–1560.
- [36] X. LIU, A. CHERTOCK, AND A. KURGANOV, An asymptotic preserving scheme for the two-dimensional shallow water equations with Coriolis forces, *J. Comput. Phys.*, 391 (2019), pp. 259–279.
- [37] G. MÉTIVIER AND S. SCHOCHET, The incompressible limit of the non-isentropic Euler equations, *Arch. Ration. Mech. Anal.*, 158 (2001), pp. 61–90.
- [38] L. MICALIZZI AND D. TORLO, A new efficient explicit deferred correction framework: analysis and applications to hyperbolic PDEs and adaptivity, *Commun. Appl. Math. Comput.*, 6 (2024), pp. 1629–1664.
- [39] L. MICALIZZI, D. TORLO, AND W. BOSCHERI, Efficient iterative arbitrary high-order methods: an adaptive bridge between low and high order, *Commun. Appl. Math. Comput.*, 7 (2025), pp. 40–77.
- [40] H. NESSYAHU AND E. TADMOR, Nonoscillatory central differencing for hyperbolic conservation laws, *J. Comput. Phys.*, 87 (1990), pp. 408–463.
- [41] S. NOELLE, G. BISPEN, K. R. ARUN, M. LUKÁČOVÁ-MEDVID’OVÁ, AND C.-D. MUNZ, A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics, *SIAM J. Sci. Comput.*, 36 (2014), pp. B989–B1024.
- [42] P. ÖFFNER, L. PETRI, AND D. TORLO, Analysis for implicit and implicit-explicit ADER and DeC methods for ordinary differential equations, advection-diffusion and advection-dispersion equations, *Appl. Numer. Math.*, 212 (2025), pp. 110–134.
- [43] R. M. PIDATELLA, G. PUPPO, G. RUSSO, AND P. SANTAGATI, Semi-conservative finite volume schemes for conservation laws, *SIAM J. Sci. Comput.*, 41 (2019), pp. B576–B600.
- [44] P. K. SWEBY, High resolution schemes using flux limiters for hyperbolic conservation laws, *SIAM J. Numer. Anal.*, 21 (1984), pp. 995–1011.
- [45] M. TANG, Second order all speed method for the isentropic Euler equations, *Kinet. Relat. Models*, 5 (2012), pp. 155–184.
- [46] E. F. TORO, *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction*, Springer-Verlag, Berlin, third ed., 2009.

- [47] E. F. TORO AND M. E. VÁZQUEZ-CENDÓN, Flux splitting schemes for the Euler equations, *Comput. & Fluids*, 70 (2012), pp. 1–12.
- [48] E. WEINAN AND C.-W. SHU, A numerical resolution study of high order essentially non-oscillatory schemes applied to incompressible flow, *J. Comput. Phys.*, 110 (1994), pp. 39–46.
- [49] J. ZEIFANG, J. SCHÜTZ, K. KAISER, A. BECK, M. LUKÁČOVÁ-MEDVID’OVÁ, AND S. NOELLE, A novel full-Euler low Mach number IMEX splitting, *Commun. Comput. Phys.*, 27 (2020), pp. 292–320.