Consiglio Nazionale delle Ricerche

Istituto di
Matematica Applicata e
Tecnologie Informatiche

PUBBLICAZIONI

Enrico Bertolazzi and Gianmarco Manzini

DIAGONALLY IMPLICIT–EXPLICIT RUNGE KUTTA METHODS FOR MULTIDIMENSIONAL HYPERBOLIC SYSTEMS.
PART II: ANALYSIS AND NUMERICAL EXPERIMENTS.

N. 17-PV 2004

Diagonally Implicit–Explicit Runge Kutta Methods for Multidimensional Hyperbolic Systems. Part II: Analysis and Numerical Experiments.

Enrico Bertolazzi [a]   Gianmarco Manzini [b]

[a]*Dipartimento di Ingegneria Meccanica e Strutturale,*
Università di Trento,
via Mesiano 77, I – 38050 Trento, Italy

[b]*Istituto di Matematica Applicata e Tecnologie Informatiche, IMATI – CNR,*
via Ferrata 1, I – 27100 Pavia, Italy

Abstract

In this paper we continue the study of the Diagonally IMplicit-EXplicit Runge-Kutta (DIMEX-RK) methods that we proposed in the first part of this work in the framework of Finite Volume methods for unstructured grids. These new numerical approximation schemes are based on a special splitting of the physical and numerical flux vector functions into a convective and a non-convective part. In the framework of DIMEX-RK schemes, the convective part is discretized in an implicit way, while we discuss the non-convective one in an explicit way. We discuss some theoretical properties of the non-linear algebraic evolution operators that are derived in the full discretization from the application of these methods to strongly convected dominated flows. A set of numerical experiments illustrates the behavior of this class of methods on reactive and non-reactive hypersonic simulations.

Key words: Finite Volume, Runge-Kutta, Implicit-Explicit, Partial Differential Equation, M-matrix, Unstructured Grid

1 Introduction

This is the second paper in a series in which we contruct and study high resolution Diagonally IMplicit-EXplicit Runge-Kutta (DIMEX-RK) schemes for numerically solving multi-dimensional hyperbolic systems in the framework of Finite Volume (FV) methods for unstructured grids. The multidimensional hyperbolic systems that we wish to focus on are related to either time-dependent or steady high speed compressible flow simulations. This kind of flows, that are strongly convected-dominated, is usually found when modelling reactive hypersonics. The mathematical model takes the conservative divergence form

$$\frac{\partial}{\partial t}\mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \mathbf{0}, \qquad \text{in } \Omega \times (0,\, T), \tag{1}$$

where Ω is a bounded open connected subset of $\mathbb{R}^d$, $d = 1, 2, 3$, $\mathbf{u}$ is a vector-valued function from $\mathbb{R}^d \times [0,\, \infty]$ into the open subset $\mathfrak{U} \subseteq \mathbb{R}^m$, and $\mathbf{F}(\mathbf{u})$ is a non-linear vector-valued mapping from $\mathfrak{U}$ into $\mathbb{R}^m$. Throughout the paper, we will refer to $\mathbf{u}$ as *the solution vector function*, to $\mathfrak{U}$ as *the set of admissible solution states*, and to $\mathbf{F}(\mathbf{u})$ as *the flux vector function*. As usual, the components at $\mathbf{F}(\mathbf{u})$ are assumed smooth — say of class C^∞.

The DIMEX-RK schemes, that we proposed in our first work [1], are based on a splitting of new conception of the numerical flux function into a convective and a non-convective part. This splitting allows us to adopt the DIMEX strategy for the time discretization. In fact, the convective term, which may be source of stiffness in this kind of problems, will be treated implicitly by mimicking the upwinding of a scalar linear flux. On the other hand, the non-convective part of the numerical flux function will be treated explicitly. In order to formally introduce this idea, let us first denote the set of d Jacobian matrices of the flux vector functions in (1) by

$$\mathbf{J}(\mathbf{u}) = \frac{\partial \mathbf{F}(\mathbf{u})}{\partial \mathbf{u}}, \qquad \text{for each } \mathbf{u} \in \mathfrak{U}.$$

Equation (1) is a multidimensional hyperbolic system of equations in divergence form if the matrix $\mathbf{J}(\mathbf{u}, \mathbf{n}) = \mathbf{n} \cdot \mathbf{J}(\mathbf{u})$ has m real eigenvalues $\lambda^{\text{min}}(\mathbf{u}, \mathbf{n}) = \lambda_1(\mathbf{u}, \mathbf{n}) \leq \cdots \leq \lambda_m(\mathbf{u}, \mathbf{n}) = \lambda^{\text{max}}(\mathbf{u}, \mathbf{n})$ and a complete set of eigenvectors for any $\mathbf{u} \in \mathfrak{U}$ and any non-zero vector $\mathbf{n} \in \mathbb{R}^d$. We will denote the minimum and maximum eigenvalues by $\lambda^{\text{min}}(\mathbf{u}, \mathbf{n})$ and $\lambda^{\text{max}}(\mathbf{u}, \mathbf{n})$. We focus our attention on problems of the form (1) whose physical flux vector function satisfies the following formal assumption.

Assumption 1 *The flux vector function* $\mathbf{F}(\mathbf{u})$ *can be split as*

$$\mathbf{F}(\mathbf{u}) = \mathbf{F}^{(c)}(\mathbf{u}) + \mathbf{F}^{(nc)}(\mathbf{u}),$$

where the convective part takes the form

$$\mathbf{F}^{(c)}(\mathbf{u}) = \mathbf{u} \otimes \mathbf{v}(\mathbf{u}),$$

$\mathbf{v}(\mathbf{u})$ *being the* convective velocity field *and satisfying*

$$\lambda^{\mathit{min}}(\mathbf{u}, \mathbf{n}) \leq \mathbf{n} \cdot \mathbf{v}(\mathbf{u}) \leq \lambda^{\mathit{max}}(\mathbf{u}, \mathbf{n}), \qquad \textit{for any } \mathbf{u} \in \mathfrak{U}, \quad \textit{and } \mathbf{n} \in \mathbb{R}^d.$$

Following Assumption 1, we consider numerical fluxes $\boldsymbol{H}(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n})$ that satisfy the following definition:

Definition 2 *A numerical flux admits a* numerical convective splitting *if it can be decomposed into the sum of a convective and non-convective part,*

$$\boldsymbol{H}(\mathbf{u}, \mathbf{v}, \mathbf{n}) = \boldsymbol{H}^{(c)}(\mathbf{u}, \mathbf{v}, \mathbf{n}) + \boldsymbol{H}^{(nc)}(\mathbf{u}, \mathbf{v}, \mathbf{n}).$$

The convective part takes the form

$$\boldsymbol{H}^{(c)}(\mathbf{u}, \mathbf{v}, \mathbf{n}) = \mathsf{a}(\mathbf{u}, \mathbf{v}, \mathbf{n})\mathbf{u} - \mathsf{a}(\mathbf{v}, \mathbf{u}, -\mathbf{n})\mathbf{v},$$

and is such that for each $\mathbf{u}$, $\mathbf{v} \in \mathfrak{U}$ *and* $\mathbf{n} \in \mathbb{R}^d$ *with* $\|\mathbf{n}\| = 1$ *the numerical convective velocity satisfies the following conditions:*

(i) $a(\mathbf{u},\mathbf{v},\mathbf{n}) \geq 0$;
(ii) $|a(\mathbf{u},\mathbf{v},\mathbf{n}) - a(\mathbf{u}',\mathbf{v}',\mathbf{n})| \leq L\left(\|\mathbf{u}-\mathbf{u}'\| + \|\mathbf{v}-\mathbf{v}'\|\right)$;
(iii) $\mathbf{v}(\mathbf{u}) \cdot \mathbf{n} = a(\mathbf{u},\mathbf{u},\mathbf{n}) - a(\mathbf{u},\mathbf{u},-\mathbf{n})$.

In [1] we discuss how many of the shock-capturing numerical fluxes developed for strongly convected-dominated problems can be re-formulated in accord with Definition 2. This allowed us to derive a new class of DIMEX-RK discretization methods in the FV framework and to guarantee high-resolution quality in the solution approximation by using the reconstruction from cell averages .

In this paper, we carry out the theoretical analysis of these DIMEX-RK finite volume methods to demonstrate some interesting properties concerning the algebraic structure of the discrete non-linear evolution operators.

The outline of the paper is as follows. In Section 2, we briefly review the semi-discrete FV formulation. In Section 3, we introduce the DIMEX-RK methods for the full discretization and the basic algebraic non-linear problem derived by the application of the method to the semi-discrete FV scheme. In Section 4, we study the theoretical properties of the non-linear algebraic evolution operators. In particular, we show the existence and uniqueness of the solution of the non-linear algebraic problem under quite general hypothesis by using the M-matrix theory. We also design efficient iterative procedures that ensure the correct formal order of accuracy in solving this non-linear algebraic problem. In Section 5, we study how non-negativity results can be obtained for these approximation methods. In Section 6, we illustrate the performance of these methods in term of computational efficiency and quality of approximation on several 1-D and 2-D numerical examples. Finally, in Section 7 conclusions are offered.

2 The semi-discrete Finite Volume formulation

2.1 Mesh notations and conventions

Our basic FV scheme is defined on a mesh that completely covers the computational domain $\Omega \in \mathbb{R}^d$. The mesh is defined as the union of a set of n non-overlapping control volumes or *cells*, that are *intervals* in the 1-D case, *triangles* in the 2-D case, and *tetrahedrons* in the 3-D case. The mesh is assumed to be *regular* and *conformal* in the sense specified by Reference [2].

Cells are conventionally labeled by an integer identifier ranging from 1 to n. The identifier is generically indicated by the index letters i, j or k. For the generic cell T_i we indicate by $|\mathsf{T}_i|$ the d dimensional measure of the cell (volume in 3-D, area in 2-D, length in 1-D).

The intersection of two cells or the intersection of a cell and the border of Ω with positive $(d-1)$ dimensional measure is called a face (edge in 2-D, point in 1-D). The internal face shared by the cells T_i and T_j is addressed by the pair ij and denoted by the symbol f_{ij}. For the sake of notation consistency, a boundary face is also addressed by a pair of indices, namely ik, i being the unique cell the face belongs to, and k a specific boundary face identifier — like a fictitious "external" cell. This convention allows us to refer to either internal or boundary faces by means of an index pair. For the generic face f_{ij}, we indicate by $|\mathsf{f}_{ij}|$ its $(d-1)-$dimensional measure (area in 3-D, length in 2D, conventionally 1 in 1-D), and by $\mathbf{n}_{ij}$ its normal vector (± 1 in 1D). The normal vector is assumed to be oriented from the cell i to the cell j when the face is internal and outward directed when the face is on the boundary.

For each cell T_i, we denote by $\sigma(i)$ the set of internal faces and by $\sigma'(i)$ the subset of the cell faces located at the boundary. The symbol $|\mathsf{T}_i|$ denotes the d-dimensional measure of the cell, i.e. the tetrahedron volume in 3-D, the triangle area in 2-D, and the interval length in 1-D. These notations and convections are suitable for practical implementation of Finite Volume solvers by using the object oriented library P2MESH [3].

2.2 Vector/Matrix notations and conventions

Let us now introduce the vector and matrix notations that will be utilized in this sections and in the following ones. The symbol $\mathbf{I}_k$ indicates the $k \times k$ identity matrix, $\mathbf{1}_k$ the k size vector all of whose components are equal to 1, while the symbol $\mathbf{0}_k$ indicates the k size vector all of whose components are equal to 0. A positive vector $\mathbf{v}$ satisfies $\mathbf{v} \gg \mathbf{0}$, a non-negative vector $\mathbf{v}$ satisfies $\mathbf{v} \geq \mathbf{0}$, where the compact notation means

$$\begin{aligned} \mathbf{v} \gg \mathbf{0} &\iff v_i > 0 \qquad \text{for all } i \\ \mathbf{v} \geq \mathbf{0} &\iff v_i \geq 0 \qquad \text{for all } i \\ \mathbf{v} > \mathbf{0} &\iff \mathbf{v} \geq 0 \qquad \text{and } \mathbf{v} \neq \mathbf{0}. \end{aligned}$$

Similar notations and definitions also apply to matrices; for instance, the matrix inequality $\mathbf{M} \leq \mathbf{N}$ means that $M_{ij} \leq N_{ij}$ for every pair ij, and a positive (non-negative) real matrix $\mathbf{M}$ is a matrix all of whose components are positive (non-negative) real numbers. Block vectors are denoted by underlined bold symbols, that is $\underline{\mathbf{u}}$ is a block vector. Their block components are indicated by indexed (non underlined) bold symbols, that is $\mathbf{u}_i$ is the i-th sub-vector block of $\underline{\mathbf{u}}$.

The symbol $\otimes$ indicates the standard tensor product, which is as follows. Given two matrices $\mathbf{A}$ and $\mathbf{B}$ of order $m \times n$ and $p \times q$, $\mathbf{A} \otimes \mathbf{B}$ is the block matrix of order $mp \times nq$ whose block i, j is $(\mathbf{A} \otimes \mathbf{B})_{i,j} = A_{ij}\mathbf{B}$. The tensor product has two noteworthy properties, see for instance Reference [4]. We just mention the one most used in the paper, that is $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$, with $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ four generic matrices (with compatible dimensions).

2.3 The basic semi-discrete FV scheme

The i-th cell-averaged solution state is denoted by $\mathbf{u}_i$, and the global collection of n cell-averaged data by $\underline{\mathbf{u}}$. Thus, this latter one is the $n \times m$-size block vector $\underline{\mathbf{u}}^T = (\mathbf{u}_1^T, \mathbf{u}_2^T, \dots, \mathbf{u}_n^T)$, whose i-th block is the m-size vector $\mathbf{u}_i$. We can also address the cell-averaged k-th equation variable for $k = 1, 2, \dots, m$ within the i-th cell, for $i = 1, 2, \dots, n$ by the notation $\mathbf{u}_i|_k = \underline{\mathbf{u}}|_{k+mi}$. This is the k-th component of the i-th vector block.

Reformulating equation (1) in an integral form for each cell of the mesh, applying the Gauss divergence theorem and introducing suitable numerical flux functions to discretize the physical flux yield the semi-discrete FV numerical scheme in the form

$$|\mathsf{T}_i| \frac{d\mathbf{u}_i}{dt} + \sum_{j\in\sigma(i)} \left(\mathsf{a}_{ij}(\underline{\mathbf{u}})\mathbf{u}_i - \mathsf{a}_{ji}(\underline{\mathbf{u}})\mathbf{u}_j\right) + \sum_{j\in\sigma(i)} \boldsymbol{H}_{ij}^{(nc)}(\underline{\mathbf{u}}) + \sum_{j'\in\sigma'(i)} \boldsymbol{H}_{ij'}^{(bc)}(\underline{\mathbf{u}}) = \mathbf{0}, \quad (2)$$

where the scalar functions a_{ij} and the vectors $\boldsymbol{H}_{ij}^{(nc)}$ are defined as follows

$$\mathsf{a}_{ij}(\underline{\mathbf{u}}) = |\mathsf{f}_{ij}| \sum_{k=1}^{N_q} \omega_k \mathsf{a}_{ij}^k(\underline{\mathbf{u}}), \quad (3a)$$

$$\boldsymbol{H}_{ij}^{(nc)}(\underline{\mathbf{u}}) = \boldsymbol{G}_{ij}^{(a)}(\underline{\mathbf{u}}) + |\mathsf{f}_{ij}| \sum_{k=1}^{N_q} \omega_k \boldsymbol{H}^{(nc)}(\mathbf{u}_i(\cdot, \mathbf{x}_{ij}^k), \mathbf{u}_j(\cdot, \mathbf{x}_{ij}^k), \mathbf{n}_{ij}), \quad (3b)$$

and

$$\mathsf{a}_{ij}^k(\underline{\mathbf{u}}) = \mathsf{a}(\mathbf{u}_i(\cdot, \mathbf{x}_{ij}^k), \mathbf{u}_j(\cdot, \mathbf{x}_{ij}^k), \mathbf{n}_{ij}), \quad (4a)$$

$$\boldsymbol{G}_{ij}^{(a)}(\underline{\mathbf{u}}) = |\mathsf{f}_{ij}| \sum_{k=1}^{N_q} \omega_k \left(\mathsf{a}_{ij}^k(\underline{\mathbf{u}})(\mathbf{u}_i(\cdot, \mathbf{x}_{ij}^k) - \mathbf{u}_i) - \mathsf{a}_{ji}^k(\underline{\mathbf{u}})(\mathbf{u}_j(\cdot, \mathbf{x}_{ij}^k) - \mathbf{u}_j)\right). \quad (4b)$$

In (3a)–(3b)–(4b), $\mathbf{x}_{ij}^k$ is the kth quadrature node on the face f_{ij}, ω_k the corresponding quadrature weight, and $\mathbf{u}_i(t, \mathbf{x})$ the solution reconstructed in the i-th cell. We assume that:

Assumption 3 *All of the weights used in quadrature formulae are* positive, *i.e.* $\omega_i > 0$ *for all* $i = 1, \dots, N_q$.

3 DIMEX-RK discretization in time

Let us rewrite equation (2) as

$$\frac{d\,\underline{\mathbf{u}}}{d\,t} = \underline{\mathbf{b}}\,(\underline{\mathbf{u}}) - \underline{\mathbf{a}}\,(\underline{\mathbf{u}})\,, \quad (5)$$

where

$$\mathbf{D} = \text{diag}(|\mathsf{T}_1|\,, \dots, |\mathsf{T}_N|), \quad \mathcal{A}(\underline{\mathbf{u}}) = \mathbf{D}^{-1}\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m, \quad \underline{\mathbf{a}}\,(\underline{\mathbf{u}}) = \mathcal{A}(\underline{\mathbf{u}})\underline{\mathbf{u}}. \quad (6)$$

The $n \times n$ matrix $\mathbf{A}(\underline{\mathbf{u}})$ is defined as

$$A_{ij}(\underline{\mathbf{u}}) = \begin{cases} \sum_{l\in\sigma(i)} \mathsf{a}_{il}(\underline{\mathbf{u}}) & \text{if } i = j; \\ -\mathsf{a}_{ji}(\underline{\mathbf{u}}) & \text{if } ij \text{ addresses a mesh face, namely } \mathsf{f}_{ij}; \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

The block vector $\underline{\mathbf{b}}\left(\underline{\mathbf{u}}\right)^T = [\,\mathbf{b}_1(\underline{\mathbf{u}})^T, \ldots, \mathbf{b}_n(\underline{\mathbf{u}})^T\,]$, whose i-th block is given by

$$|\mathsf{T}_i|\,\mathbf{b}_i(\underline{\mathbf{u}}) = -\sum_{j\in\sigma(i)} \boldsymbol{H}_{ij}^{(nc)}(\underline{\mathbf{u}}) - \sum_{j'\in\sigma'(i)} \boldsymbol{H}_{ij'}^{(bc)}(\underline{\mathbf{u}}).$$

The DIMEX-RK method provides a natural framework to discretize (5). In fact, since $\underline{\mathbf{a}}\left(\underline{\mathbf{u}}\right)$ contains the transport contribution to the flux, it can be interpreted as the *stiff* part of the system of ODEs (5), and an implicit discretization can be seek. On the other hand, the second r.h.s. term, i.e. $\underline{\mathbf{b}}\left(\underline{\mathbf{u}}\right)$, can be discretized explicitly.

The general r-stage DIMEX-RK scheme is formulated as:

— for each $i = 1, \ldots, r$ solve for $\underline{\mathbf{w}}^i$:

$$\underline{\mathbf{w}}^i + \Delta t\,\alpha_{ii}^{\mathrm{IM}}\,\underline{\mathbf{a}}\left(\underline{\mathbf{w}}^i\right) = \underline{\mathbf{u}}^n + \Delta t\sum_{j=1}^{i-1}\left(\alpha_{ij}^{\mathrm{EX}}\underline{\mathbf{b}}\left(\underline{\mathbf{w}}^j\right) - \alpha_{ij}^{\mathrm{IM}}\underline{\mathbf{a}}\left(\underline{\mathbf{w}}^j\right)\right), \qquad (8a)$$

— then compute

$$\underline{\mathbf{u}}^{n+1} = \underline{\mathbf{u}}^n + \Delta t\sum_{i=1}^{r}\left(\omega_i^{\mathrm{EX}}\underline{\mathbf{b}}\left(\underline{\mathbf{w}}^i\right) - \omega_i^{\mathrm{IM}}\underline{\mathbf{a}}\left(\underline{\mathbf{w}}^i\right)\right). \qquad (8b)$$

Specific values of $\boldsymbol{\alpha}^{\mathrm{EX}}$, $\boldsymbol{\alpha}^{\mathrm{IM}}$, $\boldsymbol{\omega}^{\mathrm{EX}}$, $\boldsymbol{\omega}^{\mathrm{IM}}$ can be found in Reference [5,6].

For completeness' sake, we report in Section 6 the value of these coefficients in a double Butcher tableau format for all the schemes that we consider in the numerical experiments.

The DIMEX-RK method in (8a)–(8b) requires the solution of r non-linear systems of the form

$$\underline{\mathbf{w}} + \Delta t\; a\,\underline{\mathbf{a}}\left(\underline{\mathbf{w}}\right) = \mathbf{r}, \qquad a > 0, \qquad (9)$$

where $\mathbf{r}$ is the right-hand-side of (8a). In view of equation (6), let us first define the map

$$\boldsymbol{\Phi}(\underline{\mathbf{w}}) = \left(\mathbf{I} + \Delta t\; a\,\mathcal{A}(\underline{\mathbf{w}})\right)^{-1}\mathbf{r}. \qquad (10)$$

Let us then observe that every solution $\underline{\mathbf{w}}$ of (9) is equivalently a fixed point of (10). The theoretical results of this paper are resumed in the two following theorems. The first one formally states the existence and uniqueness of the solution to the non-linear equation (9). The other one defines a recursive procedure to solve (9). Since both theorems are based on some theoretical properties of the discretization matrices, such as the Lipschitz continuity of $\mathcal{A}(\underline{\mathbf{w}})$, we anticipate here their statements without the proof, which will be given at the end of the next section.

Theorem 4 *The map* (10) *admits a fixed point for all* $\Delta t > 0$*. The fixed point is unique when*

$$\Delta t < \frac{1}{L\,a\,\kappa^2\,\|\mathbf{r}\|_1}, \qquad (11)$$

where

$$\kappa = \frac{\max_{i=1,2\ldots,n}|\mathsf{T}_i|}{\min_{i=1,2\ldots,n}|\mathsf{T}_i|},$$

and L is the Lipschitz constant of the map $\mathcal{A}(\mathbf{W})$.

Equation (9) depends on $\underline{\mathbf{w}}$ in a non-linear fashion, and thus the DIMEX-RK method as it has been proposed so far may be very expensive from the viewpoint of computational costs. However, the solution $\underline{\mathbf{w}}$ can be approximated up to order $\mathcal{O}\left(\Delta t^{k+1}\right)$ by a straightforward iterative procedure, as it is stated in the following theorem:

Theorem 5 *Let $\underline{\mathbf{w}}^i$ be defined iteratively for $i = 0, \dots, k$ as*

$$\begin{aligned} \underline{\mathbf{w}}^0 &= \underline{\mathbf{u}}^n, \\ \underline{\mathbf{w}}^i &= \left(\mathbf{I} + \Delta t\, a\, \mathcal{A}(\underline{\mathbf{w}}^{i-1})\right)^{-1} \mathbf{r}, \qquad i = 1, 2, \dots, k. \end{aligned} \tag{12}$$

Then, the k-th iterate $\underline{\mathbf{w}}^k$ is an $\mathcal{O}\left(\Delta t^{k+1}\right)$ approximation of $\underline{\mathbf{w}}$, which is the exact solution to (9)*.*

4 Properties of the discretization matrices $\mathbf{A}(\underline{\mathbf{u}})$, $\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m$ and proofs

In this section, we discuss some important properties of the matrices $\mathbf{A}(\underline{\mathbf{u}})$ and $\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m$ and in particular we will prove that they are singular *M-matrices* and Lipschitz continuous mappings of $\underline{\mathbf{u}}$. These results are crucial in demonstrating theorems 4 and 5, whose proofs are given at the end of this section. Let us first introduce the notations, the definitions and the basic results from the M-matrix theory. The basic properties of M-matrices are listed for reference's sake as technical lemmata without proofs. A detailed presentation of this topic is beyond the scope of the present paper; hence, we refer the interested reader to the extensive exposition given in References [7,8].

Definition 6 *Any matrix $\mathcal{M}$ of the form $\mathcal{M} = s\mathbf{I} - \mathcal{B}$, with $\mathcal{B} \geq \mathbf{0}$ is a* Z-matrix*. Moreover if $s > \rho(\mathcal{B})$ the spectral radius of $\mathcal{B}$ then $\mathcal{M}$ is an* M-matrix*, if $s = \rho(\mathcal{B})$ it is a* singular *M-matrix.*

Lemma 7 *Any matrix $\mathcal{M}$ such that $\varepsilon\mathbf{I} + \mathcal{M}$ is an M-matrix for any real number $\varepsilon > 0$ is a* singular *M-matrix.*

Lemma 8 *A matrix $\mathcal{M}$ is a Z-matrix iff $\mathcal{M}_{ij} \leq 0$ for $i \neq j$.*

Lemma 9 *Each one of these three statements is equivalent to the statement "$\mathcal{M}$ is an M-matrix":*

- (i) *$\mathcal{M}^T$ is an M-matrix;*
- (ii) *$\mathcal{M}$ is a Z-matrix and there exists a vector $\mathbf{x} \gg \mathbf{0}$ such that $\mathcal{M}\mathbf{x} \gg \mathbf{0}$;*
- (iii) *$\mathcal{M}$ is a non-singular Z-matrix and $\mathcal{M}^{-1} > \mathbf{0}$.*

Lemma 10 *Let $\mathcal{M}$ be an M-matrix and $\mathcal{D}$ a non-negative diagonal matrix, then:*

- (i) *if $\mathcal{D}$ is non-singular then $\mathcal{D}\mathcal{M}$ is an M-matrix;*

(ii) *$\mathcal{D} + \mathcal{M}$ is an M-matrix;*
(iii) *the inequality $(\mathcal{D} + \mathcal{M})^{-1} \leq \mathcal{M}^{-1}$ holds.*

From the definition given in (7) it follows that $\mathbf{A}(\underline{\mathbf{u}})$ and $\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m$ are singular matrices. We formally state this fact in Lemma 11.

Lemma 11

(i) *$\mathbf{1}_n^T$ is a left eigenvector of $\mathbf{A}(\underline{\mathbf{u}})$, with respect to the null eigenvalue;*
(ii) *$\mathbf{1}_{nm}^T = \mathbf{1}_n^T \otimes \mathbf{1}_m^T$ is a left eigenvector of $\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m$ with respect to the null eigenvalue;*

Proof.

(i) From (7) it follows that $\left[\mathbf{1}_n^T \mathbf{A}(\underline{\mathbf{u}})\right]_k = \sum_{j=1}^n A_{jk}(\underline{\mathbf{u}}) = 0$ for $k = 1, 2, \ldots, n$.
(ii) From (7) and (i) it follows that $(\mathbf{1}_n^T \otimes \mathbf{1}_m^T)(\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m) = \mathbf{1}_n^T \mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{1}_m^T \mathbf{I}_m = \mathbf{0}_n^T \otimes \mathbf{1}_m^T = \mathbf{0}_{nm}^T$.

The following proposition is among the main results of the present section.

Proposition 12

(i) *$\mathbf{A}(\underline{\mathbf{u}})$ is a singular M-matrix.*
(ii) *$\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m$ is a singular M-matrix.*

Proof.

(i) From (7) and Lemma 8 it follows that $\mathbf{A}(\underline{\mathbf{u}})$ is a Z-matrix. From Lemma 11 we have that $\mathbf{1}_n^T \mathbf{A}(\underline{\mathbf{u}}) = \mathbf{0}_n^T$ and there holds that $\mathbf{1}_n^T (\varepsilon \mathbf{I}_n + \mathbf{A}(\underline{\mathbf{u}})) = \varepsilon \mathbf{1}_n^T \gg \mathbf{0}_n^T$. By using Lemma 9$(ii)$, it follows that $\varepsilon \mathbf{I}_n + \mathbf{A}(\underline{\mathbf{u}})$ is an M-matrix for all $\varepsilon > 0$. Finally, Lemma 7 implies the statement.
(ii) The result follows by using the same arguments of the previous item and the positive vector $\mathbf{1}_n^T \otimes \mathbf{1}_m^T$ instead of $\mathbf{1}_n^T$

From the Lipschitz continuity property of the numerical flux of the numerical convective splitting there immediately turns out the following proposition.

Proposition 13 *Both $\mathbf{A}(\underline{\mathbf{u}})$ and $\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m$ are Lipschitz continuous mappings of the entry argument $\underline{\mathbf{u}}$ with respect to the* same *Lipschitz constant L.*

Proof. From Definition 2-*(ii)* and the fact that the numerical flux is usually supposed to be a Lipschitz function of its arguments, it immediately follows that

$$\|\mathbf{A}(\underline{\mathbf{u}}) - \mathbf{A}(\underline{\mathbf{v}})\|_1 = \|\mathbf{A}(\underline{\mathbf{u}}) \otimes \mathbf{I}_m - \mathbf{A}(\underline{\mathbf{v}}) \otimes \mathbf{I}_m\|_1 \leq L \, \|\underline{\mathbf{u}} - \underline{\mathbf{v}}\|_1 .$$

The following lemma illustrates how we can take advantage of the fact that all of the systems (9) are indeed generated by the finite volume discretization of numerical fluxes re-formulated along the lines of the numerical convective splitting.

Lemma 14 *The matrix $\mathcal{A}(\underline{\mathbf{w}})$*

(i) *is a singular M-matrix;*

(ii) *is a Lipschitz continuous function of its argument, that is a real positive constant exists, L, such that for every couple of vectors $\underline{\mathbf{u}}$ and $\underline{\mathbf{w}}$ the inequality*

$$\|\mathcal{A}(\underline{\mathbf{u}}) - \mathcal{A}(\underline{\mathbf{w}})\|_1 \leq L\, \|\underline{\mathbf{u}} - \underline{\mathbf{w}}\|_1$$

holds.

Proof. The first item is a direct consequence of Proposition 12, while the second one follows immediately from Proposition 13.

Then, we have the two following lemmata (see Reference [9] for the $\|\cdot\|_\infty$ estimate).

Lemma 15 *Let $\mathcal{M} = \mathbf{I} + \mathbf{B}$ be a singular M-matrix, and $\mathbf{v} \gg \mathbf{0}$ a non-negative vector such that $\mathbf{v}^T\mathbf{B} = \mathbf{0}$; then*

$$\left\|\mathcal{M}^{-1}\right\|_1 \leq \frac{v_{max}}{v_{min}}.$$

Proof. From $\mathbf{v}^T\mathbf{B} = \mathbf{0}$ it follows

$$v_{max}\mathbf{1}^T \geq \mathbf{v}^T = \mathbf{v}^T(\mathbf{I} + \mathbf{B})^{-1} = \mathbf{v}^T\mathcal{M}^{-1}; \tag{13}$$

from Lemma 9(iii) we can write,

$$\mathbf{v}^T\mathcal{M}^{-1} \geq v_{min}\mathbf{1}^T\mathcal{M}^{-1}, \tag{14}$$

and by using both (13) and (14) we prove the lemma.

Lemma 16 *The matrix $\mathbf{I} + \Delta t\; a\, \mathcal{A}(\underline{\mathbf{w}})$ is a non-singular M-matrix for every vector $\underline{\mathbf{w}}$ and its non-negative inverse verifies the inequality*

$$\left\|(\mathbf{I} + \Delta t\; a\, \mathcal{A}(\underline{\mathbf{w}}))^{-1}\right\|_1 \leq \kappa. \tag{15}$$

Proof. Because of Proposition 12(i), the term $\Delta t\; a\, \mathcal{A}(\underline{\mathbf{w}})$ is a singular M-matrix; moreover a direct calculation yields

$$\begin{aligned}(\mathbf{1}_n^T\mathbf{D} \otimes \mathbf{1}_m^T)\mathcal{A}(\underline{\mathbf{w}}) &= (\mathbf{1}_n^T\mathbf{D} \otimes \mathbf{1}_m^T)\left(\mathbf{D}^{-1}\mathbf{A}(\underline{\mathbf{w}}) \otimes \mathbf{I}_m\right),\\ &= \mathbf{1}_n^T\mathbf{A}(\underline{\mathbf{w}}) \otimes \mathbf{1}_m^T,\\ &= \mathbf{0}_n,\end{aligned}$$

and from Lemma 15 with $\mathbf{v} = \mathbf{D}\mathbf{1}_n \otimes \mathbf{1}_m$ equation (15) follows.

Lemma 17 *The following inequality holds for all $\underline{\mathbf{u}}$, $\underline{\mathbf{w}}$:*

$$\|\,\boldsymbol{\Phi}(\underline{\mathbf{u}}) - \boldsymbol{\Phi}(\underline{\mathbf{w}})\|_1 \leq \Delta t\; a\, L\, \kappa^2\, \|\underline{\mathbf{u}} - \underline{\mathbf{w}}\|_1\, \|\mathbf{r}\|_1\,. \tag{16}$$

Proof. The inequality (16) follows by noting that

$$\begin{aligned}&\boldsymbol{\Phi}(\underline{\mathbf{u}}) - \boldsymbol{\Phi}(\underline{\mathbf{w}})\\ &\quad= (\mathbf{I} + \Delta t\; a\, \mathcal{A}(\underline{\mathbf{u}}))^{-1}\, \mathbf{r} - (\mathbf{I} + \Delta t\; a\, \mathcal{A}(\underline{\mathbf{w}}))^{-1}\, \mathbf{r},\\ &\quad= \Delta t\; a\; (\mathbf{I} + \Delta t\; a\, \mathcal{A}(\underline{\mathbf{u}}))^{-1}\, (\mathcal{A}(\underline{\mathbf{w}}) - \mathcal{A}(\underline{\mathbf{u}}))\, (\mathbf{I} + \Delta t\; a\, \mathcal{A}(\underline{\mathbf{w}}))^{-1}\, \mathbf{r},\end{aligned}$$

taking the 1-norm on both sides and applying the result of Lemma 16

Finally, these are the proofs of Theorems 4 and 5.

Proof of Theorem 4. Let us notice that

$$\mathcal{K} = \{\underline{\mathbf{u}} \mid \|\underline{\mathbf{u}}\|_1 \leq \kappa \|\mathbf{r}\|_1\}$$

is a convex compact set and that from Lemma 16 there holds

$$\|\boldsymbol{\Phi}(\underline{\mathbf{u}})\|_1 \leq \left\|(\mathbf{I} + \Delta t\, a\, \mathcal{A}(\underline{\mathbf{u}}))^{-1}\right\|_1 \|\mathbf{r}\|_1 \leq \kappa \|\mathbf{r}\|_1 .$$

Thus, $\boldsymbol{\Phi}$ is a continuous map from the convex compact set $\mathcal{K}$ into $\mathcal{K}$. From the Brouwer fixed point theorem [10] it follows that the map admits a fixed point. Finally, from Lemma 17, equation (16), it follows that if Δt satisfies (11) the map $\boldsymbol{\Phi}$ is a contraction. This implies the uniqueness of the solution.

Proof of Theorem 5. By using the approximation $\mathcal{A}(\underline{\mathbf{w}}) \approx \mathcal{A}(\underline{\mathbf{u}}^n)$ in (9), we define $\underline{\mathbf{w}}^*$ as the solution of the linear algebraic system

$$(\mathbf{I} + \Delta t\, a\, \mathcal{A}(\underline{\mathbf{u}}^n))\, \underline{\mathbf{w}}^* = \mathbf{r}. \tag{17}$$

By comparing (9) and (17), we have

$$(\mathbf{I} + \Delta t\, a\, \mathcal{A}(\underline{\mathbf{u}}^n))\, (\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) = \Delta t\, a\, L\, [\mathcal{A}(\underline{\mathbf{u}}^n) - \mathcal{A}(\underline{\mathbf{w}})]\, \underline{\mathbf{w}},$$

and in view of Lemmata 16 and 14 there holds

$$\begin{aligned} \|\underline{\mathbf{w}} - \underline{\mathbf{w}}^*\|_1 &\leq \left\|(\mathbf{I} + \Delta t\, a\, \mathcal{A}(\underline{\mathbf{u}}^n))^{-1}\right\|_1 \|\mathcal{A}(\underline{\mathbf{u}}^n) - \mathcal{A}(\underline{\mathbf{w}})\|_1 \, \|\underline{\mathbf{w}}\|_1 , \\ &\leq \kappa\, \Delta t\; a\; \|\underline{\mathbf{w}} - \underline{\mathbf{u}}^n\|_1 \, \|\underline{\mathbf{w}}\|_1 . \end{aligned} \tag{18}$$

As $\|\underline{\mathbf{w}} - \underline{\mathbf{u}}^n\| = \mathcal{O}\,(\Delta t)$, it follows that $\|\underline{\mathbf{w}} - \underline{\mathbf{w}}^*\|_1 = \mathcal{O}\left(\Delta t^2\right)$.

Thus, the approximation $\mathcal{A}(\underline{\mathbf{w}}) \approx \mathcal{A}(\underline{\mathbf{u}}^n)$ in (17) is first order accurate. Nonetheless, when a better approximation of $\underline{\mathbf{w}}$ is used instead of $\underline{\mathbf{u}}^n$, namely $\underline{\mathbf{u}}^{*,k}$, such that $\left\|\underline{\mathbf{u}}^{*,k} - \underline{\mathbf{w}}\right\|_1 = \mathcal{O}\left(\Delta t^k\right)$, we can define $\underline{\mathbf{w}}^{*,k}$ as the solution of

$$\left(\mathbf{I} + \Delta t\; a\, \mathcal{A}(\underline{\mathbf{u}}^{*,k})\right) \underline{\mathbf{w}}^{*,k} = \mathbf{r}.$$

From inequality (18) with $\underline{\mathbf{u}}^n$ substituted by $\underline{\mathbf{u}}^{*,k}$ it turns out that $\underline{\mathbf{w}}^{*,k}$ is an $\mathcal{O}\left(\Delta t^{k+1}\right)$ approximation of the *true* non-linear solution $\underline{\mathbf{w}}$.

5 Some remarks about the non-negativity of the scheme

In order to discuss the non-negativity issue of these DIMEX-RK FV schemes, we shall consider in the rest of this section a simplified model problem. To this purpose, let us assume that the unknown vector $\mathbf{u}$ be transported by a (possibly non-linear) advective flux $\mathbf{u} \otimes \mathbf{v}(\mathbf{u})$ and that the strictly non-linear contribution in the physical flux is null, that is $\mathbf{F}^{(nc)}(\mathbf{u}) = \mathbf{0}$. We have

$$\frac{\partial}{\partial t}\mathbf{u} + \nabla \cdot \mathbf{u} \otimes \mathbf{v}(\mathbf{u}) = \mathbf{0}, \qquad \text{in } \Omega \times (0,\, T). \tag{19}$$

Notice that in (19) we do not make any strict assumption on the linearity of the flux, because the velocity field may still be dependent on the solution $\mathbf{u}$. If the boundary flux is non-negative, i.e. $\mathbf{u}\otimes\mathbf{v}(\mathbf{u})\cdot\mathbf{n} \geq \mathbf{0}$, we can state the following non-negativity results on the cell-averaged solution $\mathbf{u}$. For the sake of exposition, we consider $\underline{\mathbf{u}}^n$ as the cell-averaged vector of one unknown field at the time t^n. All of the results readily generalize to the more general case of a vector of unknown variables.

Proposition 18 *The simplest DIMEX-RK scheme is the* Forward–Backward Euler *method, which is first order accurate in time and takes the form*

$$\underline{\mathbf{u}}^{n+1} + \Delta t\, \underline{\boldsymbol{a}}\left(\underline{\mathbf{u}}^{n+1}\right) = \underline{\mathbf{u}}^{n} + \Delta t\, \underline{\boldsymbol{b}}\left(\underline{\mathbf{u}}^{n}\right). \tag{20}$$

(1) *Let us suppose that* $\mathbf{u}^n \geq \mathbf{0}$*. Then, the scheme in* (20)*, which is globally first order accurate when no spatial reconstruction is applied, is* unconditionally *non-negative, that is* $\mathbf{u}^{n+1} \geq \mathbf{0}$*;*

(2) *The same scheme produces a non-negative solution under the condition*

$$\left(\mathbf{I} + \Delta t C(\mathbf{u}^n)\right)\mathbf{u}^n \geq \mathbf{0}, \tag{21}$$

which clearly depends on the reconstruction matrix.

Proof.

(1) The semi-discrete form of the model equation (19) is

$$\frac{\partial}{\partial t}\mathbf{u} + \mathcal{A}(\mathbf{u})\mathbf{u} = \mathbf{0}, \tag{22}$$

and the first order accurate DIMEX scheme in (20) *without* reconstruction can be simply formulated as

$$\left(\mathbf{I} + \Delta t \mathcal{A}(\mathbf{u}^{n+1})\right)\mathbf{u}^{n+1} = \mathbf{u}^n,$$

which formally implies that

$$\mathbf{u}^{n+1} = \left(\mathbf{I} + \Delta t \mathcal{A}(\mathbf{u}^{n+1})\right)^{-1}\mathbf{u}^n.$$

Since $(\mathbf{I} + \Delta t \mathcal{A}(\mathbf{u}^{n+1}))$ is an M-matrix and from Lemma 9, item 3, $\mathbf{u}^n \geq \mathbf{0}$ implies that $\mathbf{u}^{n+1} \geq \mathbf{0}$. This result is not limited by constraints on the time step size Δt.

(2) When a spatial reconstruction is considered, (22) is discretized by applying the first order time accurate scheme (20) as

$$\left(\mathbf{I} + \Delta t \mathcal{A}(\mathbf{u}^{n+1})\right)\mathbf{u}^{n+1} = \left(\mathbf{I} + \Delta t C(\mathbf{u}^n)\right)\mathbf{u}^n.$$

The non-negativity of $\mathbf{u}^{n+1}$ can be obtained by an argument that exploits the non-negativity of the inverse of an M-matrix as in the proof of the previous item. In this case, the condition to be satisfied by the r.h.s of the equation is no more $\mathbf{u} \geq \mathbf{0}$, but the one stated in the proposition.

The condition (21) can be used to produce suitable sufficient conditions capable of ensuring the non-negativity of the cell-averaged solution. To see how this item works, let us derive one such sufficient condition. Let us consider the i-th cell equation. There holds

$$\begin{aligned}\left[\left(\mathbf{I} + \Delta t C(\mathbf{u}^n)\right)\mathbf{u}^n\right]_i &= \left(\mathbf{I}_m + \Delta t C_{ii}(\mathbf{u}^n)\right)\mathbf{u}_i^n + \sum_{j\in\sigma(i)} C_{ij}(\mathbf{u}^n)\mathbf{u}_j^n \\ &\geq \left(\mathbf{I}_m + \Delta t C_{ii}(\mathbf{u}^n)\right)\mathbf{u}_i^n - \Delta t C_{ii} M_i^n \mathbf{1}_m \geq \mathbf{0},\end{aligned}$$

where $M_i^n = \max_{j\in\sigma(i)} \mathbf{u}_j^n$. The *local* condition on the cell time-step Δt_i is

$$\Delta t_i \leq \frac{m_i^n}{\max_s C_{ii}(\mathbf{u}^n)|_{ss}(M_i^n - m_i^n)}, \qquad m_i^n = \min_{j\in\sigma(i)} \mathbf{u}_j^n.$$

This last condition implies the *global* time-step sufficient condition

$$\Delta t_{max} \leq \frac{2}{\|C(\mathbf{u}^n)\|_\infty}\left(\frac{m}{M-m}\right), \qquad m = \min_{i,n} m_i^n, \qquad M = \max_{i,n} M_i^n,$$

and we exploited the fact that $\max_{i,s} C_{ii}(\mathbf{u}^n)|_{ss} = \frac{1}{2}\left\|C(\mathbf{u}^n)\right\|_{\infty}$.

Stating general conditions to ensure non-negativity is more difficult when we consider a DIMEX-RK scheme with formal accuracy greater than 1 or when the truly non-linear part of the flux defined in (1) is non null, i.e. $\mathbf{F}^{(nc)}(\mathbf{u}) \neq \mathbf{0}$. Nonetheless, a case-by-case analysis may again produce non-negativity results, see for instance Reference [11]. For example, by using the result in Reference [12] and due to the fact that the MDP(1,2,2) can be written as the sum of an implicit and an explicit part, the non-negativity property can still be ensured under the condition *CFL* ≤ 2.

6 Numerical Results

In this section we document the performance of several representative DIMEX-RK FV schemes as far as their approximation order is concerned and experimentally investigate the computational efficiency of the DIMEX-RK approach on a complex fluid dynamics application. Tables (1-3) show the double Butcher's tableaux with the coefficients $\boldsymbol{\alpha}^{\text{IM}}$, $\boldsymbol{\omega}^{\text{IM}}$, and $\boldsymbol{\alpha}^{\text{EX}}$, $\boldsymbol{\omega}^{\text{EX}}$ used in the formulation (8a)–(8b) for all of the schemes in the numerical experiments of this section. The format is as follows

$$\begin{array}{c|c} \mathbf{c}^{\text{IM}} & \boldsymbol{\alpha}^{\text{IM}} \\ \hline & \boldsymbol{\omega}^{\text{IM}} \end{array} \qquad \begin{array}{c|c} \mathbf{c}^{\text{EX}} & \boldsymbol{\alpha}^{\text{EX}} \\ \hline & \boldsymbol{\omega}^{\text{EX}} \end{array}$$

where the tableau on the left refers to the implicit part of the approximation, while the one on the right to the explicit part of the approximation. We also use the triplet notation (s', s'', r), where the integer s' characterizes the number of stages of the implicit scheme, the integer s'' characterizes the number of stages of the explicit scheme, and r is the order of the DIMEX-RK integrator.

Table 1
First-Order DIMEX-RK Methods

FB(1,1,1)

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1 & 0 \end{array}$$

MDP(1,2,2)

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 0 & 1/2 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

6.1 Accuracy of the Methods

In the first set of experiments, we measure the order of accuracy in time and the global (spatial and temporal) order of accuracy of each approximation scheme. Since the order of accuracy is formally defined for smooth functions, we propose the following original test case for the one-dimensional compressible Euler equations which shows a smooth exact solution. The initial density and pressure fields consist in a bell-shaped pulse superimposed to a spatially constant value, which is translated by an initially constant velocity field. The extension of the one-dimensional computational domain is virtually infinite, but clearly only a small portion of it can be represented. The simulation is arrested before the pulse goes out of the finite computational domain in order not to have to take care of the boundary

Table 2
Second-Order DIMEX-RK Methods

ARS(2,2,2) [5]

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & 0 & \gamma & 0 \\ 1 & 0 & 1-\gamma & \gamma \\ \hline & 0 & 1-\gamma & \gamma \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ 1 & \delta & 1-\delta & 0 \\ \hline & \delta & 1-\delta & 0 \end{array} \qquad \begin{array}{rcl} \gamma & = & 1-\dfrac{\sqrt{2}}{2}, \\ \delta & = & 1-\dfrac{1}{2\gamma}. \end{array}$$

ARS(2,3,2) [5]

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & 0 & \gamma & 0 \\ 1 & 0 & 1-\gamma & \gamma \\ \hline & 0 & 1-\delta & \delta \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ 1 & \delta & 1-\delta & 0 \\ \hline & \delta & 1-\delta & 0 \end{array} \qquad \begin{array}{rcl} \gamma & = & 1-\dfrac{\sqrt{2}}{2}, \\ \delta & = & -\dfrac{2\sqrt{2}}{3}. \end{array}$$

LRR(3,2,2) [13]

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 \\ 1 & 0 & 0 & 3/4 & 1/4 \\ \hline & 0 & 0 & 3/4 & 1/4 \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ \hline & 0 & 1 & 0 & 0 \end{array}$$

Table 3
Third-Order DIMEX-RK Methods

ARS(2,3,3) [5]

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & 0 & \gamma & 0 \\ 1-\gamma & 0 & 1-2\gamma & \gamma \\ \hline & 0 & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ 1-\gamma & \gamma-1 & 2-2\gamma) & 0 \\ \hline & 0 & 1/2 & 1/2 \end{array} \qquad \gamma = \frac{3+\sqrt{3}}{6}.$$

ARS(4,4,3) [5]

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 2/3 & 0 & 1/6 & 1/2 & 0 & 0 \\ 1/2 & 0 & -1/2 & 1/2 & 1/2 & 0 \\ 1 & 0 & 3/2 & -3/2 & 1/2 & 1/2 \\ \hline & 0 & 3/2 & -3/2 & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 2/3 & 11/18 & 1/18 & 0 & 0 & 0 \\ 1/2 & 5/6 & -5/6 & 1/2 & 0 & 0 \\ 1 & 1/4 & 7/4 & 3/4 & -7/4 & 0 \\ \hline & 1/4 & 7/4 & 3/4 & -7/4 & 0 \end{array}$$

condition treatment.

The order of accuracy in time is numerically measured as follows. We calculate two distinct solutions with time-steps Δt and $\Delta t/2$ and we compare them with a reference solution obtained by using the time-step $\Delta t/10$. The logarithm of the ratio between differences measured in a standard L^2-norm yields the desired time convergence rate. In this case we use a mesh of 200 intervals and a piecewise constant reconstruction.

The global accuracy is estimated by running three different simulations on meshes with respectively 200, 400 and 800 intervals. The time-step size changes during the run in accord with the maximum allowable *CFL* number. We use a piecewise constant reconstruction for the first order time-stepping schemes, a linear reconstruction for the second order time-stepping schemes, and a parabolic reconstruction for the third order time-stepping schemes. The slopes in the linear reconstruction are monotonized by a standard minmod limiter, while slopes and concavities in the parabolic reconstruction are monotonized by adopting the procedure proposed in Reference [14].

Table 4 reports the orders of accuracy obtained by using the HLLE flux. The first and second columns of the table are self-explanatory. The third column shows the convergence rates in time of the DIMEX-RK schemes, and the fourth column their global (in time and space) convergence rates, all measured as explained above. From the table it is evident that all these DIMEX-RK FV achieves the formal order of accuracy that is theoretically expected. It should be pointed out that repeating these experiments by using the other numerical fluxes that meet the numerical convective splitting formulation, like the *Steger-Warming* flux splitting, the *Van Leer* flux splitting and the *AUSM+* flux splitting (see [1] for more details), produces similar results. This very little dependence on the choice of the numerical flux is quite reasonable, because we do not measure any significant issue related to the spatial discretization.

Table 4
DIMEX-RK convergence rates in time and for the full (time and space) FV discretizations.

Formal Accuracy	DIMEX-RK Method	DIMEX-RK Rate	Global rate
First-Order	ERK(1)	1.220	0.952
	FB(1,1,1)	1.126	0.930
Second-Order	ERK(2)	2.048	1.947
	MDP(1,2,2)	2.044	1.947
	ARS(2,2,2)	2.045	1.946
	ARS(2,3,2)	2.045	1.946
	LRR(3,3,2)	2.045	1.946
Third-Order	ERK(3)	3.019	2.963
	ARS(2,3,3)	2.989	2.984
	ARS(4,4,3)	3.004	2.943

6.2 Computational Efficiency

In the second set of experiments, we consider the system of 2-D reactive compressible Euler equations on three standard test cases taken from the CFD literature. Basically, a shock discontinuity is moving on a compression ramp with different inclination angles and thus forms well-known shock patterns at the final integration time. As we are only interested in measuring the CPU times for the various DIMEX-RK schemes, a detailed analysis of these shock interaction phenomena is beyond the scope of this paper. Thus, we refer the reader to Reference [11] and the bibliography therein for a thorough description of this model problems. We list below the major simulation parameters that characterize each test case. We also provide for each test case a figure that shows the approximate solution at the final time step as indicated below. The approximate solution shown in each figure is calculated by using the a piece-wise linear reconstruction, the HLLE numerical flux and the ERK(2) time integrator.

(i) *Single Mach Reflection*:
Mach number $M_s = 2.03$, compression angle 27^0, mesh of 59756 triangles, final time $t = 100\mu s$, solution shown in Figure 1;

(ii) *Complex Mach Reflection*:
Mach number $M_s = 10.37$, compression angle 10^0, mesh of 34833 triangles, final time $t = 20\mu s$, solution shown in Figure 3;

(iii) *Double Mach Reflection*:
Mach number $M_s = 8.7$, compression angle 27^0, mesh of 49179 triangles, final time $t = 24\mu s$, solution shown in Figure 2.

For a complete description of the model problem and the three test cases, we refer the reader to Reference [11] and the bibliography therein.

Table 5 illustrates the performance of the schemes ERK(2) and DIMEX MDP(1,2,2). Row *#Iter* reports the total number of iterations required to solve the non-linear algebraic system (9) by applying the iterative procedure of equation (12). The non-linear solving procedure implements a standard Richardson iterative scheme [15] coupled with an incomplete LU preconditioner [16]. Rows *Rec.*, *Chem.* and *Step* respectively detail the computational time (in seconds) needed to perform second order reconstructions, the update of the chemical terms, and the rest of the integration time-step. Finally, Row *Tot. Time* gives the total CPU time to perform the run, and Row *% Gain* the percentage gain obtained by using the DIMEX-RK scheme instead of the explicit one.

Let us remark that the reconstruction procedure, which is needed to attain second order accuracy in space, is the same for both ERK(2) and MDP(1,2,2) methods and must be performed at each time-step. The reconstruction is an important entry in the overall computational costs, but its impact is less dramatic in the DIMEX-RK case than in the explicit one. This is because the former scheme allows a greater time step size and requires less time iterations to perform the same calculation. The update of the chemical terms (Row *Chem.*) and the rest of the integration time step (Row *Step*) is less expensive for the explicit scheme, because the DIMEX-RK method requires the iterative re-solution of a set non-linear algebraic systems. Nonetheless, from the viewpoint of global computational costs (Row *Tot. Time*) the

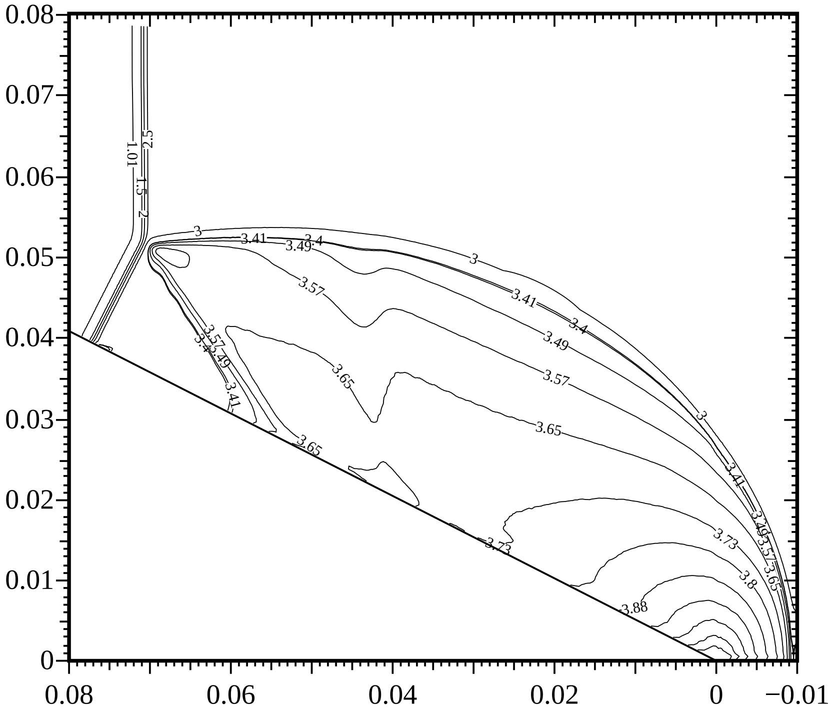

Fig. 1. Single Mach Reflection at $T = 100\mu s$

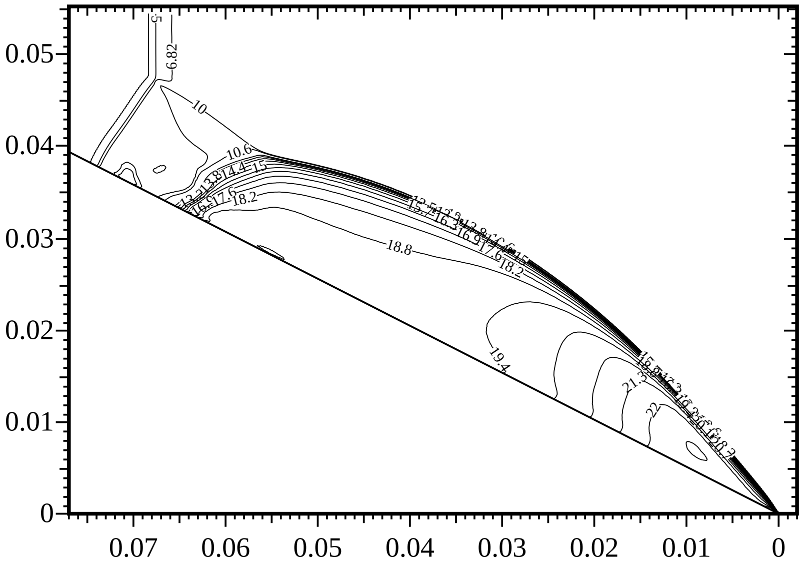

Fig. 2. Double Mach Reflection at $T = 20\mu s$

DIMEX-RK method is always more efficient. The percentage gain (Row *% Gain*) is in the range $[30,\ 60]\%$.

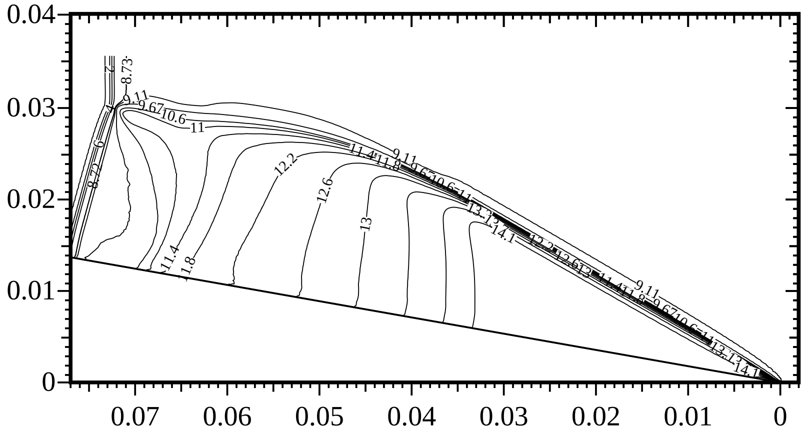

Fig. 3. Complex Mach Reflection at $T = 24\mu s$

Table 5
Perfomance of ERK(2) and MDP(1,2,2) schemes on a 2-D reactive compressible flow calculation; the CPU times are in seconds.

Problem	SINGLE		COMPLEX		DOUBLE	
Method	ERK(2)	MDP	ERK(2)	MDP	ERK(2)	MDP
#Iter.	1626	771	1512	455	1753	524
Rec. (sec)	3400	1590	1850	556	3040	910
Chem.(sec)	—	—	318	565	513	852
Step (sec)	5980	5360	3650	2540	5990	4130
Tot. Time (sec)	9370	6960	5810	3660	9540	5900
% Gain	35%		59%		61%	

7 Conclusions

In this paper, we studied some theoretical properties of the DIMEX-RK FV integrators that have been developed for multi-dimensional hyperbolic systems in [1]. We demonstrated that this approach produces a time evolution matrix operator which shows a peculiar block structure common to a wide family of numerical fluxes. The underlying block matrices are M-matrices and the analysis which can be carried on within this context allowed to show that simple and efficient resolution algorithms even for high-order schemes can be built. Although this FV scheme has been developed for unstructured mesh calculations all the theoretical results proved here can be straightforwardly extended to schemes defined on structured cartesian or curvilinear meshes.

Finally, the performance of the method has been experimentally investigated. A set of representative time integration schemes up to third order global accuracy applied to a 1-D test case has been considered to measure the approximation accuracy of the method. The computational efficiency of the DIMEX-RK approach when compared to an explicit integration time stepping scheme has been measured on a more complex application concerning with the simulation of a 2-D reactive compressible flow. All these experimental investigations illustrate that this approach performs in good accord with the theoretical predictions.

Acknowledgments

The authors are grateful to Prof. T. Ruggeri and Prof. G. Russo for their useful suggestions.

References

[1] E. Bertolazzi, G. Manzini, Diagonally Implicit–Explicit Runge Kutta methods for multidimensional hyperbolic systems. Part I: Formulation of the method, (submitted).

[2] P. G. Ciarlet, The finite element method for elliptic problems, North-Holland Publishing Company, Amsterdam, Holland, 1980.

[3] E. Bertolazzi, G. Manzini, Algorithm 817 P2MESH: generic object-oriented interface between 2-D unstructured meshes and FEM/FVM-based PDE solvers, ACM TOMS 28 (1) (2002) 101–132.

[4] P. R. Halmos, Finite-Dimensional Vector Spaces, Van Nostrand, Princeton, N.J., 1958.

[5] U. M. Ascher, S. J. Ruuth, R. J. Spiteri, Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations, Appl. Numer. Math. 25 (2-3) (1997) 151–167, special issue on time integration (Amsterdam, 1996).

[6] L. Pareschi, G. Russo, Implicit-Explicit Runge-Kutta methods and applications to hyperbolic systems with relaxation, J. Sci. Comp.To appear.

[7] O. Axelsson, Iterative solution methods, Cambridge University Press, Cambridge, 1994.

[8] A. Berman, R. J. Plemmons, Nonnegative Matrices in the Mathematical Sciences, SIAM, Philadelphia, 1994, (republished in “Classics in Applied Mathematics”).

[9] R. S. Varga, On diagonal dominance arguments for bounding $\| A^{-1} \|_\infty$, Linear Algebra and Appl. 14 (3) (1976) 211–217.

[10] E. Zeidler, Nonlinear Functional Analysis and its Applications, Springer-Verlag, New York, 1986.

[11] E. Bertolazzi, G. Manzini, A triangle-based unstructured finite volume method for chemically reactive hypersonic flows, J. Comput. Phys. 166 (2001) 84–115.

[12] Z. Horvath, Positivity of Runge-Kutta and diagonally split Runge-Kutta methods, Appl. Numer. Math. 28 (1998) 309–326.

[13] S. F. Liotta, V. Romano, G. Russo, Central schemes for balance laws of relaxation type, SIAM J. Numer. Anal. 38 (4) (2000) 1337–1356 (electronic).

[14] A. Kurganov, E. Tadmor, New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations, J. Comput. Phys. 160 (1) (2000) 241–282.

[15] R. S. Varga, Matrix iterative analysis, expanded Edition, Springer-Verlag, Berlin, 2000.

[16] G. H. Golub, C. F. Van Loan, Matrix computations, 3rd Edition, Johns Hopkins University Press, Baltimore, MD, 1996.