IMEX finite volume methods for multi-dimensional hyperbolic systems

E. Bertolazzi

Dipartimento di Ingegneria Meccanica e Strutturale, Università di Trento

via Mesiano 77, Trento, Italy. *E-mail: enrico.bertolazzi@ing.unitn.it*

G. Manzini

Istituto di Analisi Numerica – CNR

via Ferrata 1, Pavia, Italy. *E-mail: gianmarco.manzini@ian.pv.cnr.it*

Abstract

A general framework for the semi-implicit discretization of multidimensional conservative hyperbolic systems is proposed. The discretization approach is based on the method-of-line strategy. The spatial discretization uses an unstructured Finite Volume (FV) technique, and a non-oscillatory reconstruction procedure to provide a spatial accuracy of order higher than one. The time derivative is discretized by an Implicit-Explicit Runge-Kutta (IMEX-RK) stepping scheme. The resulting matrix operators are analized within the framework of the M-matrix theory. Sufficient conditions for positive-in-the-mean discrete solution are derived. It is also proved that the non-linear implicit problem, whose solution is needed by the IMEX-RK approach, always admits a unique solution under general hypothesis. Several numerical examples illustrates the behavior of the method.

1 Introduction

Convected-dominated flows are usually described in terms of the following multidimensional hyperbolic system in divergence conservative form

$$\mathbf{U}_t + \nabla \cdot \mathbf{F} = \mathbf{Q}, \qquad \text{in } \Omega \times (0,\, T), \tag{1}$$

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 $\mathbb{R}^m$. The r.h.s vector term $\mathbf{Q}$ is representative of the stiff sources that are eventually present in the flow model, and is generally used to take into account the interactions among different conservative variables.

An Initial Boundary Value Problem (IBVP) is defined by providing system (1) with an initial solution $\mathbf{U}|_{t=0} = \mathbf{U}_0$, where $\mathbf{U}_0$ is a vector function from $\mathbb{R}^d$ into $\mathfrak{U} \subseteq \mathbb{R}^m$, and with a suitable set of problem dependent boundary conditions. Let us denote by

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

the set of d Jacobian matrices of the flux vector functions in (1). Equation (1) is an hyperbolic system if the matrix $\mathbf{J}(\mathbf{U}, \mathbf{n}) = \mathbf{n} \cdot \mathbf{J}(\mathbf{U})$ has m real eigenvalues $\lambda_1(\mathbf{U}, \mathbf{n}) \leq \cdots \leq \lambda_m(\mathbf{U}, \mathbf{n})$ and a complete set of eigenvectors for any $\mathbf{U} \in \mathfrak{U}$ and $\mathbf{n} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$. *No strict hyperbolicity assumption will be retained in this work.*

The numerical schemes that we investigate in this work can be naturally built in the context of the recently developed IMEX-RK schemes [2, 1, 16] when the physical flux vector function can be decomposed into the sum of a *convective* and a *non-convective* part, that is

$$\mathbf{F}(\mathbf{U}) = \mathbf{F}^{(c)}(\mathbf{U}) + \mathbf{F}^{(nc)}(\mathbf{U}).$$

In a wide range of convected-dominated problems $\mathbf{F}^{(c)}(\mathbf{U})$ can be written in terms of a suitable non-linear *convective velocity field*, namely $\mathbf{v}(\mathbf{U})$. Hence, we introduce the following assumption.

ASSUMPTION 1 The flux vector function $\mathbf{F}(\mathbf{U})$ can be written as

$$\mathbf{F}(\mathbf{U}) = \mathbf{U} \otimes \mathbf{v}(\mathbf{U}) + \mathbf{F}^{(nc)}(\mathbf{U}), \tag{2a}$$

$$\lambda_1(\mathbf{U}, \mathbf{n}) \leq \mathbf{n} \cdot \mathbf{v}(\mathbf{U}) \leq \lambda_m(\mathbf{U}, \mathbf{n}), \tag{2b}$$

Assumption 1 is satisfied, e.g., by the Compressible Gas Dynamic Euler Equations and the Shallow Water Equations. Notice that assumption has a direct physical interpretation because the non-convective flux $\mathbf{F}^{(nc)}(\mathbf{U})$ is the pressure contribution to the physical flux RUGGIERO.

In order to define a suitable numerical flux $\boldsymbol{H}(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n})$ that approximates the normal interface flux $\mathbf{F}(\mathbf{U}) \cdot \mathbf{n}$ we assume what follows.

ASSUMPTION 2 The numerical flux $\boldsymbol{H}(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n})$ can be split in the contribution of a convective and a non-convective part as

$$\boldsymbol{H}(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}) = \boldsymbol{H}^{(c)}(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}) + \boldsymbol{H}^{(nc)}(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}),$$

where $\boldsymbol{H}^{(c)}$ is a numerical model for $\mathbf{F}^{(c)}$ and $\boldsymbol{H}^{(nc)}$ for $\mathbf{F}^{(nc)}$.

Following a method-of-line approach, we numerically solve (1) by combining a shock-capturing cell-centered FV spatial discretization of the divergence term and an IMEX-RK scheme for the semi-discrete scheme where an implicit discretization is done to $\boldsymbol{H}^{(c)}(\mathbf{U})$ and an explicit one to $\boldsymbol{H}^{(nc)}(\mathbf{U})$.

We introduce a general definition for the numerical fluxes named *generalized upwind form*, whose properties are discussed in the paper. We claim that this condition is not too restrictive because it is satisfied by many important numerical fluxes available in literature (HLLE fluxes, Steger-Warming, Van Leer, AUSM).

All of these coupled IMEX-RK FV integrators are strictly conservative, shock-capturing, formally n-th order accurate in space and time, and do not require the evaluation of any Jacobian matrix.

We demonstrate that some peculiar block structures can be identified in the time evolution matrix operator and that the matrices underlying them are M-matrices. The existence and uniqueness of the solution of the non-linear algebraic problem arising at each internal stage of the IMEX-RK scheme is proved, moreover an economic iterative approximation at the correct time accuracy can be easily estimated. We outline that simplicity and efficiency are of fundamental importance in devising effective numerical methods for real problems, especially in the 3-D case.

Since the simplest global first-order accurate scheme coincides with the Forward-Backward Euler scheme for Ordinary Differential Equations, it readily follows that the lowest order approximation is unconditionally stable and positive. Both these features are lost when the order is increased in time and space, but we can demonstrate that positivity still holds under a CFL-like condition which is comparable to the one ensuring the calculation stability.

2 The generalized upwind form

Let us consider the transport of a scalar quantity u by the constant vector field $\mathbf{v}$:

$$u_t + \nabla \cdot (\mathbf{v}\, u) = 0. \tag{3}$$

The first-order upwind flux for this problem reads as

$$H(u, v, \mathbf{n}) = a(\mathbf{n})u - a(-\mathbf{n})v, \tag{4}$$

where $a(\mathbf{n}) = (\mathbf{v}\cdot\mathbf{n})^+$ and $(x)^+ = (x+|x|)/2$. The "upwind" velocity $a(\mathbf{n})$ is such that

$$a(\mathbf{n}) \geq 0, \qquad \text{(non-negativity)}$$
$$\mathbf{v}\cdot\mathbf{n} = a(\mathbf{n}) - a(-\mathbf{n}). \qquad \text{(consistency)}$$

Using flux (4) in an FV discretization of (3) and taking properly into account the boundary conditions, the resulting numerical scheme has several interesting numerical properties such as, for example, that it satisfies a discrete maximum principle. Let us introduce the following basic definition to generalize these issues.

Definition 1 A numerical flux $\boldsymbol{H}(\mathbf{U},\mathbf{V},\mathbf{n}) = \boldsymbol{H}^{(c)}(\mathbf{U},\mathbf{V},\mathbf{n}) + \boldsymbol{H}^{(nc)}(\mathbf{U},\mathbf{V},\mathbf{n})$ admits a *generalized upwind form* when it satisfies

$$\boldsymbol{H}^{(c)}(\mathbf{U},\mathbf{V},\mathbf{n}) = \boldsymbol{a}(\mathbf{U},\mathbf{V},\mathbf{n})\mathbf{U} - \boldsymbol{a}(\mathbf{V},\mathbf{U},-\mathbf{n})\mathbf{V}, \tag{5}$$

and

$$\boldsymbol{a}(\mathbf{U},\mathbf{V},\mathbf{n}) \geq 0 \qquad \text{(non-negativity)}$$
$$\mathbf{v}(\mathbf{U})\cdot\mathbf{n} = \boldsymbol{a}(\mathbf{U},\mathbf{U},\mathbf{n}) - \boldsymbol{a}(\mathbf{U},\mathbf{U},-\mathbf{n}) \qquad \text{(consistency)}$$
$$|\boldsymbol{a}(\mathbf{U},\mathbf{V},\mathbf{n}) - \boldsymbol{a}(\mathbf{U}',\mathbf{V}',\mathbf{n})| \leq L\left(\|\mathbf{U}-\mathbf{U}'\| + \|\mathbf{V}-\mathbf{V}'\|\right) \qquad \text{(Lipschitzianity)}$$

Many important fluxes in literature fulfill the condition required by definition 1. For instance, the Rusanov-like numerical fluxes, which has the following *viscosity form*

$$\boldsymbol{H}(\mathbf{U}_L,\mathbf{U}_R,\mathbf{n}) = \frac{1}{2}\left[\mathbf{n}\cdot\mathbf{F}(\mathbf{U}_L) + \mathbf{n}\cdot\mathbf{F}(\mathbf{U}_R)\right] - \frac{1}{2}\xi(\mathbf{U}_L,\mathbf{U}_R,\mathbf{n})(\mathbf{U}_R-\mathbf{U}_L),$$

can be rewritten in generalized upwind form, provided that

$$2\,\boldsymbol{a}(\mathbf{U}_L,\mathbf{U}_R,\mathbf{n}) = \mathbf{v}(\mathbf{U}_L)\cdot\mathbf{n} + \xi(\mathbf{U}_L,\mathbf{U}_R,\mathbf{n}),$$
$$2\,\boldsymbol{H}^{(nc)}(\mathbf{U}_L,\mathbf{U}_R,\mathbf{n}) = \mathbf{F}^{(nc)}(\mathbf{U}_L)\cdot\mathbf{n} + \mathbf{F}^{(nc)}(\mathbf{U}_R)\cdot\mathbf{n}.$$

The HLLE-like numerical fluxes discussed in References [5, 6, 7, 17] can be re-formulated in generalized upwind form by setting

$$(S_R - S_L)\boldsymbol{a}(\mathbf{U}_L,\mathbf{U}_R,\mathbf{n}) = S_R\left(\mathbf{v}(\mathbf{U}_L)\cdot\mathbf{n} - S_L\right),$$
$$(S_R - S_L)\boldsymbol{H}^{(nc)}(\mathbf{U}_L,\mathbf{U}_R,\mathbf{n}) = \left(S_R\mathbf{F}^{(nc)}(\mathbf{U}_L) - S_L\mathbf{F}^{(nc)}(\mathbf{U}_R)\right)\cdot\mathbf{n},$$

where $S_R = S(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n})$ and $S_L = -S(\mathbf{U}_R, \mathbf{U}_L, -\mathbf{n})$. Some possible choices of S can be found in [3]. For example, using

$$S(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}) = \max\{\lambda_m(\mathbf{U}_L, \mathbf{n}),\ \lambda_m(\mathbf{U}_R, \mathbf{n}),\ 0\}$$

we obtain the *local Lax-Friedrichs* numerical flux [9, 10]. It is interesting to notice that also the Steger & Warming flux splitting [18], the Van Leer flux splitting [11, 12, 19] and the AUSM+ flux splitting [13, 20] are writable in accord with definition 1. See for major details Reference [3].

3 The semi-discrete Finite Volume formulation

The computational mesh is defined as the set of n non-overlapping *cells* covering the domain $\Omega \in \mathbb{R}^d$. Cells are *intervals* in the 1-D case, *triangles* in 2-D and *tetrahedrons* in 3-D. The mesh is assumed to be *regular* and *conformal* in the sense specified by [4].

Cells are conventionally labeled by an integer identifier ranging from 1 to n. For the generic cell $\mathfrak{C}_i$ we indicate by $|\mathfrak{C}_i|$ the d dimensional measure of the cell (volume in 3-D, area in 2-D, length in 1-D).

The intersection with 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 cells $\mathfrak{C}_i$ and $\mathfrak{C}_j$ will be addressed by the pair ij and denoted by the symbol $\mathfrak{f}_{ij}$. A boundary face will also be 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. For the generic face $\mathfrak{f}_{ij}$, we indicate by $|\mathfrak{f}_{ij}|$ 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 cell i to cell j if the face is internal and outward directed if the face is on the boundary.

For each cell $\mathfrak{C}_i$, we will denote by $\sigma(i)$ the set of its adjacent cells and by $\sigma'(i)$ the subset of its faces that are on the boundary.

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 $\underline{\mathbf{U}}^T = (\mathbf{U}_1^T, \mathbf{U}_2^T, \ldots, \mathbf{U}_n^T)$, whose i-th block is the m-size vector $\mathbf{U}_i$. For equation (1) the semi-discrete FV numerical scheme takes the form

$$|\mathfrak{C}_i| \frac{d\mathbf{U}_i}{dt} + \sum_{j \in \sigma(i)} \boldsymbol{H}_{ij}(\underline{\mathbf{U}}) + \sum_{j' \in \sigma'(i)} \boldsymbol{H}_{ij'}^{(bc)}(\underline{\mathbf{U}}) = \mathbf{0}, \qquad i = 1, 2, ..., n \qquad (6)$$

The terms $\boldsymbol{H}_{ij}(\underline{\mathbf{U}})$ and $\boldsymbol{H}^{(bc)}_{ij'}(\underline{\mathbf{U}})$ denote the numerical fluxes estimated on the internal and on the boundary faces. Accuracy of order higher than one is formally achieved by using a suitable numerical quadrature for the numerical flux integral and a proper reconstruction procedure to extrapolate face values. Let us write the numerical flux integral as

$$\boldsymbol{H}_{ij}(\underline{\mathbf{U}}) = |\mathfrak{f}_{ij}| \sum_{k=1}^{N_q} \omega_k \boldsymbol{H}(\mathbf{U}^k_{ij}(\underline{\mathbf{U}}), \mathbf{U}^k_{ji}(\underline{\mathbf{U}}), \mathbf{n}_{ij}), \qquad \mathbf{U}^k_{ij}(\underline{\mathbf{U}}) = \mathbf{U}_i(\cdot, \mathbf{x}^k_{ij})$$

where $\mathbf{x}^k_{ij}$ is the k^{th} quadrature point on the face $\mathfrak{f}_{ij}$ and ω_k the quadrature weight, $\mathbf{U}_i(t, \mathbf{x})$ is the reconstructed solution relative the i^{th} element and $\mathbf{U}^k_{ij}(\underline{\mathbf{U}})$ is the face extrapolated value at the quadrature point $\mathbf{x}^k_{ij}$.

ASSUMPTION 3 The reconstructed solution $\mathbf{U}_i(\cdot, \mathbf{x})$ within the i-th cell for $i = 1, \ldots, n$ and $j \in \sigma(i)$ satisfies

$$\min\{\mathbf{U}_i, \mathbf{U}_j\} \leq \mathbf{U}^k_{ij}(\underline{\mathbf{U}}) \leq \max\{\mathbf{U}_i, \mathbf{U}_j\}, \qquad k = 1, \ldots, N_q.$$

ASSUMPTION 4 All of the weights used in the quadrature formulae are *positive*, i.e. $\omega_i > 0$ for all $i = 1, \ldots, N_q$.

This last assumption is quite naturally satisfied by almost all of the quadrature formulae, such as the Gaussian ones.

Introducing the generalized upwind form (5) into (6) yields the final form of the basic semi-discrete FV scheme:

$$\begin{aligned} |\mathfrak{C}_i| \frac{d\mathbf{U}_i}{dt} + & \sum_{j \in \sigma(i)} \left(\mathsf{a}_{ij}(\underline{\mathbf{U}})\mathbf{U}_i - \mathsf{a}_{ji}(\underline{\mathbf{U}})\mathbf{U}_j \right) \\ & + \sum_{j \in \sigma(i)} \left(\widetilde{\boldsymbol{H}}^{(c)}_{ij}(\underline{\mathbf{U}}) + \boldsymbol{H}^{(nc)}_{ij}(\underline{\mathbf{U}}) \right) + \sum_{j' \in \sigma'(i)} \boldsymbol{H}^{(bc)}_{ij'}(\underline{\mathbf{U}}) = \mathbf{0}, \end{aligned} \tag{7}$$

where

$$\mathsf{a}^k_{ij}(\underline{\mathbf{U}}) = \mathsf{a}(\mathbf{U}^k_{ij}(\underline{\mathbf{U}}), \mathbf{U}^k_{ji}(\underline{\mathbf{U}}), \mathbf{n}_{ij}), \tag{8a}$$

$$\mathsf{a}_{ij}(\underline{\mathbf{U}}) = |\mathfrak{f}_{ij}| \sum_{k=1}^{N_q} \omega_k \mathsf{a}^k_{ij}(\underline{\mathbf{U}}), \tag{8b}$$

$$\widetilde{\boldsymbol{H}}^{(c)}_{ij}(\underline{\mathbf{U}}) = |\mathfrak{f}_{ij}| \sum_{k=1}^{N_q} \omega_k (\mathsf{a}^k_{ij}(\underline{\mathbf{U}})(\mathbf{U}^k_{ij}(\underline{\mathbf{U}}) - \mathbf{U}_i) - \mathsf{a}^k_{ji}(\underline{\mathbf{U}})(\mathbf{U}^k_{ji}(\underline{\mathbf{U}}) - \mathbf{U}_j)) \tag{8c}$$

$$\boldsymbol{H}^{(nc)}_{ij}(\underline{\mathbf{U}}) = |\mathfrak{f}_{ij}| \sum_{k=1}^{N_q} \omega_k \boldsymbol{H}^{(nc)}(\mathbf{U}^k_{ij}(\underline{\mathbf{U}}), \mathbf{U}^k_{ji}(\underline{\mathbf{U}}), \mathbf{n}_{ij}). \tag{8d}$$

Notice that the a-term in (7) has been split in two contributions, namely a_{ij} and $\widetilde{\boldsymbol{H}}_{ij}^{(c)}$. The former one is the scalar a-function estimated in the quadrature nodes by using the cell reconstructions and contributes to the linear advective term in (5). The latter one takes into account how much the reconstructed values at quadrature nodes differ from the cell-averages solutions and added to the $\boldsymbol{H}^{(nc)}$-term contributes to the non-linear term in (5). When a 1-st order piecewise constant reconstruction is considered we have that $\widetilde{\boldsymbol{H}}_{ij}^{(c)} = 0$.

3.1 Matrix notation for the semi-discrete scheme

Block vectors and block matrices will be denoted by underlined bold symbols, and their blocks by a proper set of subscript indices. For example, $\underline{\mathbf{U}}$ and $\underline{\widetilde{\mathbf{A}}}$ indicate a block vector and a block matrix, and $\mathbf{U}_i$ and $\widetilde{\mathbf{A}}_{ij}$ are the i-th sub-vector block of $\underline{\mathbf{U}}$ and the ij-th block of $\underline{\widetilde{\mathbf{A}}}$. Let us introduce the $n \times n$ matrix $\mathbf{A}(\underline{\mathbf{U}})$ defined as

$$A_{ij}(\underline{\mathbf{U}}) = \begin{cases} -a_{ji}(\underline{\mathbf{U}}) & \text{if } ij \text{ addresses a mesh face, namely } \mathfrak{f}_{ij}, \\ \sum_{l \in \sigma(i)} a_{il}(\underline{\mathbf{U}}), & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

the diagonal matrix $\mathbf{D} = \mathrm{diag}(|\mathfrak{C}_1|, \ldots, |\mathfrak{C}_N|)$, whose i-th component is the d-measure of the cell $\mathfrak{C}_i$, and the block vector $\underline{\mathbf{b}}(\underline{\mathbf{U}})^T = [\,\mathbf{b}_1(\underline{\mathbf{U}})^T, \ldots, \mathbf{b}_n(\underline{\mathbf{U}})^T\,]$, whose i-th block is given by

$$\mathbf{b}_i(\underline{\mathbf{U}}) = -\sum_{j \in \sigma(i)} \widetilde{\boldsymbol{H}}_{ij}^{(c)}(\underline{\mathbf{U}}) - \sum_{j' \in \sigma'(i)} \boldsymbol{H}_{ij'}^{(bc)}(\underline{\mathbf{U}}). \tag{10}$$

In view of (6) and using relations (8a-d), we can reformulate the $\widetilde{\boldsymbol{H}}_{ij}^{(c)}$ contribution of (8c). From assumption 3 it follows that there exists $m \times m$ diagonal matrices $\boldsymbol{\alpha}_{ij}^k$ such that

$$\mathbf{U}_{ij}^k(\underline{\mathbf{U}}) = (\mathbf{I}_m - \boldsymbol{\alpha}_{ij}^k)\mathbf{U}_i + \boldsymbol{\alpha}_{ij}^k \mathbf{U}_j, \qquad \mathbf{0}_m \leq \boldsymbol{\alpha}_{ij}^k \leq \mathbf{I}_m,$$

and substituting this latter expression in (8b) we obtain

$$\sum_{j \in \sigma(i)} \widetilde{\boldsymbol{H}}_{ij}^{(c)} = \sum_{j \in \sigma(i)} \widetilde{\mathbf{a}}_{ij}(\underline{\mathbf{U}})(\underline{\mathbf{U}}_i - \underline{\mathbf{U}}_j),$$

where

$$\widetilde{\mathbf{a}}_{ij}(\underline{\mathbf{U}}) = |\mathfrak{f}_{ij}| \sum_{k=1}^{N_q} \omega_k (a_{ij}^k(\underline{\mathbf{U}})\boldsymbol{\alpha}_{ij}^k + a_{ji}^k(\underline{\mathbf{U}})\boldsymbol{\alpha}_{ji}^k).$$

From assumption 4 and the definition of $a_{ij}^k(\underline{\mathbf{U}})$ given in (8a), there follows that the $m \times m$-size diagonal matrices $\widetilde{\mathbf{a}}_{ij}(\underline{\mathbf{U}})$ are non-negative, that is $\widetilde{\mathbf{a}}_{ij}(\underline{\mathbf{U}}) = \widetilde{\mathbf{a}}_{ji}(\underline{\mathbf{U}}) \geq \mathbf{0}$. Let us now introduce the $nm \times nm$-size block matrix $\widetilde{\underline{\mathbf{A}}}(\underline{\mathbf{U}})$, whose $m \times m$-size block $\widetilde{\mathbf{A}}_{ij}(\underline{\mathbf{U}})$ is defined as

$$\widetilde{\mathbf{A}}_{ij}(\underline{\mathbf{U}}) = \begin{cases} -\widetilde{\mathbf{a}}_{ji}(\underline{\mathbf{U}}) & \text{if } ij \text{ addresses a mesh face, namely } \mathfrak{f}_{ij}, \\ \sum_{l \in \sigma(i)} \widetilde{\mathbf{a}}_{il}(\underline{\mathbf{U}}) & \text{if } i = j. \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Finally, by using the definitions introduced in (9–11) and the diagonal matrix $\mathbf{D}$, we can reformulate the semi-discrete FV scheme (6) in the more compact matrix form

$$(\mathbf{D} \otimes \mathbf{I}_m) \frac{d\,\underline{\mathbf{U}}}{d\,t} = \left[\widetilde{\underline{\mathbf{A}}}(\underline{\mathbf{U}}) - \mathbf{A}(\underline{\mathbf{U}}) \otimes \mathbf{I}_m\right] \underline{\mathbf{U}} + \underline{\mathbf{b}}(\underline{\mathbf{U}}). \tag{12}$$

4 IMEX-RK discretization in time

Let us rewrite equation (12) as

$$\frac{d\,\underline{\mathbf{U}}}{d\,t} = -\underline{\mathbf{a}}\,(\underline{\mathbf{U}}) + \underline{\mathbf{b}}\,(\underline{\mathbf{U}})\,, \tag{13}$$

where

$$\begin{aligned} \mathcal{A}(\underline{\mathbf{U}}) &= \mathbf{D}^{-1}\mathbf{A}(\underline{\mathbf{U}}) \otimes \mathbf{I}_m, \\ \underline{\mathbf{a}}\,(\underline{\mathbf{U}}) &= \mathcal{A}(\underline{\mathbf{U}})\underline{\mathbf{U}}, \\ \underline{\mathbf{b}}\,(\underline{\mathbf{U}}) &= \left(\mathbf{D}^{-1} \otimes \mathbf{I}_m\right) \left(\underline{\mathbf{b}}(\underline{\mathbf{U}}) + \widetilde{\underline{\mathbf{A}}}(\underline{\mathbf{U}})\underline{\mathbf{U}}\right). \end{aligned} \tag{14}$$

The IMEX-RK method provides a natural framework to discretize (13). In fact, since $\underline{\mathbf{a}}\,(\underline{\mathbf{U}})$ contains the transport contribution to the flux, it can be interpreted as the *stiff* part of the system of ODEs (13), and an implicit discretization can be sought. On the other hand, the second r.h.s. term, i.e. $\underline{\mathbf{b}}\,(\underline{\mathbf{U}})$, can be discretized explicitly.

The simplest IMEX-RK scheme is the semi-implicit Euler method, which is 1-st order accurate in time. It takes the form

$$\underline{\mathbf{U}}^{n+1} + \Delta t\,\underline{\mathbf{a}}\left(\underline{\mathbf{U}}^{n+1}\right) = \underline{\mathbf{U}}^{n} + \Delta t\,\underline{\mathbf{b}}\left(\underline{\mathbf{U}}^{n}\right). \tag{15}$$

The general r-stage IMEX-RK scheme can be formulated as:

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

$$\underline{\mathbf{W}}^i + \Delta t\, \alpha'_{ii}\, \underline{\mathbf{a}}\left(\underline{\mathbf{W}}^i\right) = \underline{\mathbf{U}}^n + \Delta t \sum_{j=1}^{i-1} \left(\alpha_{ij} \underline{\mathbf{b}}\left(\underline{\mathbf{W}}^j\right) - \alpha'_{ij} \underline{\mathbf{a}}\left(\underline{\mathbf{W}}^j\right)\right), \quad (16a)$$

— then compute

$$\underline{\mathbf{U}}^{n+1} = \underline{\mathbf{U}}^n + \Delta t \sum_{i=1}^{r+1} \left(\omega_i \underline{\mathbf{b}}\left(\underline{\mathbf{W}}^i\right) - \omega'_i \underline{\mathbf{a}}\left(\underline{\mathbf{W}}^i\right)\right). \quad (16b)$$

Equations (16a-b) will be used used for positivity considerations. Specific values of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$, $\boldsymbol{\omega}$, $\boldsymbol{\omega}'$ can be found in [1, 16].

The IMEX-RK method in (16a-b) requires the solution of r non-linear systems of the form

$$\underline{\mathbf{W}} + \Delta t\, a\, \underline{\mathbf{a}}\left(\underline{\mathbf{W}}\right) = \mathbf{r}, \qquad a > 0, \quad (17)$$

where $\mathbf{r}$ is the right-hand-side of (16a). Let us define the map

$$\boldsymbol{\Phi}(\underline{\mathbf{W}}) = \left(\mathbf{I} + \Delta t\, a\, \mathcal{A}(\underline{\mathbf{W}})\right)^{-1} \mathbf{r}, \quad (18)$$

where $\mathcal{A}(\underline{\mathbf{W}})$ has been introduced in (14); $\underline{\mathbf{W}}$, then, is a solution to (17) if and only if it is a fixed point of (18).

The main results of this section are summarized in the two following theorems.

Theorem 1 *The map* (18) *admits a fixed point for all* $\Delta t > 0$. *The fixed point is unique if*

$$\Delta t < \frac{1}{L\, a\, \kappa^2 \left\|\mathbf{r}\right\|_1}, \qquad \kappa = \frac{\max_{i=1,2\ldots,n} \left|\mathfrak{C}_i\right|}{\min_{i=1,2\ldots,n} \left|\mathfrak{C}_i\right|},$$

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

PROOF see [3].

Equation (17) depends on $\underline{\mathbf{W}}$ in a non-linear fashion, and thus the IMEX-RK method as it has been proposed so far may be very expensive from the computational point of view. However, as stated in the following theorem, the solution $\underline{\mathbf{W}}$ can be approximated up to order $\mathcal{O}\left(\Delta t^{k+1}\right)$ by a straightforward iterative procedure.

Theorem 2 *Let* $\underline{\mathbf{W}}^i$ *for* $i = 0, \ldots, k$ *be defined iteratively as*

$$\begin{aligned} \underline{\mathbf{W}}^0 &= \underline{\mathbf{U}}^n, \\ \underline{\mathbf{W}}^i &= \left(\mathbf{I} + \Delta t\, a\, \mathcal{A}(\underline{\mathbf{W}}^{i-1})\right)^{-1} \mathbf{r}, \qquad i = 1, 2, \ldots, k. \end{aligned} \quad (19)$$

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* (17).

PROOF see [3].

In view of Theorem 2, the solution of the non-linear systems (17) can be substituted by the solution of a *fixed* number of linear systems whose coefficient matrix is an M-matrix. For example, the first-order accurate-in-time semi-implicit Euler scheme (15) can be re-formulated as

$$(\mathbf{I} + \Delta t\, \mathcal{A}(\underline{\mathbf{U}}^n))\underline{\mathbf{U}}^{n+1} = \underline{\mathbf{U}}^n + \Delta t\, \underline{\mathbf{b}}\,(\underline{\mathbf{U}}^n)\,,$$

without any loss of accuracy.

After straightforward algebraic manipulations, it turns out that the recursive procedure (19) formally requires the solution of the k linear algebraic problems

$$\left(\mathcal{M}(\underline{\mathbf{W}}^{i-1}) \otimes \mathbf{I}_m\right) \underline{\mathbf{W}}^i = (\mathbf{D} \otimes \mathbf{I}_m)\mathbf{r} \qquad i = 1, 2, \ldots, k, \tag{20}$$

where

$$\mathcal{M}(\underline{\mathbf{W}}^{i-1}) = \mathbf{D} + \Delta t\, a\, \mathbf{A}(\underline{\mathbf{W}}^{i-1}). \tag{21}$$

This fact is remarkable and allows a noteworthy simplification of the whole solution procedure. Indeed, for any $\mathbf{W}$ there holds that

$$(\mathcal{M}(\mathbf{W}) \otimes \mathbf{I})^{-1} = \mathcal{M}(\mathbf{W})^{-1} \otimes \mathbf{I},$$

and it is clear that at any step i the scheme (19) requires the formal inversion of the matrix (21), which is the same for all of the m linear systems in (20). This matrix has size $n \times n$, thus smaller than the whole coefficient matrix $\mathcal{M}(\mathbf{W}) \otimes \mathbf{I}$, whose size is $m\,n \times m\,n$.

The matrix $\mathcal{M}(\mathbf{W})$ is an M-matrix; the proof of this property is given in [3]. This fact has some very important consequences on computational efficiency in solving the linear algebraic problems (20).

Let us first recall that the M-matrix $\mathbf{A}$ such that $\mathbf{y}^T\mathbf{A} \geq \mathbf{0}$ for some vector $\mathbf{y} \gg \mathbf{0}$ admits an LU factorization whose triangular factors $\mathbf{L}$ and $\mathbf{U}$ are also M-matrices, see Reference [8]. Thus, no numerical pivoting is necessary to ensure stability in the LU factorization process, as would be the case of a general matrix, see Reference [15].

Finally, notice that the structural pattern of the matrix $\mathcal{M}(\mathbf{W})$, i.e. its non-zeros coefficients, only depends on the topological neighborhood relationships among the cells of the mesh. That is, a matrix non-zero element

always corresponds to a connection between two adjacent cells of the mesh. If the mesh does not change in the time stepping calculation, i.e. no automatic grid adaptation is carried out during the run, the non-zero pattern of $\mathcal{M}(\mathbf{W})$ must remain constant. Then, it follows that the symbolic pivoting of the matrix, i.e. the re-ordering of rows and columns to reduce and control the fill-in phenomenon of the factorization, can be performed only once at the beginning of each calculation. When a grid adaption strategy is incorporated, the symbolic pivoting must be recalculated only after updating the mesh. This fact may clearly have a strong impact in reducing both computational costs and memory storage if direct algebraic methods for sparse matrices are used to solve (20).

The previous arguments based on the M-matrix nature of $\mathcal{M}(\mathbf{W})$, still remain valid for incomplete factorizations, i.e. when a direct re-solution method is used to pre-condition an iterative method.

To this purpose we recall that the incomplete LU-decomposition of an M-matrix is at least as stable as the complete decomposition without any numerical pivoting – see Theorem 3.2 in Reference [14].

5 Numerical Results

In this section we focus on the performance of several representative IMEX-RK FV schemes as far as their approximation order in space and time is concerned. The coefficients defining all of the schemes are given in Reference [3].

The approximation orders are estimated on a 1-D test case for the compressible Euler equations. Both the initial density and pressure fields consists of a smooth bell-shaped pulse superimposed to a spatially constant value. This density and pressure pulse is translated by an initially constant velocity field. The extension of the 1-D domain is virtually infinite, but clearly only a small portion of it is computationally represented. The simulation is arrested before the pulse goes out the finite computational domain so as there is no need for a special treatment of the boundary conditions. We propose this test case because it shows a smooth analytical solution, the pulse translated to a new position, which is thus suitable for comparisons with numerical approximations. Major details are in Reference [3].

The time order accuracy 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.

Tables 1-2 reports the time accuracy orders that we measured by using second- and third-order IMEX time stepping schemes. All of the four numerical fluxes discussed in Section 2, that are the *HLLE* flux, the *Steger-Warming* flux splitting, the *Van Leer* flux splitting and the *AUSM+* flux splitting produce about the same results. The ones reported in Tables 1-2 are obtained by using the *HLLE* flux.

The global accuracy is estimated by running three different simulations on meshes with respectively 200, 400 and 800 intervals. The time-step changes during the run in accord with the maximum allowable CFL number.

We use 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 [10].

The results are given in Tables 3-4.

Both time and global convergence orders are quite close to the theoretical ones.

Table 1: Time accuracy convergence rates of second-order accurate schemes.

	ERK(2)	Midpoint(1,2,2)	ARS(2,2,2)	ARS(2,3,2)	LRR(3,3,2)
order	2.048	2.044	2.045	2.045	2.045

Table 2: Time accuracy convergence rates of third-order accurate schemes.

	ERK(3)	ARS(2,3,3)	ARS(4,4,3)
order	3.019	2.989	3.004

6 Conclusions

This paper consider semi-implicit approaches on unstructured meshes in discretizing multidimensional hyperbolic systems. The method that we propose

Table 3: Global convergence rates of second-order accurate schemes.

	HLLE	SW	Van Leer	AUSM+
ERK2	1.947	1.948	1.949	1.950
Midpoint(1,2,2)	1.947	1.947	1.949	1.949
ARS(2,2,2)	1.946	1.947	1.949	1.949
ARS(2,3,2)	1.946	1.948	1.949	1.949
LRR(3,3,2)	1.946	1.947	1.949	1.948

Table 4: Global convergence rates of third-order accurate schemes.

	HLLE	SW	Van Leer	AUSM+
ERK3	2.963	2.982	2.942	2.953
ARS(2,3,3)	2.984	2.940	2.995	2.970
ARS(4,4,3)	2.943	3.091	2.974	2.960

has been developed to perform integration of IBVPs in conservative divergence form. It is based on a special splitting of the physical flux vector function into a convective and a non-convective part. In the framework of IMEX-RK schemes, the convective part is discretized in an implicit way, while the non-convective one in an explicit way.

It can be demonstrated that this approach produces a time evolution matrix operator which shows a peculiar block structure common to a wide family of spatial FV discretization. The underlying block matrices are M-matrices and an analysis of the method can be carried out within this context, thus resulting in simple and efficient resolution algorithms even for high-order schemes.

The IMEX-RK FV schemes proposed in this paper are especially suitable when an implicit scheme must be used even in evolutive problems. This occurs, for instance, when fast reactive processes take place in the flow or when flows of very high speed are confined in restricted regions and at the same time of low speed elsewhere. Regions of high speed flows may in fact drastically reduce the time step size allowable in an explicit formulation,

consequently degrading the performance of a numerical flow solver from the viewpoint of the computational costs. On the other hand, an implicit first-order discretization in time produces a poor accurate approximation of the time-dependent solution. Thus, higher order semi-implicit techniques should be devised, which would greatly improve the overall accuracy of the model predictions.

Finally, we emphasize that all of the theoretical results presented in this paper are valid for both 2-D and 3-D unstructured meshes.

References

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

[2] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 797–823.

[3] E. Bertolazzi and G. Manzini, *High-order IMEX-RK finite volume methods for multidimensional hyperbolic systems*, Tech. Rep. 1202, I.A.N.-C.N.R., 2001.

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

[5] S. F. Davis, *Simplified second-order Godunov-type methods*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 445–473.

[6] B. Einfeldt, *On Godunov-type methods for gas dynamics*, SIAM J. Numer. Anal., 25 (1988), pp. 294–318.

[7] B. Einfeldt, C.-D. Munz, P. L. Roe, and B. Sjögreen, *On Godunov-type methods near low densities*, J. Comput. Phys., 92 (1991), pp. 273–295.

[8] R. E. Funderlic and R. J. Plemmons, *LU decomposition of M-matrices by elimination without pivoting*, Linear Algebra and Appl., 41 (1981), pp. 99–110.

[9] A. Kurganov and G. Petrova, *Central schemes and contanct discontinuities*, M2AN. Mathematical Modelling and Numerical Analysis, 34 (2000), pp. 1259–1275.

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

[11] B. V. Leer, *Towards the ultimate conservative difference scheme*, J. Comput. Phys., 32 (1979), pp. 101–136.

[12] B. V. Leer, *Flux-vector splitting for the Euler equations*, Lecture Notes in Physics, 170 (1982), pp. 507–512.

[13] M.-S. Liou, *A sequel to AUSM: AUSM$^+$*, J. Comput. Phys., 129 (1996), pp. 364–382.

[14] J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.

[15] M. Neumann, *On the Schur complement and the LU-factorization of a matrix*, Linear and Multilinear Algebra, 9 (1980/81), pp. 241–254.

[16] L. Pareschi and G. Russo, *Implicit-explicit runge-kutta schemesfor stiff systems of differential equations*, in Recent Trends in Numerical Analysis, L.Brugnano and D.Trigiante, eds., vol. 3, 2000, pp. 269–289.

[17] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes. II*, J. Comput. Phys., 83 (1989), pp. 32–78.

[18] J. L. Steger and R. F. Warming, *Flux-vector splitting of the inviscid gas dynamic equations with application to finite-difference methods*, Journal of Computational Physics, 40 (1981), pp. 263–293.

[19] B. van Leer, *Lecture Notes in Physics*, vol. 170, Sringer Verlag, 1992.

[20] Y. Wada and M.-S. Liou, *An accurate and robust flux splitting scheme for shock and contact discontinuities*, SIAM J. Sci. Comput., 18 (1997), pp. 633–657.