An asymptotic preserving scheme for the M1 model on conical meshes

Xavier Blanc, Philippe Hoch, Clément Lasuen

An asymptotic preserving scheme for the M_1 model on conical meshes

Xavier Blanc[1], Philippe Hoch[2], Clément Lasuen[2]

[1]Université de Paris, CNRS, Sorbonne Université,
Laboratoire Jacques-Louis Lions (LJLL), F-75006 Paris,
xavier.blanc@u-paris.fr

[2]CEA, DAM, DIF, F-91297 Arpajon, France
philippe.hoch@cea.fr clement.lasuen@cea.fr

January 21, 2022

Abstract

This work focuses on the design of a $2D$ numerical scheme for the M_1 model on conical meshes. This model is nonlinear and approximates the firsts moments of the radiative transfert equation using an entropic closure. Besides, this model admits a diffusion limit as the cross section increases. It is important for the numerical scheme to be consistent with this limit, that is to say, it has to be *asymptotic preserving* or AP. Such a scheme already exists on polygonal meshes and our work consisted in adapting it to conical meshes. After having introduced conical meshes, we explain the construction of the scheme. It is based on an analogy between the M_1 model and the Euler gas dynamic system. We also present a second order reconstruction procedure and we apply it on both polygonal and conical meshes. Moreover, we prove that the scheme converges toward a limit scheme in the diffusion limit. In the last section, some numerical test cases are given so as to compare the polygonal and conical schemes. The limit scheme is studied and we observed numerically that it is consistent with the diffusion equation. Eventually, the limit scheme is compared to a limit scheme coming from another moment model for the radiative transfer equation (namely, the P_1 model).

Contents

1 Introduction **3**

2 Conical meshes **4**
2.1 Rational quadratic Bezier curve . . . 4
2.2 Computing the area of a conical cell . . . 6

3 Reformulation of the model **9**

4 Numerical method **9**
4.1 Partially implicit time discretisation . . . 13
4.2 Second order reconstruction . . . 14
4.3 Diffusion limit scheme . . . 15

5 Theoretical study of the scheme **16**
5.1 Notation . . . 16
5.2 Assumptions on the mesh . . . 19
5.3 Proof of lemma 5.2 . . . 20
5.3.1 First step . . . 20
5.3.2 Second step . . . 20
5.3.3 Conclusion of the proof . . . 21
5.3.4 Continuity of $\mathbf{u}_{\text{dof}}/\varepsilon$ . . . 22
5.4 Proof of lemma 5.3 . . . 22
5.5 Proof of lemma 5.4 . . . 22
5.6 Proof of lemma 5.5 . . . 23
5.7 Proof of lemma 5.6 . . . 24
5.8 Proof of theorem 5.7 . . . 24
5.9 Invertibility of the matrix β^r . . . 26

6 Numerical results **26**
6.1 Propagation of a Dirac mass . . . 27
6.2 Streaming regime test cases . . . 27
6.2.1 Transport of a step function . . . 28
6.2.2 Smooth solution . . . 28
6.2.3 Singular solution . . . 30
6.3 Diffusion regime . . . 31
6.3.1 $1D$ test case . . . 31
6.3.2 Propagation of a Dirac mass in the diffusion limit . . . 32
6.3.3 Fundamental solution of the diffusion equation . . . 36
6.4 Comparison with the limit scheme of another model . . . 40

7 Conclusion **45**

8 Annex **46**
8.1 Upwind scheme for the transport equation . . . 46

1 Introduction

In this article, we focus on the following system, called M_1 system:

$$\begin{cases} \partial_t E + \dfrac{1}{\varepsilon}\text{div } \mathbf{F} = 0, \\ \partial_t \mathbf{F} + \dfrac{1}{\varepsilon}\text{div } P = -\dfrac{\sigma}{\varepsilon^2}\mathbf{F}, \end{cases} \tag{1}$$

where ε is a positive coefficient, $\sigma > 0$ is the opacity. The time variable is denoted by $t \geq 0$ and the space variable is $\mathbf{x} \in \mathbb{R}^2$. The unknowns of system (1) are the radiative energy $E(t, \mathbf{x})$ and the radiative flux $\mathbf{F}(t, \mathbf{x}) \in \mathbb{R}^2$. The pressure tensor $P(t, \mathbf{x}) \in \mathbb{R}^{2\times 2}$ depends on E and $\mathbf{F}$ and writes :

$$P = \frac{E}{2}\left((1-\chi(\mathbf{f}))\, I_2 + (3\chi(\mathbf{f})-1)\,\frac{\mathbf{f}\otimes\mathbf{f}}{\|\mathbf{f}\|^2}\right), \qquad \chi(\mathbf{f}) = \frac{3+4\|\mathbf{f}\|^2}{5+2\sqrt{4-3\|\mathbf{f}\|^2}}, \tag{2}$$

where $\mathbf{f} = \mathbf{F}/E$ is the dimensionless flux, I_2 is the identity matrix of size 2 and $\chi(\mathbf{f})$ is called the Eddington factor. Furthermore the model satisfies $E(t, \mathbf{x}) > 0$, $\|\mathbf{f}(t, \mathbf{x})\| \leq 1$ and as ε goes to 0, the radiative flux $\mathbf{F}$ vanishes and E converges toward the solution of the diffusion equation :

$$\partial_t \tilde{E} - \text{div}\left(\frac{1}{3\sigma}\nabla\tilde{E}\right) = 0. \tag{3}$$

This can be seen using a standard Hilbert expansion in powers of ε. A rigorous proof is given in the $1D$ case in [1]. It is important for a numerical scheme that discretises system (1) to be consistent with this diffusion limit. Such a scheme is called *asymptotic preserving* (AP).

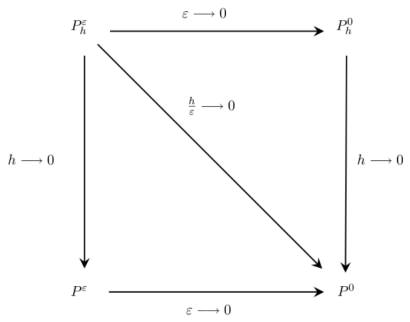

Figure 1: Definition of an AP scheme.

Figure 1 illustrates this property. The discretisation parameter is denoted by h, ε is the physical parameter aimed at vanishing and which reflects the convergence of the model P^ε toward the limit model P^0. A scheme P_h^ε consistant with P^ε is said to be *asymptotic preserving* is the scheme P_h^0 computed in the limit $\varepsilon \to 0$ is consistent with the limit model P^0.

Originally, the M_1 model was derived and studied in [2]. We refer to [3, 4] (see also [5]) for its mathematical properties, and to [6, 7] for related modelling considerations. As we already mentioned, the M_1 model is based on the radiative transfer equation. Taking the two first moments of this equation with respect to the velocity, we find system (1), where the matrix P is *a priori* not known. Then, assuming that I is the minimum entropy distribution with first moments equal to E and $\mathbf{F}$, allows for an explicit expression of I, hence of P, as functions of E and F, namely (2). In [2], numerical tests were given in dimension 1, with an HLL-type scheme. Such a discretization is in general not asymptotic preserving [8]. In [9], an AP finite volume scheme was proposed for this model, in dimension 1, using an upwind discretization, together with ideas of [10]. In [5, 11, 12], modifications of HLL-type fluxes were proposed to derive an AP scheme in dimension 2 on cartesian grids. A discontinuous Garlerkin approach satisfying positivity and flux limitation in dimension 1 was also considered in [13]. In [14, 15] (see also [16]), an AP scheme on deformed meshes in

dimension 2 was proposed, based on nodal Riemann solvers [17, 18] and the method of Jin and Levermore [8]. This method is our starting point for deriving an asymptotic preserving scheme on conical meshes.
Some recent works have also focused on higher order AP schemes for the M_1 model, as for instance [19], with a Discontinuous Galerkin approach.
The method of entropy minimum closure has also been extended, first in the radiation setting, by including more moments (this gives the so-called M_N model). Note however that only $N = 1$ allows for an explicit expression of the closure. In all other cases, the closure must be performed through a numerical method, inducing an additional numerical cost and loss of accuracy. For these aspects, we refer to [20, 21] and the references therein. The case of rarefied gas (in which the underlying kinetic equation is the linearized Boltzmann equation instead of –here– the radiative transfer equation) was considered in [22], and the case of electronic M1 model was studied in [23, 24, 25].
The M_1 model (1), together with its generalizations M_N, are known to exhibit unphysical shocks. This is why a modified version of the M_N models have been proposed in [26], with a discontinuous Galerkin discretization.

The paper is organised as follows. In section 2, we define conical meshes and give some useful properties. Once the model is reformulated (section 3), the design of the numerical scheme is described (section 4). Some properties are given and a second order recontruction procedure is proposed. Section 5 is devoted to the theoretical analysis of the scheme : we prove that the energy remains positive under a CFL condition. We also give a rigorous proof of the AP property of the scheme (we mention here that the previous proofs were formal and used a standard Hilbert expansion). Eventually, some numerical tests are presented in section 6. The present scheme is compared to the polygonal scheme and its properties (positivity, convergence) are illustrated by numerical simulations.

In order to make the algebra clearer, vectors are denoted in **bold** in the rest of the paper.

2 Conical meshes

In this section, we define conical meshes. We follow the presentation of [27] [28]. Important geometrical properties are presented and a quadrature formula is given in order to compute the integral of a smooth function on a conical cell.

2.1 Rational quadratic Bezier curve

A rational quadratic Bezier curve is a curve $\{\mathbf{M}^\omega(q),\ q \in [0,1]\}$ such that :

$$\mathbf{M}^\omega(q) = \frac{(1-q)^2\mathbf{M}_0 + 2\omega q(1-q)\mathbf{M}_1 + q^2\mathbf{M}_2}{(1-q)^2 + 2\omega q(1-q) + q^2}, \tag{4}$$

where $\mathbf{M}_0$ and $\mathbf{M}_2$ are the extremities, $\mathbf{M}_1$ is the control point and $\omega \geq 0$ a scalar weight. The curve is said to be :

- degenerate if $\omega = 0$,
- elliptic if $\omega \in]0,1[$,
- parabolic if $\omega = 1$,
- hyperbolic if $\omega > 1$.

In the first case ($\omega = 0$), the curve is equal to the segment $[\mathbf{M}_0, \mathbf{M}_2]$. Note that the control point $\mathbf{M}_1$ almost never belongs to the curve (except in some very particular cases). This is the main drawback of this parametrisation. This is why we prefer to parametrise the curve in a different way, using a point that lies on the curve. The point is named the *shoulder point* and is defined by :

$$\mathbf{S} = \mathbf{M}^\omega(0.5) = \frac{1}{2}(\mathbf{Q}_0 + \mathbf{Q}_2), \qquad \mathbf{Q}_0 = \frac{1}{1+\omega}(\omega\mathbf{M}_1 + \mathbf{M}_0), \qquad \mathbf{Q}_2 = \frac{1}{1+\omega}(\omega\mathbf{M}_1 + \mathbf{M}_2).$$

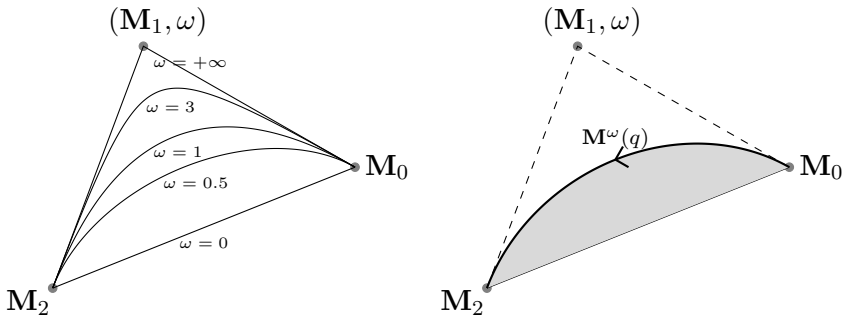

Figure 2: Rational quadratic Bezier curve.

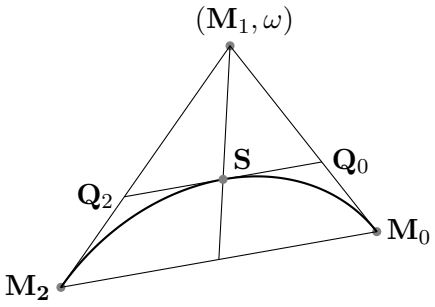

Figure 3: Shoulder point.

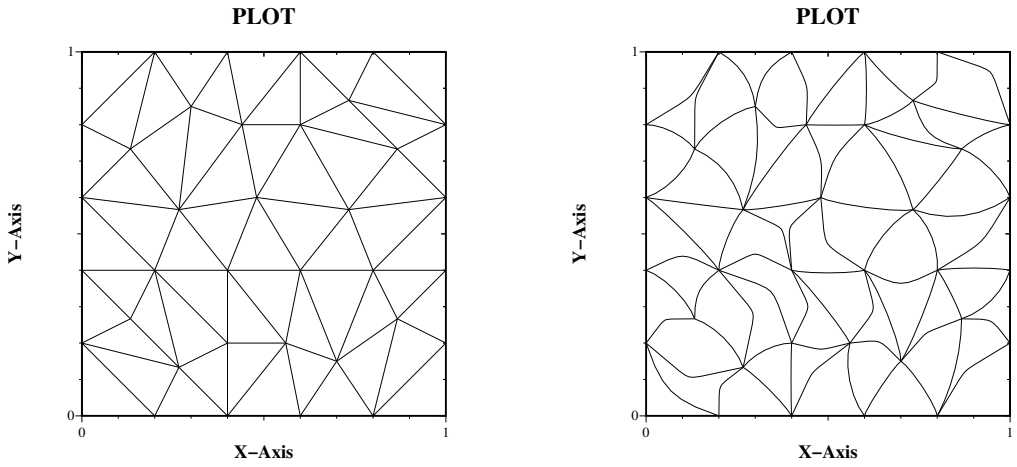

Figure 4: Polygonal (left) and conical (right) meshes.

2.2 Computing the area of a conical cell

Let Ω_j be a conical cell which center is denoted by $\mathbf{x}_j$. The area of Ω_j can be computed using the following formula :

$$|\Omega_j| = \int_{\Omega_j} d\mathbf{x} = \frac{1}{2} \int_{\partial\Omega_j} \langle \mathbf{x}(\mathbf{s}) - \mathbf{x}_j, \mathbf{N}(\mathbf{x}(\mathbf{s})) \rangle d\mathbf{s}, \tag{5}$$

where $\mathbf{N}(\mathbf{x}(\mathbf{s}))$ is the unit normal vector to the edge at the curvilinear coordinate $\mathbf{s} \in \partial\Omega_j$ and $d\mathbf{s}$ is the surface measure on $\partial\Omega_j$. In order to clarify the algebra, we define the following notation :

- $(\mathbf{x}_r)_r$ the coordinates of the vertices of the cell j,
- $(\mathbf{M}_{1,r+1/2})_{r+1/2}$ the coordinates of the control points of the cell j,
- $(\mathbf{x}_{r+1/2})_{r+1/2}$ the coordinates of the shoulder points of the cell j,
- $\sum_r g_j^r$: sum over all the vertices of the cell j of the quantity g (g_j^r being the evaluation of the function g on the vertex r in cell j),
- $\sum_{r+1/2} g_j^{r+1/2}$: sum over all the shoulder points of the cell j of the quantity g,
- $\sum_{\text{dof}} g_j^{\text{dof}} = \sum_r g_j^r + \sum_{r+1/2} g_j^{r+1/2}$: sum over all the degrees of freedom (dof) of the cell j of the quantity g,
- $N_{\text{dof}} = \sum_i 1$: number of cells that contains the given degree of freedom dof,
- $\sum_i g_i^{\text{dof}}$: sum, for a given degree of freedom, over all the cells that contains this degree of freedom,

- $\sum_{i\in\mathcal{V}_j} g_i$: sum, for a given cell j, over the neighboring cells (those which share a *dof* with j, note that $j \in \mathcal{V}_j$),
- $\sum_{j*} g_j$: sum over all the cells of the mesh,
- $\sum_{\text{dof}*} g^{\text{dof}}$: sum over all the degrees of freedom (nodes and shoulder points) of the mesh.

The integral (5) can be computed exactly using the coordinates of the vertices $(\mathbf{x}_r)_r$ and the control points $(\mathbf{M}_{1,r+1/2})_{r+1/2}$ (cf [28]) :

$$|\Omega_j| = \frac{1}{2}\sum_r \langle \mathbf{C}_j^r, \mathbf{x}_r - \mathbf{x}_j\rangle + \frac{1}{2}\sum_{r+1/2,\ \text{control points}} \left\langle \mathbf{C}_j^{r+1/2}, \mathbf{M}_{1,r+1/2} - \mathbf{x}_j \right\rangle, \tag{6}$$

it can also be expressed in terms of $(\mathbf{x}_r)_r$ and $(\mathbf{x}_{r+1/2})_{r+1/2}$ (cf [28]) :

$$|\Omega_j| = \frac{1}{2}\sum_r \langle \tilde{\mathbf{C}}_j^r, \mathbf{x}_r - \mathbf{x}_j\rangle + \frac{1}{2}\sum_{r+1/2,\ \text{shoulder points}} \left\langle \tilde{\mathbf{C}}_j^{r+1/2}, \mathbf{x}_{r+1/2} - \mathbf{x}_j \right\rangle = \frac{1}{2}\sum_{\text{dof}} \langle \tilde{\mathbf{C}}_j^{\text{dof}}, \mathbf{x}_{\text{dof}} - \mathbf{x}_j\rangle . \tag{7}$$

The coefficients $\mathbf{C}_j^r$, $\mathbf{C}_j^{r+1/2}$, $\tilde{\mathbf{C}}_j^r$ and $\tilde{\mathbf{C}}_j^{r+1/2}$ depend on the geometry of the cell and are given by :

$$\mathbf{C}_j^r = \frac{1}{2}\left((1-f(\omega_{r-1/2}))\mathbf{N}_{r-1,r} + (1-f(\omega_{r+1/2}))\mathbf{N}_{r,r+1} + f(\omega_{r-1/2})\mathbf{N}_{r-1/2,r} + f(\omega_{r+1/2})\mathbf{N}_{r,r+1/2}\right),$$

$$\mathbf{C}_j^{r+1/2} = \frac{f(\omega_{r+1/2})}{2}(\mathbf{N}_{r,r+1/2} + \mathbf{N}_{r+1/2,r+1}),$$

$$\tilde{\mathbf{C}}_j^r = \frac{1}{2}\left((1-h(\omega_{r-1/2}))\tilde{\mathbf{N}}_{r-1,r} + (1-h(\omega_{r+1/2}))\tilde{\mathbf{N}}_{r,r+1} + h(\omega_{r-1/2})\tilde{\mathbf{N}}_{r-1/2,r} + h(\omega_{r+1/2})\tilde{\mathbf{N}}_{r,r+1/2}\right), \tag{8}$$

and :

$$\tilde{\mathbf{C}}_j^{r+1/2} = \frac{h(\omega_{r+1/2})}{2}(\tilde{\mathbf{N}}_{r,r+1/2} + \tilde{\mathbf{N}}_{r+1/2,r+1}),$$

with :

$$h(\omega) = f(\omega)\frac{1+\omega}{\omega}, \qquad f(\omega) = \begin{cases} \dfrac{2\omega}{1-\omega^2}\left(\dfrac{1}{\sqrt{1-\omega^2}}\arctan\left(\sqrt{\dfrac{1-\omega}{1+\omega}}\right) - \dfrac{\omega}{2}\right) & \text{if } \omega \in [0,1[, \\ \dfrac{2}{3} \quad \text{if } \omega = 1, \\ \dfrac{\omega}{\omega^2-1}\left(\omega + \dfrac{1}{\sqrt{\omega^2-1}}\log\left(\omega - \sqrt{\omega^2-1}\right)\right) & \text{if } \omega > 1. \end{cases} \tag{9}$$

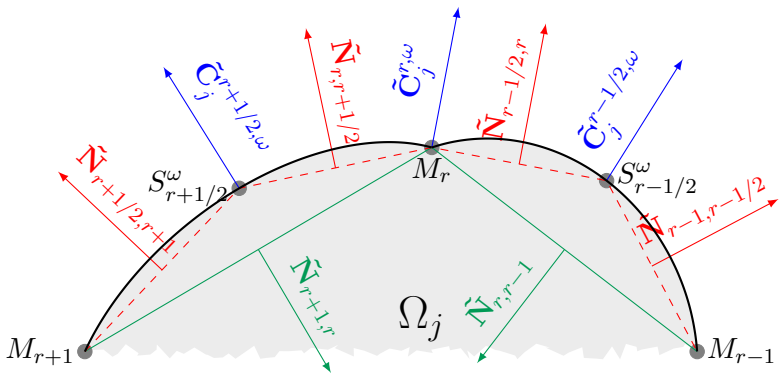

Figure 5: Normal vector from a degree of freedom defined on boundary cell Ω_j. Two types : endpoints $\mathbf{M}_r$ denoted by $\tilde{\mathbf{C}}_j^r$ or shoulder point denoted by $\tilde{\mathbf{C}}_j^{r+1/2}$

Figure 2.2 displays the curves of functions f and h (9).

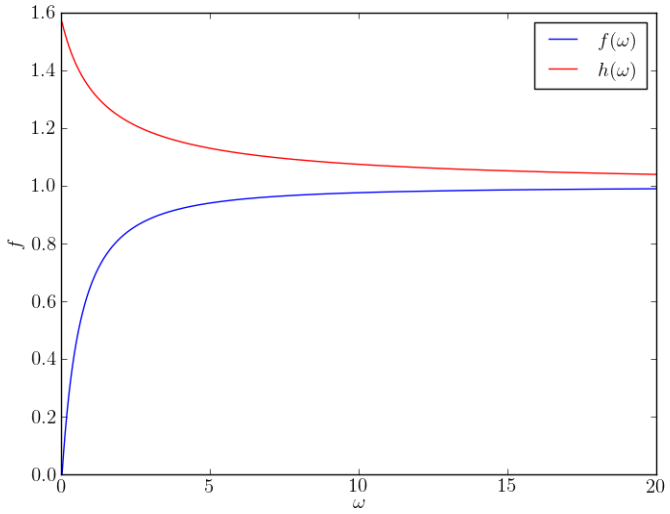

Figure 6: functions f and h.

The following properties are satisfied :

- for any cell j :

$$\sum_{\text{dof}} \tilde{\mathbf{C}}_j^{\text{dof}} = \mathbf{0},$$

- for any inner degree of freedom dof :

$$\sum_{i} \tilde{\mathbf{C}}_i^{\text{dof}} = \mathbf{0}. \tag{10}$$

As a conclusion, for any given scalar valued function g and any vector valued function $\mathbf{g}$, we approximate :

$$\begin{cases} \displaystyle\int_{\Omega_j} \nabla g = \int_{\partial\Omega_j} g\mathbf{N} \approx \sum_{\text{dof}} g_j^{\text{dof}} \tilde{\mathbf{C}}_j^{\text{dof}}, \\ \displaystyle\int_{\Omega_j} \text{div } \mathbf{g} = \int_{\partial\Omega_j} \langle \mathbf{g}, \mathbf{N} \rangle \approx \sum_{\text{dof}} \left\langle \mathbf{g}_j^{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle, \end{cases} \tag{11}$$

where g_j^{dof} and $\mathbf{g}_j^{\text{dof}}$ are the values of the functions g and $\mathbf{g}$ at the degree of freedom dof in the cell j. Formulas (11) are equalities if g and $\mathbf{g}$ are affine functions.

3 Reformulation of the model

Using the ideas from [29] and [16], system (1) can be written under the form of the gas dynamic system. This system writes :

$$\begin{cases} \partial_t \rho + \dfrac{1}{\varepsilon}\text{div}(\rho\mathbf{u}) = 0, \\ \partial_t(\rho\mathbf{u}) + \dfrac{1}{\varepsilon}\text{div}(\rho\mathbf{u}\otimes\mathbf{u}) + \dfrac{1}{\varepsilon}\nabla q = 0, \\ \partial_t(\rho e) + \dfrac{1}{\varepsilon}\text{div}(\rho e\mathbf{u}) + \dfrac{1}{\varepsilon}\text{div}(q\mathbf{u}) = 0, \\ \partial_t(\rho s) + \dfrac{1}{\varepsilon}\text{div}(\rho s\mathbf{u}) \geq 0, \end{cases} \tag{12}$$

where the density ρ, the velocity $\mathbf{u}$, the massic energy e and the massic entropy s are the unknowns of the system. The pressure q is a function of e and $\mathbf{u}$.

In order to relate the M_1 model (1) (2) to system (12), we define :

$$q := \frac{1-\chi(\mathbf{f})}{2}E, \qquad \mathbf{u} := \frac{3\chi(\mathbf{f})-1}{2\|\mathbf{f}\|^2}\mathbf{f} = \frac{2}{3-\chi(\mathbf{f})}\mathbf{f}, \qquad k := \frac{3-\chi(\mathbf{f})}{2}E.$$

The following relations hold :

$$(E+q)\mathbf{u} = \mathbf{F}, \qquad \mathbf{F} = k\mathbf{u}, \qquad P = qI_2 + \mathbf{F}\otimes\mathbf{u}.$$

Eventually system (1) can be written in the following form :

$$\begin{cases} \partial_t E + \dfrac{1}{\varepsilon}\text{div}(E\mathbf{u}) + \dfrac{1}{\varepsilon}\text{div}(q\mathbf{u}) = 0, \\ \partial_t \mathbf{F} + \dfrac{1}{\varepsilon}\text{div}(\mathbf{F}\otimes\mathbf{u}) + \dfrac{1}{\varepsilon}\nabla q = -\dfrac{\sigma}{\varepsilon^2}\mathbf{F}. \end{cases} \tag{13}$$

The structure of system (13) is quite similar to the one of system (12), for which numerical methods exist (see [30] and [17]). We use similar ideas to define a numerical solution for system (13). The major differences are that the velocity $\mathbf{u}$ and the pressure q depend on the unknowns E and $\mathbf{F}$ and that there is no density ρ.

4 Numerical method

In this section we introduce the numerical scheme : it is an adaptation of the method of [16] to conical meshes. We give here the details of its construction. We want it to satisfy the following properties :

- positivity of the radiative energy : $E > 0$,
- limitation of the flux : $\|\mathbf{f}\| \leq 1$,

- asymptotic preserving : the scheme obtained when ε vanishes is consistant with the limit diffusion equation (3).

Our numerical scheme is based on the reformulation of the M_1 model that is described in section 3. System (13) is integrated over cell Ω_j. We denote by E_j and $\mathbf{F}_j$ the averages of E and $\mathbf{F}$ over cell Ω_j, thus leading to :

$$\begin{cases} |\Omega_j|\ \partial_t E_j + \dfrac{1}{\varepsilon}\displaystyle\int_{\partial\Omega_j} E\langle \mathbf{u}, \mathbf{N}\rangle + \dfrac{1}{\varepsilon}\int_{\partial\Omega_j} q\langle \mathbf{u}, \mathbf{N}\rangle = 0, \\ |\Omega_j|\ \partial_t \mathbf{F}_j + \dfrac{1}{\varepsilon}\displaystyle\int_{\partial\Omega_j} \langle \mathbf{u}, \mathbf{N}\rangle \mathbf{F} + \dfrac{1}{\varepsilon}\int_{\partial\Omega_j} q\ \mathbf{N} = -\dfrac{\sigma}{\varepsilon^2}\displaystyle\int_{\Omega_j} \mathbf{F}. \end{cases} \tag{14}$$

First, we adapt the Gosse and Toscani method [31] to the M_1 model and we use the ideas from [16]. The source term is discretised using the node values :

$$\int_{\Omega_j} \mathbf{F} \approx \sum_{\text{dof}} \beta_j^{\text{dof}} \mathbf{F}_{\text{dof}} = \sum_{\text{dof}} \beta_j^{\text{dof}}\ k_{\text{dof}}\ \mathbf{u}_{\text{dof}}, \tag{15}$$

since $\mathbf{F} = k\mathbf{u}$. The matrix β_j^{dof} is given by :

$$\beta_j^{\text{dof}} = \tilde{\mathbf{C}}_j^{\text{dof}} \otimes (\mathbf{x}_{\text{dof}} - \mathbf{x}_j). \tag{16}$$

The advection terms ($\text{div}(E\mathbf{u})$ and $\text{div}(\mathbf{F}\otimes\mathbf{u})$) are discretised using an upwind scheme (94) :

$$\begin{cases} \displaystyle\int_{\Omega_j} \text{div}(E\mathbf{u}) = \int_{\partial\Omega_j} E\,\langle \mathbf{u}, \mathbf{N}\rangle \approx \sum_{R_j^+} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle E_j + \sum_{R_j^-} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle E_{k(\text{dof})}, \\ \displaystyle\int_{\Omega_j} \text{div}(\mathbf{F}\otimes\mathbf{u}) = \int_{\partial\Omega_j} \langle \mathbf{u}, \mathbf{N}\rangle\,\mathbf{F} \approx \sum_{R_j^+} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle \mathbf{F}_j + \sum_{R_j^-} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle \mathbf{F}_{k(\text{dof})}, \end{cases}$$

with :

$$R_j^+ = \{\text{dof},\ \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle > 0\}, \qquad R_j^- = \{\text{dof},\ \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle < 0\}.$$

The remaining integrals in (14) are approximated using formula (11) :

$$\begin{cases} \displaystyle\int_{\partial\Omega_j} q\langle \mathbf{u}, \mathbf{N}\rangle \approx \sum_{\text{dof}} q_j^{\text{dof}} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle, \\ \displaystyle\int_{\partial\Omega_j} q\ \mathbf{N} \approx \sum_{\text{dof}} q_j^{\text{dof}} \tilde{\mathbf{C}}_j^{\text{dof}}. \end{cases}$$

In the end, the scheme associated to (14) writes :

$$\begin{cases} |\Omega_j|\ \partial_t E_j + \dfrac{1}{\varepsilon}\left[\displaystyle\sum_{R_j^+} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle E_j + \sum_{R_j^-} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle E_{k(\text{dof})}\right] + \dfrac{1}{\varepsilon}\displaystyle\sum_{\text{dof}} q_j^{\text{dof}} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle = 0, \\ |\Omega_j|\ \partial_t \mathbf{F}_j + \dfrac{1}{\varepsilon}\left[\displaystyle\sum_{R_j^+} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle \mathbf{F}_j + \sum_{R_j^-} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle \mathbf{F}_{k(\text{dof})}\right] + \dfrac{1}{\varepsilon}\displaystyle\sum_{\text{dof}} q_j^{\text{dof}} \tilde{\mathbf{C}}_j^{\text{dof}} = -\dfrac{\sigma}{\varepsilon^2}\displaystyle\sum_{\text{dof}} \beta_j^{\text{dof}}\ k_{\text{dof}}\ \mathbf{u}_{\text{dof}}. \end{cases} \tag{17}$$

Note that scheme (17) is similar to the JLb scheme from [27]. The computation of $\mathbf{u}_{\text{dof}}$ and q_j^{dof} is described below. Coefficient k_{dof} is given by :

$$k_{\text{dof}} = \frac{1}{N_{\text{dof}}} \sum_i k_i, \qquad k_i = \frac{3 - \chi(\mathbf{f}_i)}{2} E_i, \tag{18}$$

where N_{dof} is the number of cells that contains the degree of freedom ($N_{\text{dof}} = 2$ for a shoulder point that is not on the boundary).

Now we apply the ideas used in [29] for developing a $1D$ solver. We impose the following relation between the pressures and the velocities at the degree of freedom and at the center of the cell :

$$q_j^{\text{dof}} + r_j^{\text{dof}} \left\langle \mathbf{u}_{\text{dof}},\ \tilde{\mathbf{n}}_j^{\text{dof}} \right\rangle = q_j + r_j^{\text{dof}} \left\langle \mathbf{u}_j,\ \tilde{\mathbf{n}}_j^{\text{dof}} \right\rangle \tag{19}$$

where :

$$\tilde{\mathbf{n}}_j^{\text{dof}} = \frac{1}{\|\tilde{\mathbf{C}}_j^{\text{dof}}\|} \tilde{\mathbf{C}}_j^{\text{dof}}, \qquad q_j = \frac{1 - \chi(\mathbf{f}_j)}{2} E_j, \qquad \mathbf{u}_j = \frac{2}{3 - \chi(\mathbf{f}_j)} \mathbf{f}_j. \tag{20}$$

Several definitions of r_j^{dof} are possible :

$$r_j^{\text{dof}} = r_j = \frac{4}{\sqrt{3}} \frac{E_j}{3 + \|\mathbf{u}_j\|^2}, \qquad \text{or : } r_j^{\text{dof}} = r_{\text{dof}} = \frac{1}{N_{\text{dof}}} \sum_i r_i. \tag{21}$$

Equation (19) is equivalent to :

$$q_j^{\text{dof}}\ \tilde{\mathbf{C}}_j^{\text{dof}} = q_j \tilde{\mathbf{C}}_j^{\text{dof}} + r_j^{\text{dof}} \alpha_j^{\text{dof}} (\mathbf{u}_j - \mathbf{u}_{\text{dof}}),$$

with :

$$\alpha_j^{\text{dof}} = \tilde{\mathbf{C}}_j^{\text{dof}} \otimes \tilde{\mathbf{n}}_j^{\text{dof}}. \tag{22}$$

The flux is modified using the Jin-Levermore [8] method. The equilibrium state obtained by setting to 0 the time derivatives and the transport terms is given by $\nabla q = -(\sigma k/\varepsilon)\mathbf{u}$. It is then added to the fluxes :

$$\left[q_j^{\text{dof}}\ + \langle (\nabla q)_{\text{dof}}\ , \mathbf{x}_j - \mathbf{x}_{\text{dof}} \rangle\right] \tilde{\mathbf{C}}_j^{\text{dof}} = q_j \tilde{\mathbf{C}}_j^{\text{dof}} + r_j^{\text{dof}} \alpha_j^{\text{dof}} (\mathbf{u}_j - \mathbf{u}_{\text{dof}}),$$

which can be written :

$$\left[q_j^{\text{dof}}\ - \frac{\sigma}{\varepsilon} k_{\text{dof}} \langle \mathbf{u}_{\text{dof}}\ , \mathbf{x}_j - \mathbf{x}_{\text{dof}} \rangle\right] \tilde{\mathbf{C}}_j^{\text{dof}} = q_j \tilde{\mathbf{C}}_j^{\text{dof}} + r_j^{\text{dof}} \alpha_j^{\text{dof}} (\mathbf{u}_j - \mathbf{u}_{\text{dof}}).$$

Thus we define, using definition (16) :

$$\mathbf{G}_j^{\text{dof}} := q_j \tilde{\mathbf{C}}_j^{\text{dof}} + r_j^{\text{dof}} \alpha_j^{\text{dof}} (\mathbf{u}_j - \mathbf{u}_{\text{dof}}) - \frac{\sigma}{\varepsilon} \beta_j^{\text{dof}}\ k_{\text{dof}}\ \mathbf{u}_{\text{dof}}. \tag{23}$$

The scheme eventually writes :

$$\begin{cases} |\Omega_j|\ \partial_t E_j + \dfrac{1}{\varepsilon} \left[\displaystyle\sum_{R_j^+} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle E_j + \sum_{R_j^-} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle E_{k(\text{dof})} \right] + \dfrac{1}{\varepsilon} \displaystyle\sum_{\text{dof}} \left\langle \mathbf{u}_{\text{dof}}, \mathbf{G}_j^{\text{dof}} \right\rangle = 0, \\ |\Omega_j|\ \partial_t \mathbf{F}_j + \dfrac{1}{\varepsilon} \left[\displaystyle\sum_{R_j^+} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle \mathbf{F}_j + \sum_{R_j^-} \left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle \mathbf{F}_{k(\text{dof})} \right] + \dfrac{1}{\varepsilon} \displaystyle\sum_{\text{dof}} \mathbf{G}_j^{\text{dof}} = -\frac{\sigma}{\varepsilon^2} \sum_{\text{dof}} \beta_j^{\text{dof}}\ k_{\text{dof}}\ \mathbf{u}_{\text{dof}}. \end{cases} \tag{24}$$

The second line of system (24) can be simplified using (23) :

$$\frac{1}{\varepsilon} \sum_{\text{dof}} \mathbf{G}_j^{\text{dof}} + \frac{\sigma}{\varepsilon^2} \sum_{\text{dof}} \beta_j^{\text{dof}}\ k_{\text{dof}}\ \mathbf{u}_{\text{dof}} = \frac{1}{\varepsilon} \underbrace{\sum_{\text{dof}} q_j \tilde{\mathbf{C}}_j^{\text{dof}}}_{=\mathbf{0}} + \frac{1}{\varepsilon} \sum_{\text{dof}} r_j^{\text{dof}} \alpha_j^{\text{dof}}\ (M_{\text{dof}} \mathbf{u}_j - \mathbf{u}_{\text{dof}}) \tag{25}$$

$$+\frac{1}{\varepsilon}\left(\sum_{\text{dof}} r_j^{\text{dof}}\alpha_j^{\text{dof}} - \sum_{\text{dof}} r_j^{\text{dof}}\ \alpha_j^{\text{dof}} M_{\text{dof}}\right)\mathbf{u}_j,$$

where M_r is defined by :

$$M_r = \left(\sum_i r_i^r\alpha_i^r + \frac{\sigma}{\varepsilon}k_r\beta_i^r\right)^{-1}\left(\sum_i r_i^r\alpha_i^r\right), \tag{26}$$

and $M_{r+1/2}$ is defined below, see (39).
The semi-discrete system writes :

$$\begin{cases} |\Omega_j|\partial_t E_j + \dfrac{1}{\varepsilon}\left[\displaystyle\sum_{R_j^+}\left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}}\right\rangle E_j + \sum_{R_j^-}\left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}}\right\rangle E_{k(\text{dof})}\right] + \dfrac{1}{\varepsilon}\displaystyle\sum_{\text{dof}}\left\langle \mathbf{u}_{\text{dof}}, \mathbf{G}_j^{\text{dof}}\right\rangle = 0, \\ |\Omega_j|\partial_t \mathbf{F}_j + \dfrac{1}{\varepsilon}\left[\displaystyle\sum_{R_j^+}\left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}}\right\rangle \mathbf{F}_j + \sum_{R_j^-}\left\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}}\right\rangle \mathbf{F}_{k(\text{dof})}\right] + \dfrac{1}{\varepsilon}\displaystyle\sum_{\text{dof}}\mathbf{G}_j^{\text{dof}\ *} \\ = -\dfrac{1}{\varepsilon}\dfrac{1}{k_j}\left(\displaystyle\sum_{\text{dof}} r_j^{\text{dof}}\ \alpha_j^{\text{dof}}(I - M_{\text{dof}})\right)\mathbf{F}_j, \end{cases} \tag{27}$$

with :

$$\mathbf{G}_j^{\text{dof}\ *} = r_j^{\text{dof}}\ \alpha_j^{\text{dof}}(M_{\text{dof}}\mathbf{u}_j - \mathbf{u}_{\text{dof}}). \tag{28}$$

We impose, for every inner degree of freedom dof :

$$\sum_i \mathbf{G}_i^r = \sum_i \mathbf{G}_i^{r+1/2} = \mathbf{0}. \tag{29}$$

Equation (29) ensures the conservativity of the scheme. It allows also to compute $\mathbf{u}_{\text{dof}}$ for some given $(q_i)_i$, $(r_i)_i$ and $(\mathbf{u}_i)_i$. Using definition (23), it writes :

$$\left(\sum_i r_i^{\text{dof}}\alpha_i^{\text{dof}} + \frac{\sigma}{\varepsilon}k_{\text{dof}}\beta_i^{\text{dof}}\right)\mathbf{u}_{\text{dof}} = \sum_i q_i\tilde{\mathbf{C}}_i^{\text{dof}} + r_i^{\text{dof}}\alpha_i^{\text{dof}}\mathbf{u}_i. \tag{30}$$

We have, for any inner node r :

$$\left(\sum_i r_i^r\alpha_i^r + \frac{\sigma}{\varepsilon}k_r\beta_i^r\right)\mathbf{u}_r = \sum_i q_i\tilde{\mathbf{C}}_i^r + r_i^r\alpha_i^r\mathbf{u}_i. \tag{31}$$

The proof of the invertibility of the matrix $\sum_i r_i\alpha_i^r + \frac{\sigma}{\varepsilon}k_r\beta_i^r$ is given below. However, for an inner shoulder point, the matrix is not invertible. Denoting by i the unique cell different from j that contains the shoulder point $r+1/2$, we have :

$$\tilde{\mathbf{C}}_j^{r+1/2} + \tilde{\mathbf{C}}_i^{r+1/2} = \mathbf{0},$$

and the matrix of equation (30) simplifies into :

$$r_i^{r+1/2}\alpha_i^{r+1/2} + r_j^{r+1/2}\alpha_j^{r+1/2} + \frac{\sigma}{\varepsilon}k_{r+1/2}(\beta_i^{r+1/2} + \beta_j^{r+1/2}) = \tilde{\mathbf{C}}_j^{r+1/2}\otimes\mathbf{D}_{r+1/2}, \tag{32}$$

with :

$$\mathbf{D}_{r+1/2} = (r_j^{r+1/2} + r_i^{r+1/2})\tilde{\mathbf{n}}_j^{r+1/2} + \frac{\sigma}{\varepsilon}k_{r+1/2}(\mathbf{x}_i - \mathbf{x}_j). \tag{33}$$

Therefore, the matrix of (32) has rank 1. The right hand side of (30) can be written under the form $b_{r+1/2}\tilde{\mathbf{C}}_j^{r+1/2}$, with :

$$b_{r+1/2} = q_j - q_i + \left\langle \tilde{\mathbf{n}}_j^{r+1/2}, r_j^{r+1/2}\mathbf{u}_j + r_i^{r+1/2}\mathbf{u}_i \right\rangle. \tag{34}$$

Using ideas from [27], $\mathbf{u}_{r+1/2}$ is computed using the following formulas :

$$\langle \mathbf{u}_{r+1/2}, \mathbf{D}_{r+1/2} \rangle = b_{r+1/2}, \qquad \langle \mathbf{u}_{r+1/2}, (\mathbf{D}_{r+1/2})^{\perp} \rangle = \frac{1}{2} \left\langle \mathbf{u}_r + \mathbf{u}_{r+1}, (\mathbf{D}_{r+1/2})^{\perp} \right\rangle, \tag{35}$$

where, for any $\xi \in \mathbb{R}^2$:

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \qquad \xi^{\perp} = \begin{pmatrix} -\xi_2 \\ \xi_1 \end{pmatrix}, \qquad (\xi \quad \xi^{\perp}) = \begin{pmatrix} \xi_1 & -\xi_2 \\ \xi_2 & \xi_1 \end{pmatrix}. \tag{36}$$

The linear system solved by $\mathbf{u}_{r+1/2}$ can be written under the form :

$$\left(A_{r+1/2} + \frac{\sigma}{\varepsilon} B_{r+1/2} \right) \mathbf{u}_{r+1/2} = \mathbf{y}_{r+1/2},$$

with :

$$A_{r+1/2} = (r_j^{r+1/2} + r_i^{r+1/2}) \begin{pmatrix} \tilde{\mathbf{n}}_j^{r+1/2} & \left(\tilde{\mathbf{n}}_j^{r+1/2}\right)^{\perp} \end{pmatrix}^T, \qquad B_{r+1/2} = k_{r+1/2} \begin{pmatrix} \mathbf{x}_i - \mathbf{x}_j & (\mathbf{x}_i - \mathbf{x}_j)^{\perp} \end{pmatrix}^T, \tag{37}$$

$$\mathbf{y}_{r+1/2} = \begin{pmatrix} b_{r+1/2} \\ \frac{1}{2}\langle \mathbf{u}_r + \mathbf{u}_{r+1}, \mathbf{D}_{r+1/2}^{\perp} \rangle \end{pmatrix}, \tag{38}$$

thus allowing us to define $M_{r+1/2}$ by :

$$M_{r+1/2} = \left(A_{r+1/2} + \frac{\sigma}{\varepsilon} B_{r+1/2} \right)^{-1} A_{r+1/2}. \tag{39}$$

Remark 1. *As noted in [27] and [28], if the mesh is conical degenerate (all the weights ω are set to 0), the scheme under study is not equal to the classical polygonal scheme. Indeed, the contribution of the shoulder points is nonzero and the two schemes may produce quite different results. This remark is highlighted by the test cases below. Indeed, a parasite mode is observed with the classical polygonal node solver, but it is no longer present with the conical degenerate scheme.*

Remark 2. *The definition of M_{dof} in (26) and (39) is arbitrary. Indeed, equation (25) is satisfied for any choice of the matrix M_{dof}. However, as mentioned in [32], this particular choice (26) allows to recover the Gosse-Toscani scheme [31] in the 1D case.*

4.1 Partially implicit time discretisation

Only the source term in (27) is chosen implicit. The obtained scheme is thus easy to implement and its stability condition is not too restrictive.

$$\begin{cases} |\Omega_j| \dfrac{E_j^{n+1} - E_j^n}{\Delta t} + \dfrac{1}{\varepsilon} \left[\displaystyle\sum_{R_j^+} \left\langle \mathbf{u}_{\text{dof}}^n, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle E_j^n + \sum_{R_j^-} \left\langle \mathbf{u}_{\text{dof}}^n, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle E_{k(\text{dof})}^n \right] + \dfrac{1}{\varepsilon} \displaystyle\sum_{\text{dof}} \left\langle \mathbf{u}_{\text{dof}}^n, \left(\mathbf{G}_j^{\text{dof}}\right)^n \right\rangle = 0, \\ |\Omega_j| \dfrac{\mathbf{F}_j^{n+1} - \mathbf{F}_j^n}{\Delta t} + \dfrac{1}{\varepsilon} \left[\displaystyle\sum_{R_j^+} \left\langle \mathbf{u}_{\text{dof}}^n, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle \mathbf{F}_j^n + \sum_{R_j^-} \left\langle \mathbf{u}_{\text{dof}}^n, \tilde{\mathbf{C}}_j^{\text{dof}} \right\rangle \mathbf{F}_{k(\text{dof})}^n \right] + \dfrac{1}{\varepsilon} \displaystyle\sum_{\text{dof}} \left(\mathbf{G}_j^{\text{dof}\,*}\right)^n \\ = -\dfrac{1}{\varepsilon} \left(\displaystyle\sum_{\text{dof}} \left(r_j^{\text{dof}}\right)^n \ \alpha_j^{\text{dof}} (I - M_{\text{dof}}^n) \right) \dfrac{1}{k_j^n} \mathbf{F}_j^{n+1}. \end{cases} \tag{40}$$

The polygonal scheme depends only on the values located at the nodes of the mesh (cf [16]) :

$$\begin{cases} |\Omega_j|\dfrac{E_j^{n+1}-E_j^n}{\Delta t}+\dfrac{1}{\varepsilon}\left[\displaystyle\sum_{r,l_{jr}\langle \mathbf{u}_r^n,\mathbf{n}_{jr}\rangle>0} l_{jr}\left\langle \mathbf{u}_r^n,\mathbf{n}_{jr}\right\rangle E_j^n+\sum_{r,l_{jr}\langle \mathbf{u}_r^n,\mathbf{n}_{jr}\rangle<0} l_{jr}\left\langle \mathbf{u}_r^n,\mathbf{n}_{jr}\right\rangle E_{k(r)}^n+\sum_r \left\langle \mathbf{u}_r^n,\left(\mathbf{G}_j^r\right)^n\right\rangle\right]=0, \\ |\Omega_j|\dfrac{\mathbf{F}_j^{n+1}-\mathbf{F}_j^n}{\Delta t}+\dfrac{1}{\varepsilon}\left[\displaystyle\sum_{r,l_{jr}\langle \mathbf{u}_r^n,\mathbf{n}_{jr}\rangle>0} l_{jr}\left\langle \mathbf{u}_r^n,\mathbf{n}_{jr}\right\rangle \mathbf{F}_j^n+\sum_{r,l_{jr}\langle \mathbf{u}_r^n,\mathbf{n}_{jr}\rangle<0} l_{jr}\left\langle \mathbf{u}_r^n,\mathbf{n}_{jr}\right\rangle \mathbf{F}_{k(r)}^n+\sum_r \left(\mathbf{G}_j^{r\ *}\right)^n\right] \\ =-\dfrac{1}{\varepsilon}\left(\displaystyle\sum_r \left(r_j^r\right)^n \hat{\alpha}_j^r(I_d-M_r^n)\right)\dfrac{1}{k_j^n}\mathbf{F}_j^{n+1}, \end{cases} \tag{41}$$

with :

$$\mathbf{n}_{jr}=\frac{1}{2l_{jr}}\begin{pmatrix} y_{r+1}-y_{r-1} \\ x_{r-1}-x_{r+1}\end{pmatrix}, \qquad l_{jr}=\frac{1}{2}||\mathbf{x}_{r+1}-\mathbf{x}_{r-1}||, \qquad \hat{\alpha}_j^r=l_{jr}\mathbf{n}_{jr}\otimes\mathbf{n}_{jr}.$$

4.2 Second order reconstruction

In this section, we explain how to modify the scheme (40) so as to make it second order in space while ensuring that the reconstructed values of the energy are positive and that the reconstructed flux is limited. To this aim, we adapt the ideas of [28]. Only the advection terms are modified (that is to say $\mathrm{div}(E\mathbf{u})$ and $\mathrm{div}(\mathbf{F}\otimes\mathbf{u})$). Writing them as functions of E and $\mathbf{f}$, system (13) becomes :

$$\begin{cases} \partial_t E+\dfrac{1}{\varepsilon}\mathrm{div}(E\mathbf{u})+\dfrac{1}{\varepsilon}\mathrm{div}(q\mathbf{u})=0, \\ \partial_t(E\mathbf{f})+\dfrac{1}{\varepsilon}\mathrm{div}(E\mathbf{f}\otimes\mathbf{u})+\dfrac{1}{\varepsilon}\nabla q=-\dfrac{\sigma}{\varepsilon^2}E\mathbf{f}. \end{cases}$$

For the sake of clarity, we will consider the following linear transport equations :

$$\begin{cases} \partial_t E+\mathrm{div}(E\mathbf{a})=0, \\ \partial_t(E\mathbf{f})+\mathrm{div}(E\mathbf{f}\otimes\mathbf{a})=0, \end{cases}$$

where the velocity $\mathbf{a}$ is given. A first order scheme (such as the upwind scheme (94)) approximates the solution with some piecewise constant function. Here the approximation is built using piecewise affine functions :

$$\begin{cases} P_j^1(\mathbf{x},E)=E_j+\langle(\nabla E)_j,\mathbf{x}-\mathbf{x}_j\rangle, \\ P_j^1(\mathbf{x},\mathbf{F})=\mathbf{F}_j+(\nabla\mathbf{F})_j\cdot(\mathbf{x}-\mathbf{x}_j). \end{cases}$$

The exponent 1 means that the polynomial has degree 1. Then the gradient of E is limited so as to have : $P_j^1(\mathbf{x},E)\geq 0$, thus we write :

$$P_j^1(\mathbf{x},E)=E_j+\langle(\tilde{\nabla E})_j,\mathbf{x}-\mathbf{x}_j\rangle=E_j+\alpha_{j,E}\langle(\nabla E)_j,\mathbf{x}-\mathbf{x}_j\rangle,$$

where $\alpha_{j,E}$ is the limiter (see [33]). Therefore we have $P^1(\mathbf{x},E)\geq 0$. Moreover, using the Leibniz formula, one can write :

$$P_j^1(\mathbf{x},\mathbf{F})=\mathbf{F}_j+(\nabla\mathbf{F})_j\cdot(\mathbf{x}-\mathbf{x}_j)=E_j\mathbf{f}_j+E_j(\nabla\mathbf{f})_j\cdot(\mathbf{x}-\mathbf{x}_j)+\langle\nabla E_j,\mathbf{x}-\mathbf{x}_j\rangle\mathbf{f}_j.$$

The dimensionless flux $\mathbf{f}$ is thus approximated by :

$$R_j^1(\mathbf{x},\mathbf{f})=\frac{P^1(\mathbf{x},\mathbf{F})}{P^1(\mathbf{x},E)}=\mathbf{f}_j+\frac{E_j}{P^1(\mathbf{x},E)}(\nabla\mathbf{f})_j\cdot(\mathbf{x}-\mathbf{x}_j),$$

the letter R reminds that $\mathbf{f}$ is approximated by some rational fraction. The jacobian of $\mathbf{f}$ is then limited so as to have $\left\|R_j^1(\mathbf{x},\mathbf{f})\right\|\leq 1$:

$$R_j^1(\mathbf{x},\mathbf{f}) = \mathbf{f}_j + \frac{E_j}{P^1(\mathbf{x},E)}\alpha_{j,\mathbf{f}}(\nabla\mathbf{f})_j\cdot(\mathbf{x}-\mathbf{x}_j),$$

where $\alpha_{j,\mathbf{f}}$ may be a matrix or scalar limiter (its computation is based on convex hulls, cf [34]). Finally, $\mathbf{F}$ is approximated by :

$$P_j^1(\mathbf{x},\mathbf{F}) = P_j^1(\mathbf{x},E)R_j^1(\mathbf{x},\mathbf{f}).$$

Therefore, values of the unknowns at the degrees of freedom can be computed :

$$E_j^{\text{dof}} = P_j^1(\mathbf{x}_{\text{dof}},E), \qquad \mathbf{F}_j^{\text{dof}} = P_j^1(\mathbf{x}_{\text{dof}},\mathbf{F}).$$

These values are then used instead of the cell-centered values :

$$|\Omega_j|\partial_t E_j + \sum_{R_j^+}\langle\mathbf{a}_{\text{dof}},\tilde{\mathbf{C}}_j^{\text{dof}}\rangle E_j^{\text{dof}} + \sum_{R_j^-}\langle\mathbf{a}_{\text{dof}},\tilde{\mathbf{C}}_j^{\text{dof}}\rangle E_{k(\text{dof})} = 0,$$

with :

$$E_{k(\text{dof})} = \frac{1}{\sum_{I_{\text{dof}}^+}\langle\mathbf{a}_{\text{dof}},\tilde{\mathbf{C}}_i^{\text{dof}}\rangle}\sum_{I_{\text{dof}}^+}\langle\mathbf{a}_{\text{dof}},\tilde{\mathbf{C}}_i^{\text{dof}}\rangle E_i^{\text{dof}},$$

and the same values are computed for $\mathbf{F}$.

Remark 3. *Another method would consist in directly limiting the jacobian of* $\mathbf{F}$ *so as to verify* $\left\|P_j^1(\mathbf{x},\mathbf{F})\right\| \leq P_j^1(\mathbf{x},E)$*. However the first method gives much better results in practice. Indeed, the limiter of the vector quantity (*$\mathbf{F}$ *or* $\mathbf{f}$*) is based on the computation of the convex hulls of the values in the neighboring cells. Therefore it is easier to ensure* $\|\mathbf{f}\| \leq 1$ *(since the bound is constant) than* $\|\mathbf{F}\| \leq E$ *(since the value of* E *in one cell does not necessarily lie in the convex hull of the values of the neighboring cells).*

4.3 Diffusion limit scheme

In this section we describe the limit scheme, that is to say the scheme that is obtained as ε goes to 0 in (40). It reads as :

$$|\Omega_j|\frac{E_j^{n+1}-E_j^n}{\Delta t} + \left[\sum_{R_j^+}\left\langle\mathbf{u}_{\text{dof}}^n,\tilde{\mathbf{C}}_j^{\text{dof}}\right\rangle E_j^n + \sum_{R_j^-}\left\langle\mathbf{u}_{\text{dof}}^n,\tilde{\mathbf{C}}_j^{\text{dof}}\right\rangle E_{k(\text{dof})}^n\right] + \sum_{\text{dof}}\left\langle\frac{E_j^n}{3}\tilde{\mathbf{C}}_j^{\text{dof}} - \frac{4\sigma}{3}E_{\text{dof}}^n\beta_j^{\text{dof}}\mathbf{u}_{\text{dof}}^n,\mathbf{u}_{\text{dof}}^n\right\rangle = 0, \tag{42}$$

with $E_{\text{dof}}^n = \sum_i E_i^n/N_{\text{dof}}$ and :

$$\left(\sum_i\beta_i^r\right)\ \mathbf{u}_r^n = \frac{1}{4\sigma E_r^n}\sum_i E_i^n\tilde{\mathbf{C}}_i^r, \qquad \begin{cases}\langle\mathbf{u}_{r+1/2}^n,(\mathbf{x}_i-\mathbf{x}_j)\rangle = \dfrac{E_j^n-E_i^n}{4\sigma E_{r+1/2}^n},\\ \langle\mathbf{u}_{r+1/2}^n,(\mathbf{x}_i-\mathbf{x}_j)^\perp\rangle = \dfrac{1}{2}\langle\mathbf{u}_r^n+\mathbf{u}_{r+1}^n,(\mathbf{x}_i-\mathbf{x}_j)^\perp\rangle.\end{cases}$$

It is an extension to conical meshes of the limit scheme of [16] :

$$\begin{cases}|\Omega_j|\dfrac{E_j^{n+1}-E_j^n}{\Delta t} + \displaystyle\sum_{r,l_{jr}\langle\mathbf{u}_r^n,\mathbf{n}_{jr}\rangle>0} l_{jr}\langle\mathbf{u}_r^n,\mathbf{n}_{jr}\rangle E_j^n + \sum_{r,l_{jr}\langle\mathbf{u}_r^n,\mathbf{n}_{jr}\rangle<0} l_{jr}\langle\mathbf{u}_r^n,\mathbf{n}_{jr}\rangle E_{k(r)}^n\\ \quad + \dfrac{E_j^n}{3}\displaystyle\sum_r l_{jr}\langle\mathbf{u}_r^n,\mathbf{n}_{jr}\rangle - \frac{4\sigma}{3}\sum_r E_r^n\langle\beta_j^r\mathbf{u}_r^n,\mathbf{u}_r^n\rangle = 0,\\ \beta_r\ \mathbf{u}_r^n = \dfrac{1}{4\sigma E_r^n}\displaystyle\sum_i E_i^n l_{ir}\mathbf{n}_{ir}.\end{cases} \tag{43}$$

In addition, it is proved in [16] that the scheme (43) is consistent with the diffusion equation (3). In the next section, we give a rigorous proof of the convergence of the scheme (40) toward the limit scheme (42) as ε vanishes. In the previous works (such as [16]) this AP property was formally proved using a standard Hilbert expansion.

5 Theoretical study of the scheme

5.1 Notation

We denote by J the number of cells of the mesh. The numerical solution at iteration n is denoted by $E^{n,\varepsilon} = (E_j^{n,\varepsilon})_{j<J}$ and $\mathbf{F}^{n,\varepsilon} = (\mathbf{F}_j^{n,\varepsilon})_{j<J}$. The scheme (40) can be written as :

$$\left(E^{n+1,\varepsilon}, \mathbf{F}^{n+1,\varepsilon}\right) = z(\varepsilon, E^{n,\varepsilon}, \mathbf{F}^{n,\varepsilon}),$$

where $z :]0,1] \times \mathfrak{D} \to \mathbb{R}^J \times \left(\mathbb{R}^2\right)^J$ is a function defined by :

$$\left(E^{(1)}, \mathbf{F}^{(1)}\right) = z(\varepsilon, E, \mathbf{F}) = (z_1(\varepsilon, E, \mathbf{F}), z_2(\varepsilon, E, \mathbf{F})), \qquad \mathfrak{D} = \{(E, \mathbf{F}) \in \mathbb{R}^J \times \left(\mathbb{R}^2\right)^J, \ \forall j < J, \ E_j > 0, \quad \|\mathbf{F}_j\| \leq E_j\}$$

and :

$$E_j^{(1)} = [z_1(\varepsilon, E, \mathbf{F})]_j = E_j - \frac{\Delta t}{|\Omega_j|}\left[\underbrace{\sum_{R_j^+} \frac{1}{\varepsilon}\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}}\rangle E_j + \sum_{R_j^-} \frac{1}{\varepsilon}\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}}\rangle E_{k(\text{dof})}}_{\text{advection terms}} + \frac{1}{\varepsilon}\sum_{\text{dof}}\langle \mathbf{u}_{\text{dof}}, \mathbf{G}_j^{\text{dof}}\rangle\right], \tag{44}$$

$$\tilde{A}_j \mathbf{F}_j^{(1)} = \tilde{A}_j [z_2(\varepsilon, E, \mathbf{F}]_j = \mathbf{Y}_j, \tag{45}$$

with :

$$\mathbf{Y}_j = \mathbf{F}_j - \frac{\Delta t}{|\Omega_j|}\left[\underbrace{\sum_{R_j^+} \frac{1}{\varepsilon}\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}}\rangle \mathbf{F}_j + \sum_{R_j^-} \frac{1}{\varepsilon}\langle \mathbf{u}_{\text{dof}}, \tilde{\mathbf{C}}_j^{\text{dof}}\rangle \mathbf{F}_{k(\text{dof})}}_{\text{advection terms}} + \frac{1}{\varepsilon}\sum_{\text{dof}} \mathbf{G}_j^{\text{dof}\ *}\right], \tag{46}$$

and :

$$\tilde{A}_j = I + \frac{\Delta t}{\varepsilon |\Omega_j| k_j} \sum_{\text{dof}} r_j^{\text{dof}} \alpha_j^{\text{dof}} (I - M_{\text{dof}}). \tag{47}$$

Moreover, equation (30) writes :

$$\left(A_{\text{dof}} + \frac{\sigma}{\varepsilon} B_{\text{dof}}\right) \mathbf{u}_{\text{dof}} = \mathbf{y}_{\text{dof}}, \qquad \frac{1}{\varepsilon}\mathbf{u}_{\text{dof}} = \left(\varepsilon A_{\text{dof}} + \sigma B_{\text{dof}}\right)^{-1} \mathbf{y}_{\text{dof}}, \tag{48}$$

with, for a given vertex r :

$$A_r = \sum_i r_i^r \alpha_i^r, \qquad B_r = k_r \beta_r, \qquad \beta_r = \sum_i \beta_i^r, \qquad \mathbf{y}_r = \sum_i q_i \tilde{\mathbf{C}}_i^r + r_i^r \alpha_i^r \mathbf{u}_i, \tag{49}$$

and for a given shoulder point $r + 1/2$:

$$A_{r+1/2} = (r_j^{r+1/2} + r_i^{r+1/2}) \begin{pmatrix} \tilde{\mathbf{n}}_j^{r+1/2} & \left(\tilde{\mathbf{n}}_j^{r+1/2}\right)^\perp \end{pmatrix}^T, \qquad B_{r+1/2} = k_{r+1/2} \begin{pmatrix} \mathbf{x}_i - \mathbf{x}_j & (\mathbf{x}_i - \mathbf{x}_j)^\perp \end{pmatrix}^T,$$

$$\mathbf{y}_{r+1/2} = \begin{pmatrix} b_{r+1/2} \\ \frac{1}{2}\langle \mathbf{u}_r + \mathbf{u}_{r+1}, \mathbf{D}_{r+1/2}^\perp \rangle \end{pmatrix},$$

where β_i^r, α_i^r, q_i, r_i^{dof}, $\mathbf{u}_i$, k_{dof}, $b_{r+1/2}$ and $\mathbf{D}_{r+1/2}$ are respectively defined in (16), (22), (20), (18), (34) and (33).

The coefficients $\mathbf{u}_{\text{dof}}$, $\mathbf{G}_j^{\text{dof}}$, $\mathbf{G}_j^{\text{dof}\,*}$ and M_{dof} depend on ε, E and $\mathbf{F}$. For a given $(E, \mathbf{F}) \in \mathfrak{D}$, we denote by $E_{\min}$ and $E_{\max}$ the extrema of E.

Eventually, we assume periodic boundary conditions are imposed.

Lemma 5.1. *Let $(E, \mathbf{F}) \in \mathfrak{D}$ and $\sigma \geq 0$ and $\varepsilon > 0$, then the following properties are satisfied :*

$$\sum_{j*} \sum_{dof} \left\langle \mathbf{u}_{dof}, \mathbf{G}_j^{dof} \right\rangle = \sum_{dof*} \sum_{i} \left\langle \mathbf{u}_{dof}, \mathbf{G}_i^{dof} \right\rangle = 0.$$

Besides, the advection part of the scheme being conservative (cf section 8.1), the scheme is conservative :

$$\sum_{j*} |\Omega_j| E_j^{(1)} = \sum_{j*} |\Omega_j| E_j.$$

Here we choose : $r_j^{\text{dof}} = r_{\text{dof}}$ and thus, for any vertex r :

$$A_r = r_r \alpha_r, \qquad \alpha_r = \sum_i \alpha_i^r.$$

This choice allows us to prove the following results.

Lemma 5.2. *There exists a universal positive constant denoted by $C_{5.2}$ such that, for any $(E, \mathbf{F}) \in \mathfrak{D}$ and any $\varepsilon > 0$, the following inequalities hold true :*

$$E_j - C_{5.2} \frac{\Delta t}{\sigma h^2 + \varepsilon h} \sum_{i \in \mathcal{V}_j} E_i \leq E_j^{(1)} \leq E_j + C_{5.2} \frac{\Delta t}{\sigma h^2 + \varepsilon h} \sum_{i \in \mathcal{V}_j} E_i,$$

Remark 4. *A sufficient condition condition for ensuring the positivity of the energy therfore write :*

$$\Delta t < \frac{1}{C_{5.2}} (\sigma h^2 + \varepsilon h) \min_j \left\{ \frac{E_j}{\sum_{i \in \mathcal{V}_j} E_i} \right\}.$$

Lemma 5.3. *There exists a universal positive constant denoted by $C_{5.3}$ such that, for any $(E, \mathbf{F}) \in \mathfrak{D}$ and any $\varepsilon > 0$, the following inequality holds true :*

$$\|\mathbf{Y}_j\| \leq E_j + C_{5.3} \frac{\Delta t}{\sigma h^2 + \varepsilon h} \sum_{i \in \mathcal{V}_j} E_i$$

The following lemma proves that the scheme is well defined.

Lemma 5.4. *There exists a constant $C_{5.4} > 0$ such that, for any $\varepsilon > 0$, $\sigma \geq 0$ and $(E, \mathbf{F}) \in \mathfrak{D}$, if Δt satisfies :*

$$\Delta t < C_{5.4}\, \varepsilon \frac{\sigma h + \varepsilon}{\sigma} \min_j \left\{ \frac{E_j}{\sum_{i \in \mathcal{V}_j} E_i} \right\},$$

then the matrix $\tilde{A}_j$ in (47) is invertible.

The following lemma is more interesting since it proves the invertibility of the matrix in the diffusion regime ($\varepsilon \ll 1$) for a given timestep Δt.

Lemma 5.5. *There exists a universal positive constant $C_{5.5}$ such that, for any $(E, \mathbf{F}) \in \mathfrak{D}$, any $\sigma \geq 0$ and for $\varepsilon > 0$ such that :*

$$\varepsilon C_{5.5} \frac{h}{\Delta t} \left(1 + \frac{\Delta t}{\sigma h^2 + \varepsilon h} \frac{1}{E_j} \sum_{i \in \mathcal{V}_j} E_i \right) < 1,$$

the matrix $\tilde{A}_j$ in (45) is invertible and the following inequalities hold true :

$$\left\| \mathbf{F}_j^{(1)} \right\| \leq C_{5.5} \frac{\varepsilon h}{\Delta t} \left(E_j + \frac{\Delta t}{\sigma h^2 + \varepsilon h} \sum_{i \in \mathcal{V}_j} E_i \right), \qquad \left\| \mathbf{f}_j^{(1)} \right\| \leq C_{5.5} \frac{\varepsilon h}{\Delta t} \frac{E_j + \frac{\Delta t}{\sigma h^2 + \varepsilon h} \sum_{i \in \mathcal{V}_j} E_i}{E_j - \frac{\Delta t}{\sigma h^2 + \varepsilon h} \sum_{i \in \mathcal{V}_j} E_i}.$$

Lemma 5.6. *For $\delta \geq 0$ we define :*

$$\mathfrak{D}_\delta = \{ (E, \mathbf{F}) \in \mathbb{R}^J \times \left(\mathbb{R}^2 \right)^J, \ \forall j < J, \ E_j > \delta, \quad \|\mathbf{F}_j\| \leq E_j \}.$$

For any $\delta > 0$, there exists $\varepsilon_{\min}^\delta > 0$ such that z is continuous on $[0, \varepsilon_{\min}^\delta[\times \mathfrak{D}_\delta$.

Theorem 5.7. *Let $N \in \mathbb{N}$ be the number of iterations and $(E^{n,\varepsilon}, \mathbf{F}^{n,\varepsilon})_{n<N} \in \mathfrak{D}^N$ be the solution to (40) for a given $\varepsilon > 0$. Assume that, for $\varepsilon > 0$ small enough :*

$$\exists \delta > 0, \ \forall n < N, \ (E^{n,\varepsilon}, \mathbf{F}^{n,\varepsilon}) \in \mathfrak{D}_\delta, \tag{50}$$

and that the initial conditions is of the form :

$$E^{0,\varepsilon} = E(t = 0) \ \textit{(independent of } \varepsilon), \qquad \mathbf{F}^{0,\varepsilon} = \mathbf{0}.$$

We define the limit scheme $(E^{n,0})_{n<N}$ by :

$$\forall n < N, \ E^{n+1,0} = z_1(0, E^{n,\varepsilon}, \mathbf{0}).$$

Then, for any $n < N$:

$$\mathbf{F}^{n,\varepsilon} \xrightarrow[\varepsilon \to 0]{} \mathbf{0}, \qquad E^{n,\varepsilon} \xrightarrow[\varepsilon \to 0]{} E^{n,0}$$

and $(E^{n,0})_{n<N}$ is the solution to (42).

Remark 5. *We do not know* a priori *if assumption (50) is satisfied in general. However, if we assume that :*

$$\forall n < N, \ \Delta t \leq \frac{1}{2} C_{5.2}^2 (\sigma h^2 + \varepsilon h) \frac{E_{\min}^n}{E_{\max}^n}, \tag{51}$$

and that $C_{5.2}$ is chosen large enough so as to ensure : $\sum_{i \in \mathcal{V}_j} E_i \leq C_{5.2} E_{\max}$ for any cell j and any E, then (50) holds true. Note however that this assumption is very restrictive (see below). Indeed, according to lemma 5.2, if (51) is true then :

$$\forall j < J, \ \forall n < N, \ \frac{1}{2} E_{\min}^n \leq E_j^{n+1} \leq \frac{3}{2} E_{\max}^n.$$

Since the scheme is conservative and the energy is positive, there exists a constant $C_4 > 0$ such that : $E_{\max}^n \leq C_4$ for all n. Therefore :

$$\frac{E_{\min}^n}{E_{\max}^n} \geq 2^{-n} C_4^{-1} \frac{E_{\min}^0}{E_{\max}^0}.$$

So a way to ensure (51) is to impose :

$$\Delta t \leq \tilde{C}_4(\sigma h^2 + \varepsilon h)\frac{E^0_{\min}}{E^0_{\max}}2^{-N}, \qquad \tilde{C}_4 = \frac{C_{5.2}^2}{2C_4} \tag{52}$$

Reminding that the timestep Δt, the number of iterations N and the final time T are such that : $T = \Delta t N$ then (52) reads as :

$$\Delta t 2^{\frac{T}{\Delta t}} \leq \tilde{C}_4(\sigma h^2 + \varepsilon h)\frac{E^0_{\min}}{E^0_{\max}}.$$

The left-hand side is minimal in Δt for $\Delta t = T\ln(2)$ and thus :

$$T \leq \underbrace{2^{-\frac{1}{\ln(2)}}\frac{1}{\ln(2)}\tilde{C}_4}_{:=C_4'}(\sigma h^2 + \varepsilon h)\frac{E^0_{\min}}{E^0_{\max}}.$$

In other words, for a given $h > 0$, we can only reach a time $T \leq C_4'(\sigma h^2 + \varepsilon h)E^0_{\min}/E^0_{\max}$.

5.2 Assumptions on the mesh

First we need some assumptions on the regularity of the mesh. We denote by h the characteristic length of the mesh ($h = \Delta x$ for a cartesian mesh). We assume that there exists a numerical constant $C_1 > 0$ independent of h such that, for any cell j, any dof, any node r and any shoulder point $r + 1/2$:

$$|\Omega_j| \leq C_1 h^2, \qquad \frac{1}{|\Omega_j|} \leq \frac{C_1}{h^2}, \qquad N_{\text{dof}} \leq C_1, \qquad \text{Card}(\mathcal{V}_j) \leq C_1, \tag{53}$$

$$\|\tilde{\mathbf{C}}_j^{\text{dof}}\| \leq C_1 h, \qquad \|\alpha_j^{\text{dof}}\| \leq C_1 h, \qquad \|\beta_j^{\text{dof}}\| \leq C_1 h^2, \tag{54}$$

and :

$$\|\beta_r^{-1}\| \leq \frac{C_1}{h^2}, \qquad \left\|\begin{pmatrix}\mathbf{x}_i - \mathbf{x}_j & (\mathbf{x}_i - \mathbf{x}_j)^\perp\end{pmatrix}^{-1}\right\| \leq \frac{C_1}{h}. \tag{55}$$

Therefore we have :

$$\left\|\beta_r^{-1}\tilde{\mathbf{C}}_j^r\right\| \leq \frac{C_1^2}{h}, \qquad \left\|\beta_r^{-1}\alpha_j^r\right\| \leq \frac{C_1^2}{h}, \tag{56}$$

Without loss of generality, we assume that C_1 is large enough and simplify :

$$\left\|\beta_r^{-1}\tilde{\mathbf{C}}_j^r\right\| \leq \frac{C_1}{h}, \qquad \left\|\beta_r^{-1}\alpha_j^r\right\| \leq \frac{C_1}{h}, \tag{57}$$

We also assume that β_r is positive definite for any vertex r and :

$$\forall \xi \in \mathbb{R}^2,\ \langle \xi, \beta_r \xi\rangle \geq \frac{1}{C_1}h^2\|\xi\|^2, \qquad \langle \xi, \alpha_r \xi\rangle \geq \frac{1}{C_1}h\|\xi\|^2, \tag{58}$$

and for any shoulder point $r + 1/2$:

$$\forall \lambda \geq 0, \left\|\tilde{\mathbf{n}}_j^{r+1/2} + \lambda(\mathbf{x}_i - \mathbf{x}_j)\right\| \geq \frac{1}{C_1}(1 + \lambda h), \tag{59}$$

where j and i are the indices of the cells that contain the shoulder point $r + 1/2$.

5.3 Proof of lemma 5.2

5.3.1 First step

Using $1/3 \leq \chi(\mathbf{f}_j) \leq 1$ and the definitions (18) and (20), it can be proved that there exists a constant $\tilde{C}_1 > 0$ independent of $(\varepsilon, E, \mathbf{F})$ such that, for any cell j and any dof :

$$q_j \leq \tilde{C}_1 E_j, \qquad \|\mathbf{u}_j\| \leq \tilde{C}_1, \qquad r_j \leq \tilde{C}_1 E_j, \qquad k_j \leq \tilde{C}_1 E_j, \qquad k_{\text{dof}} \leq \tilde{C}_1 \sum_i E_i, \tag{60}$$

and :

$$q_j \leq \tilde{C}_1 k_j, \qquad \frac{1}{\tilde{C}_1} k_j \leq r_j \leq \tilde{C}_1 k_j, \qquad \frac{1}{\tilde{C}_1} k_{\text{dof}} \leq r_{\text{dof}} \leq \tilde{C}_1 k_{\text{dof}}. \tag{61}$$

Without loss of generality, we assume $\tilde{C}_1 = C_1$.

5.3.2 Second step

In this part we prove an estimate on $\mathbf{u}_{\text{dof}}$ of the form :

$$\|\mathbf{u}_{\text{dof}}\| \leq C_{62} \frac{\varepsilon}{\sigma h + \varepsilon}, \tag{62}$$

for some universal constant $C_{62} > 0$. To this aim, we show an estimate of the form :

$$\|(\varepsilon A_{\text{dof}} + \sigma B_{\text{dof}})^{-1}\| \leq \frac{C_{63}}{h^{m_{\text{dof}}}(\varepsilon + \sigma h)} \frac{1}{\sum_i E_i}, \tag{63}$$

for some universal constant $C_{63} > 0$, with $m_{\text{dof}} = 1$ if dof is a vertex, 0 if it is a shoulder point. Here $\sum_i E_i$ stands for the sum of all the cells containing the degree of freedom dof. The proof uses the facts that the matrices A_{dof} and B_{dof} satisfy :

$$\|A_{\text{dof}}\| \leq C_1^3 h^{m_{\text{dof}}} \sum_i E_i, \qquad \|B_{\text{dof}}\| \leq C_1^3 h^{m_{\text{dof}}+1} \sum_i E_i. \tag{64}$$

First case : the vertices

We define :

$$\tilde{A}_r = \frac{1}{h \sum_i E_i} A_r, \qquad \tilde{B}_r = \frac{1}{h^2 \sum_i E_i} B_r,$$

then :

$$\frac{1}{h(\sigma h + \varepsilon) \sum_i E_i} (\varepsilon A_r + \sigma B_r) = \frac{\varepsilon}{\sigma h + \varepsilon} \tilde{A}_r + \frac{\sigma h}{\sigma h + \varepsilon} \tilde{B}_r, \tag{65}$$

and, due to (58) and (61):

$$\langle \mathbf{x}, \tilde{A}_r \mathbf{x} \rangle \geq \frac{1}{C_1^2} \|\mathbf{x}\|^2, \qquad \langle \mathbf{x}, \tilde{B}_r \mathbf{x} \rangle \geq \frac{1}{C_1^2} \|\mathbf{x}\|^2,$$

Therefore (65) writes as a convex combination of $\tilde{A}_r$ and $\tilde{B}_r$ and we have, for any node r :

$$\forall \lambda \in [0,1], \ \left(\lambda \tilde{A}_r + (1-\lambda)\tilde{B}_r\right) \in \left\{ M | \ \|M\| \leq C_1^3, \min_{\|\mathbf{x}\|=1} \langle \mathbf{x}, M\mathbf{x} \rangle \geq \frac{1}{C_1^2} \right\} := \mathcal{M},$$

Since $\mathcal{M}$ is a compact subset of the set of invertible matrices and since the mapping that gives the inverse of a matrix is a continuous mapping, there exists a constant $C_{\mathcal{M}} > 0$ such that :

$$\forall \lambda \in [0,1], \ \forall r, \ \|(\lambda \tilde{A}_r + (1-\lambda)\tilde{B}_r)^{-1}\| \leq C_{\mathcal{M}},$$

which proves (63) in the case of vertices.

Second case : the shoulder points

One has :

$$\varepsilon A_{r+1/2} + \sigma B_{r+1/2} = \varepsilon \begin{pmatrix} \mathbf{D}_{r+1/2} & \mathbf{D}_{r+1/2}^{\perp} \end{pmatrix}.$$

Since $\begin{pmatrix} \mathbf{D}_{r+1/2} & \mathbf{D}_{r+1/2}^{\perp} \end{pmatrix} / \|\mathbf{D}_{r+1/2}\|$ is a rotation matrix, it is invertible and there exists a universal constant $C_2 > 0$ such that :

$$\|\mathbf{D}_{r+1/2}\| \left\| \begin{pmatrix} \mathbf{D}_{r+1/2} & \mathbf{D}_{r+1/2}^{\perp} \end{pmatrix}^{-1} \right\| \leq C_2,$$

and using (59) one can write :

$$\|\mathbf{D}_{r+1/2}\| = \left\| r_{r+1/2} \tilde{\mathbf{n}}_j^{r+1/2} + k_{r+1/2} \frac{\sigma}{\varepsilon} (\mathbf{x}_i - \mathbf{x}_j) \right\| \geq r_{r+1/2} \frac{1}{C_1} \left(1 + \frac{k_{r+1/2}}{r_{r+1/2}} \frac{\sigma}{\varepsilon} h \right),$$

and thus :

$$\|\mathbf{D}_{r+1/2}\| \geq \frac{1}{C_1^2} \sum_i E_i \left(1 + \frac{\sigma h}{\varepsilon} \right), \qquad \left\| \left(\varepsilon A_{r+1/2} + \sigma B_{r+1/2} \right)^{-1} \right\| \leq C_1^2 C_2 \frac{1}{\sigma h + \varepsilon} \frac{1}{\sum_i E_i}.$$

Therefore (63) is proved. Moreover, on can easily check that : $\|\mathbf{y}_r\| \leq C_1^2 h \sum_i E_i$. Therefore, using equation (48), we have :

$$\frac{1}{\varepsilon} \|\mathbf{u}_r\| \leq \frac{C_{\mathcal{M}} C_1^2}{\varepsilon + \sigma h},$$

and using the definition of $\mathbf{y}_{r+1/2}$ (38), one eventually has :

$$\|\mathbf{y}_{r+1/2}\| \leq C_1^2 \sum_i E_i, \qquad \frac{1}{\varepsilon} \|\mathbf{u}_{r+1/2}\| \leq \frac{C_1^4 C_2}{\varepsilon + \sigma h}.$$

5.3.3 Conclusion of the proof

The fluxes are defined by :

$$\mathbf{G}_j^{\text{dof}} := q_j \tilde{\mathbf{C}}_j^{\text{dof}} + r_{\text{dof}} \alpha_j^{\text{dof}} \mathbf{u}_j - r_{\text{dof}} \alpha_j^{\text{dof}} \mathbf{u}_{\text{dof}} - \frac{\sigma}{\varepsilon} \beta_j^{\text{dof}} \, k_{\text{dof}} \, \mathbf{u}_{\text{dof}},$$

therefore, according to (60), (54) and (62), there exists a constant $C_{66} > 0$ such that :

$$\left\| \mathbf{G}_j^{\text{dof}} \right\| \leq C_{66} \left(E_j h + \sum_{i \in \mathcal{V}_j} E_i h + \sum_{i \in \mathcal{V}_j} E_i h \frac{\varepsilon}{\sigma h + \varepsilon} + \frac{\sigma}{\varepsilon} h^2 \sum_{i \in \mathcal{V}_j} E_i \frac{\varepsilon}{\sigma h + \varepsilon} \right) \leq C_{66} h \sum_{i \in \mathcal{V}_j} E_i, \tag{66}$$

hence, thanks to (62) :

$$\frac{1}{\varepsilon} \left| \langle \mathbf{u}_{\text{dof}}, \mathbf{G}_j^{\text{dof}} \rangle \right| \leq C_{62} C_{66} \frac{h}{\sigma h + \varepsilon} \sum_{i \in \mathcal{V}_j} E_i.$$

The advection terms in (44) can also be bounded from above by $C' h \sum_{i \in \mathcal{V}_j} E_i / (\sigma h + \varepsilon)$, for some universal constant $C' > 0$, thus the right hand side of (44) can be bounded from above by :

$$E_j + C_{5.2} \frac{h}{\sigma h + \varepsilon} \sum_{i \in \mathcal{V}_j} E_i \frac{\Delta t}{h^2},$$

thus leading to :

$$E_j - C_{5.2}\frac{\Delta t}{\sigma h^2+\varepsilon h}\sum_{i\in\mathcal{V}_j} E_i \le E_j^{(1)} \le E_j + C_{5.2}\frac{\Delta t}{\sigma h^2+\varepsilon h}\sum_{i\in\mathcal{V}_j} E_i, \tag{67}$$

and a sufficient condition for keeping the energy positive writes :

$$\Delta t < \frac{1}{C_{5.2}}(\sigma h^2+\varepsilon h)\min_j\left\{\frac{E_j}{\sum_{i\in\mathcal{V}_j} E_i}\right\}. \tag{68}$$

5.3.4 Continuity of $\mathbf{u}_{\mathrm{dof}}/\varepsilon$

The previous computation shows that the mapping $]0,+\infty[\times\mathfrak{D} \ni (\varepsilon, E, \mathbf{F}) \to \mathbf{u}_{\mathrm{dof}}/\varepsilon$ admits a continuous extension on $[0,+\infty[\times\mathfrak{D}$.

5.4 Proof of lemma 5.3

Let $(E,\mathbf{F}) \in \mathfrak{D}$. Using the definition of M_{dof} (26) (39), one can write :

$$\frac{1}{\varepsilon}M_{\mathrm{dof}} = \left(\varepsilon A_{\mathrm{dof}} + \sigma B_{\mathrm{dof}}\right)^{-1} A_{\mathrm{dof}}. \tag{69}$$

According to (63) and (64), there exists a constant $C_{70} > 0$ such that :

$$\|M_{\mathrm{dof}}\| \le C_{70}\frac{\varepsilon}{\sigma h+\varepsilon}. \tag{70}$$

In addition, using the definition of $\mathbf{G}_j^{\mathrm{dof}\,*}$ (28), one can prove the existence of a constant $C_{71} > 0$ such that :

$$\left\|\mathbf{G}_j^{\mathrm{dof}\,*}\right\| \le C_{71}\frac{\varepsilon h}{\sigma h+\varepsilon}\sum_{i\in\mathcal{V}_j} E_i. \tag{71}$$

The advection terms in (45) are also bounded from above by $C'h\sum_{i\in\mathcal{V}_j} E_i/(\sigma h+\varepsilon)$, thus the right hand side of (45) $\mathbf{Y}_j$ satisfies :

$$\|\mathbf{Y}_j\| \le E_j + C_{5.3}\frac{\Delta t}{\sigma h^2+\varepsilon h}\sum_{i\in\mathcal{V}_j} E_i.$$

5.5 Proof of lemma 5.4

According to equation (47), we can write $\tilde{A}_j$ as :

$$\tilde{A}_j = I + \Delta t A'_j, \qquad A'_j = \frac{1}{\varepsilon|\Omega_j|k_j}\sum_{\mathrm{dof}} r_{\mathrm{dof}}\alpha_j^{\mathrm{dof}}(I - M_{\mathrm{dof}}), \tag{72}$$

In addition, one has :

$$I - M_{\mathrm{dof}} = \sigma\left(\varepsilon A_{\mathrm{dof}} + \sigma B_{\mathrm{dof}}\right)^{-1} B_{\mathrm{dof}}.$$

According to (63) and (64), there exists a constant $C_{73} > 0$ such that :

$$\|I - M_{\mathrm{dof}}\| \le C_{73}\frac{\sigma h}{\sigma h+\varepsilon}. \tag{73}$$

and therefore there exists a constant $C_{74} > 0$ such that :

$$\|A'_j\| \le C_{74}\frac{1}{\varepsilon}\frac{\sigma}{\sigma h+\varepsilon}\frac{\sum_{i\in\mathcal{V}_j} E_i}{E_j}, \tag{74}$$

thus if :

$$\Delta t \frac{C_{74}}{\varepsilon} \frac{\sigma}{\sigma h + \varepsilon} \frac{\sum_{i \in \mathcal{V}_j} E_i}{E_j} < 1, \tag{75}$$

then $\tilde{A}_j$ is invertible.

Moreover, under the condition (75), one can prove the following inequalities :

$$\left\| \tilde{A}_j^{-1} \right\| \leq \frac{1}{1 - C_{74} \frac{\Delta t}{\varepsilon} \frac{\sigma}{\sigma h + \varepsilon} \frac{1}{E_j} \sum_{i \in \mathcal{V}_j} E_i},$$

$$\left\| \mathbf{F}_j^{(1)} \right\| \leq \frac{1}{1 - C_{74} \frac{\Delta t}{\varepsilon} \frac{\sigma}{\sigma h + \varepsilon} \frac{1}{E_j} \sum_{i \in \mathcal{V}_j} E_i} \left(E_j + C_{5.3} \frac{\Delta t}{\sigma h^2 + \varepsilon h} \sum_{i \in \mathcal{V}_j} E_i \right),$$

and :

$$\left\| \mathbf{f}_j^{(1)} \right\| \leq \frac{1}{1 - C_{74} \frac{\Delta t}{\varepsilon} \frac{\sigma}{\sigma h + \varepsilon} \frac{1}{E_j} \sum_{i \in \mathcal{V}_j} E_i} \frac{E_j + C_{5.3} \Delta t \sum_{i \in \mathcal{V}_j} E_i \frac{1}{\sigma h^2 + \varepsilon h}}{E_j - C_{5.3} \Delta t \sum_{i \in \mathcal{V}_j} E_i \frac{1}{\sigma h^2 + \varepsilon h}}$$

5.6 Proof of lemma 5.5

First we prove that the matrix $\tilde{A}_j$ of (45) is invertible if ε is small enough, for a given Δt. We can write it as :

$$\tilde{A}_j = \frac{1}{\varepsilon} \left(A_j + \varepsilon H_j \right), \qquad A_j = \frac{\Delta t}{|\Omega_j| k_j} \sum_{\text{dof}} r_{\text{dof}} \alpha_j^{\text{dof}}, \qquad H_j = I - \frac{\Delta t}{|\Omega_j| k_j} \sum_{\text{dof}} r_{\text{dof}} \alpha_j^{\text{dof}} \frac{M_{\text{dof}}}{\varepsilon}. \tag{76}$$

For $(E, \mathbf{F}) \in \mathfrak{D}$, note that :

$$C_1^2 \frac{\sum_{i \in \mathcal{V}_j} E_i}{E_j} \geq \frac{r_{\text{dof}}}{k_j} \geq \frac{1}{C_1^2}.$$

Moreover, the matrix A_j is symmetric positive definite and :

$$\forall \xi \in \mathbb{R}^2, \langle \xi, A_j \xi \rangle \geq \frac{1}{C_1^2} \frac{\Delta t}{h^2} \left\langle \xi, \left(\sum_{\text{dof}} \alpha_j^{\text{dof}} \right) \xi \right\rangle \geq \frac{1}{C_1^3} \frac{\Delta t}{h} \|\xi\|^2,$$

and thus A_j is invertible and there exists a universal constant C_{77} such that :

$$\|A_j^{-1}\| \leq C_{77} \max_{\lambda \in S_p(A_j)} \frac{1}{\lambda} \leq C_{77} C_1^3 \frac{h}{\Delta t}. \tag{77}$$

Moreover, one can write :

$$\tilde{A}_j = \frac{1}{\varepsilon} \left(A_j + \varepsilon H_j \right) = \frac{1}{\varepsilon} A_j \left(I + \varepsilon A_j^{-1} H_j \right), \tag{78}$$

and H_j satisfies :

$$\|H_j\| \leq C_{79} \left(1 + \frac{\Delta t}{\sigma h^2 + \varepsilon h} \frac{1}{E_j} \sum_{i \in \mathcal{V}_j} E_i \right). \tag{79}$$

Therefore, if :

$$\varepsilon \, C_{80} \frac{h}{\Delta t} \left(1 + \frac{\Delta t}{\sigma h^2 + \varepsilon h} \frac{1}{E_j} \sum_{i \in \mathcal{V}_j} E_i \right) < 1, \qquad C_{80} = C_{79} C_{77} C_1^3, \tag{80}$$

then $\left(I+\varepsilon A_j^{-1}H_j\right)$ is invertible and so is $\tilde{A}_j$ (using equation (78)). In addition one can write :

$$\left(I+\varepsilon A_j^{-1}H_j\right)^{-1}=\sum_{l=0}^{+\infty}(-\varepsilon)^l\left(A_j^{-1}H_j\right)^l,$$

and if :

$$\varepsilon C_{79}C_{77}C_1^3\frac{h}{\Delta t}\left(1+\frac{\Delta t}{\sigma h^2+\varepsilon h}\frac{1}{E_j}\sum_{i\in\mathcal{V}_j}E_i\right)\leq\frac{1}{2},\qquad\text{then :}\quad\left\|\left(I+\varepsilon A_j^{-1}H_j\right)^{-1}\right\|\leq\sum_{l=0}^{+\infty}2^{-l}=2.$$

Eventually :

$$\left\|\tilde{A}_j^{-1}\right\|\leq 2C_{77}C_1^3\frac{\varepsilon h}{\Delta t},$$

and using lemma 5.3, we have :

$$\left\|\mathbf{F}_j^{(1)}\right\|\leq C_{81}\frac{\varepsilon h}{\Delta t}\left(E_j+\frac{\Delta t}{\sigma h^2+\varepsilon h}\sum_{i\in\mathcal{V}_j}E_i\right),\tag{81}$$

The dimensionless flux $\mathbf{f}_j^{(1)}=\mathbf{F}_j^{(1)}/E_j^{(1)}$ can therefore be bounded by :

$$\left\|\mathbf{f}_j^{(1)}\right\|\leq C_{81}\frac{\varepsilon h}{\Delta t}\frac{E_j+\Delta t\sum_{i\in\mathcal{V}_j}E_i\frac{1}{\sigma h^2+\varepsilon h}}{E_j-C_{5.2}\Delta t\sum_{i\in\mathcal{V}_j}E_i\frac{1}{\sigma h^2+\varepsilon h}},$$

which is smaller than 1 if ε is small enough.

5.7 Proof of lemma 5.6

Since the scheme is conservative and the energy is positive (by assumption), there exists a constant $C_{5.6}>0$ such that, for any cell j :

$$\sum_{i\in\mathcal{V}_j}E_i\leq C_{5.6}\frac{\|E^0\|_{L^1}}{h^2},\qquad\text{with :}\quad\left\|E^0\right\|_{L^1}=\sum_{j*}|\Omega_j|\left|E_j^0\right|=\sum_{j*}|\Omega_j|E_j^0.\tag{82}$$

hence :

$$\frac{1}{E_j}\sum_{i\in\mathcal{V}_j}E_i\leq C_{5.6}\frac{\|E^0\|_{L^1}}{\delta h^2}.$$

Define :

$$\varepsilon_{\min}^{\delta}:=\frac{\Delta t}{C_{80}h}\left(1+\frac{\Delta t}{\sigma h^2+\varepsilon h}C_{5.6}\frac{\|E\|_{L^1}}{\delta h^2}\right)^{-1}.$$

Equations (69) and (76) show that $M_{\text{dof}}/\varepsilon$ and $\tilde{A}_j^{-1}$ admit continuous extensions on $[0,\varepsilon_{\min}^{\delta}[\times\mathfrak{D}_\delta$. The other coefficients (q_j, $\mathbf{u}_j$ etc) are also continuous with respect to $(E,\mathbf{F})$, thus z is continuous on $[0,\varepsilon_{\min}^{\delta}[\times\mathfrak{D}_\delta$.

5.8 Proof of theorem 5.7

Let $(E^{n,\varepsilon},\mathbf{F}^{n,\varepsilon})_{n\leq N}$ be a numerical solution (N being the number of iterations) that is to say :

$$\forall n<N,\ (E^{n+1,\varepsilon},\mathbf{F}^{n+1,\varepsilon})=z(\varepsilon,E^{n,\varepsilon},\mathbf{F}^{n,\varepsilon}),$$

which can be writen :

$$\forall n < N,\ \begin{cases} E^{n+1,\varepsilon} = z_1(\varepsilon, E^{n,\varepsilon}, \mathbf{F}^{n,\varepsilon}) \\ \mathbf{F}^{n+1,\varepsilon} = z_2(\varepsilon, E^{n,\varepsilon}, \mathbf{F}^{n,\varepsilon}). \end{cases}$$

According to (82, one has, for all $n < N$ and $j < J$ and $\varepsilon > 0$:

$$E_j^{n,\varepsilon} \leq C_{5.6} \frac{\|E^0\|_{L^1}}{h^2},$$

and according to lemma (5.4), there exists a constant $C_{5.7}$ depending on the mesh, on δ and σ such that :

$$\forall n < N,\ \forall j < J,\ \left\|\mathbf{F}_j^{n,\varepsilon}\right\| \leq C_{5.7}\varepsilon,$$

leading to : $(\mathbf{F}^{n,\varepsilon})_{n\leq N} \xrightarrow[\varepsilon\to 0]{} 0$.

The diffusion limit $(E^{n,0})_{n\leq N}$ is defined by :

$$\forall n < N,\ E^{n+1,0} = z_1(0, E^{n,\varepsilon}, \mathbf{0}). \tag{83}$$

Then, since z and z_1 are continuous :

$$E^{1,\varepsilon} = z_1(\varepsilon, E(t=0), \mathbf{0}) \xrightarrow[\varepsilon\to 0]{} z_1(0, E(t=0), \mathbf{0}) \tag{84}$$

Thus, by induction :

$$E^{n,\varepsilon} \xrightarrow[\varepsilon\to 0]{} E^{n,0}. \tag{85}$$

Indeed, if $E^{n,\varepsilon} \xrightarrow[\varepsilon\to 0]{} E^{n,0}$ then :

$$E^{n+1,\varepsilon} = z_1(\varepsilon, \underbrace{E^{n,\varepsilon}}_{\to E^{n,0}}, \underbrace{\mathbf{F}^{n,\varepsilon}}_{\to \mathbf{0}}) \xrightarrow[\varepsilon\to 0]{} z_1(0, E^{n,0}, \mathbf{0}) = E^{n+1,0}.$$

As a conclusion, property (85) is true for any $n \leq N$.

In addition, by assumption , there exists a constant $\delta > 0$ such that :

$$\forall \varepsilon \in]0, \varepsilon_{\min}^{\delta}[, \forall n \leq N, \forall j < J,\ E_j^{n,\varepsilon} \geq \delta.$$

Thus :

$$\forall n \leq N,\ \forall j < J,\ E_j^{n,0} \geq \delta.$$

Therefore $E^{n,0}$ does not vanish and :

$$q_j^{n,\varepsilon} \xrightarrow[\varepsilon\to 0]{} \frac{E_j^{n,0}}{3}, \qquad k_{\text{dof}}^{n,\varepsilon} \xrightarrow[\varepsilon\to 0]{} \frac{4E_{\text{dof}}^{n,0}}{3} = k_{\text{dof}}^{n,0}, \qquad \mathbf{u}_j^{n,\varepsilon} \xrightarrow[\varepsilon\to 0]{} \mathbf{0}, \qquad \mathbf{y}_r^{n,\varepsilon} \xrightarrow[\varepsilon\to 0]{} \frac{1}{3}\sum_i E_i^{n,0}\tilde{\mathbf{C}}_i^r = \mathbf{y}_r^{n,0}$$

$$\mathbf{y}_{r+1/2}^{n,\varepsilon} \xrightarrow[\varepsilon\to 0]{} \begin{pmatrix} (E_j^{n,0} - E_i^{n,0})/3 \\ \sigma k_{r+1/2}^{n,0} \langle \mathbf{u}_r^{n,0} + \mathbf{u}_{r+1}^{n,0}, (\mathbf{x}_i - \mathbf{x}_j)^{\perp} \rangle / 2 \end{pmatrix} = \mathbf{y}_{r+1/2}^{n,0}, \qquad \frac{1}{\varepsilon}\mathbf{u}_{\text{dof}}^{n,\varepsilon} \xrightarrow[\varepsilon\to 0]{} (B_{\text{dof}}^{n,0})^{-1}\mathbf{y}_{\text{dof}}^{n,0} = \mathbf{u}_{\text{dof}}^{n,0},$$

which writes, for a given node r :

$$\sigma k_r^{n,0}\beta_r \mathbf{u}_r^{n,0} = \frac{1}{3}\sum_i E_i^{n,0}\tilde{\mathbf{C}}_i^r,$$

and for a given shoulder point $r + 1/2$:

$$\begin{cases} \langle \mathbf{u}^{n,0}_{r+1/2}, (\mathbf{x}_i - \mathbf{x}_j)\rangle = \dfrac{E^{n,0}_j - E^{n,0}_i}{3\sigma k^{n,0}_{r+1/2}}, \\ \langle \mathbf{u}^{n,0}_{r+1/2}, (\mathbf{x}_i - \mathbf{x}_j)^\perp\rangle = \dfrac{1}{2}\langle \mathbf{u}^{n,0}_r + \mathbf{u}^{n,0}_{r+1}, (\mathbf{x}_i - \mathbf{x}_j)^\perp\rangle. \end{cases}$$

As a conclusion, $(E^{n,0})_{n\leq N}$ is solution to (42).

5.9 Invertibility of the matrix β^r

In this section, we give some suffisant conditions to ensure that the matrix $\beta^r = \sum_i \beta^r_i$, with $\beta^r_j = \tilde{\mathbf{C}}^r_j \otimes (\mathbf{x}_r - \mathbf{x}_j)$, is invertible. In [16], this result is proved, under some conditions on the mesh, for the polygonal scheme. Here we prove this result for a conical degenerate mesh ($\omega = 0$) and thus the result still holds true for ω small enough since the determinant is a continuous mapping. For $\omega = 0$ the coefficient $\tilde{\mathbf{C}}^r_j$ writes, according to equation (8) :

$$\tilde{\mathbf{C}}^r_j = \frac{1}{2}\left((1-\frac{\pi}{2})\tilde{\mathbf{N}}_{r-1,r} + (1-\frac{\pi}{2})\tilde{\mathbf{N}}_{r,r+1} + \frac{\pi}{2}\tilde{\mathbf{N}}_{r-1/2,r} + \frac{\pi}{2}\tilde{\mathbf{N}}_{r,r+1/2}\right)$$

where $\tilde{\mathbf{N}}_{r-1,r}$ is the normal vector to $[\mathbf{x}_r, \mathbf{x}_{r+1}]$ and $\tilde{\mathbf{N}}_{r,r+1/2}$ is the normal vector to $[\mathbf{x}_r, \mathbf{x}_{r+1/2}]$. Since the shoulder point is given by $\mathbf{S}_{r+1/2} = (\mathbf{M}_r + \mathbf{M}_{r+1})/2$ then :

$$\tilde{\mathbf{N}}_{r-1/2,r} = \frac{1}{2}\tilde{\mathbf{N}}_{r-1,r}, \qquad \tilde{\mathbf{N}}_{r,r+1/2} = \frac{1}{2}\tilde{\mathbf{N}}_{r,r+1}$$

and :

$$\tilde{\mathbf{C}}^r_j = \frac{1}{2}\left(1-\frac{\pi}{4}\right)(\tilde{\mathbf{N}}_{r-1,r} + \tilde{\mathbf{N}}_{r,r+1}) = \left(1-\frac{\pi}{4}\right)\frac{1}{2}(\mathbf{x}_{r+1} - \mathbf{x}_{r-1})^\perp = \left(1-\frac{\pi}{4}\right)\underbrace{\mathbf{C}^r_j}_{\text{polygonal}}$$

thus $\beta_r = (1-\pi/4)\beta^{\text{polygonal}}_r$ is invertible.

6 Numerical results

This section is dedicated to several numerical test cases. First we show some examples for which the conical scheme (40) gives much better results than the polygonal one. We also describe two particular regimes of the M_1 model for which exact solutions can be computed : the streaming regime (section 6.2) and the diffusion regime (section 6.3). These examples allow to compute the convergence rate of the scheme.
In some test cases, the radiative energy can be 0, then we define $\mathbf{f}$ as :

$$\mathbf{f} = \begin{cases} \dfrac{\mathbf{F}}{E} \text{ if } E > 0, \\ \mathbf{0} \text{ else.} \end{cases}$$

When no other precision is given, the timestep is chosen as : $\Delta t = (\Delta x)^2/10$, with $\Delta x = 1/N_x$, N_x being the number of cells in abscissa. The number of cells in the y direction is denoted by N_y.

For the $1D$ test cases (the solution does not depend on y) (sections 6.2.1, 6.2.3 and 6.3.1), the solutions are computed using the conical degenerate ($\omega = 0$) scheme only. We observed very few differences with the polygonal scheme.

In practice, conical meshes are computed in the following ways :

- for cartesian, Kershaw type (21) and Voronoi type (16) meshes, a control point is added to each edge at a distance worth 20% of its length,
- for a radial mesh (18), a control point is added to each edge and the weights ω of the edges are chosen so the edges are circles center at $(0.5, 0.5)$ (cf [27]).

6.1 Propagation of a Dirac mass

The initial data is a Dirac mass for E (86) on a cell j located at the center of the domain and we compare the polygonal and conical schemes on several meshes.

$$E(0,x,y) = \begin{cases} \dfrac{1}{|\Omega_j|} \text{ if } (x,y) \in \Omega_j, \\ 0 \text{ else,} \end{cases} \qquad \mathbf{F}(t=0) = \mathbf{0}. \tag{86}$$

This model is close to the streaming regime of section 6.2. The final time is $t = 0.2$ and $\sigma = 0$, $\varepsilon = 1$. The simulation failed when using the polygonal scheme (41) (we obtained values of $\|\mathbf{f}\|$ larger than $4/3$, thus making it impossible to compute the Eddington factor (2). However, the conical scheme (40) proved to be more stable since the values of $\|\mathbf{f}\|$ remained smaller than $4/3$. Figure 7 shows the results.

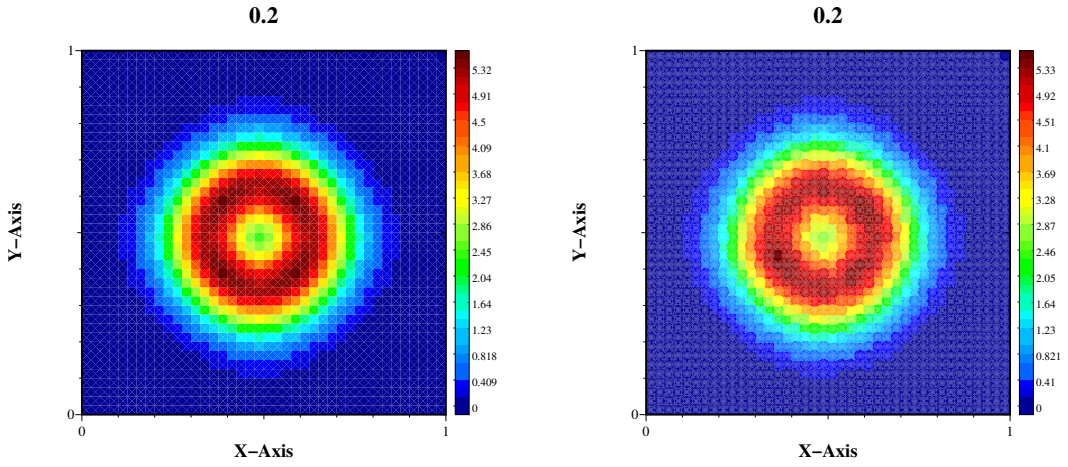

Figure 7: Numerical solution for $N_x = N_y = 40$ using the conical degenerate (left) and conical parabolic (right) scheme at $t = 0.2$ with initial data (86).

6.2 Streaming regime test cases

This regime is charaterised by $\sigma = 0$, $\varepsilon = 1$ and $\|\mathbf{f}\| = 1$. Then $q = 0$ and $\mathbf{u} = \mathbf{f}$. Periodic boundary conditions are imposed.

Remark 6. *If $\mathbf{F} = E\mathbf{a}$ with $\mathbf{a}$ constant, $\|\mathbf{a}\| = 1$, then $(E, \mathbf{F})$ is solution of the M_1 model if and only if E satisfies :*

$$\partial_t E + div(E\mathbf{a}) = \partial_t E + \langle \nabla E, \mathbf{a} \rangle = 0, \qquad \textit{that is to say :} \qquad E(t,\mathbf{x}) = E(0, \mathbf{x} - t\mathbf{a}).$$

In such a case, the scheme (40) reduces to the upwind scheme (94) and : $\forall n$, $\mathbf{F}^n = E^n\mathbf{a}$. This property is used in the first and the second test cases of this section.

6.2.1 Transport of a step function

This test case comes from [16]. The initial data is :

$$E(0,x,y) = F_x(0,x,y) = \begin{cases} 1 \text{ si } x \in [0.4, 0.6], \\ 0.0001 \text{ sinon}, \end{cases} \qquad F_y = 0, \tag{87}$$

and $\sigma = 0$, $\varepsilon = 1$. The quantities E and $\mathbf{F}$ are transported at velocity $\mathbf{a} = (1,0)$. Figure 8 (right) shows the absolute L^1 error on E as a function of the space step Δx. The number of cells in the y direction, denoted N_y, is set at 1. The rate of convergence is 0.5, which is expected for a discontinuous solution.

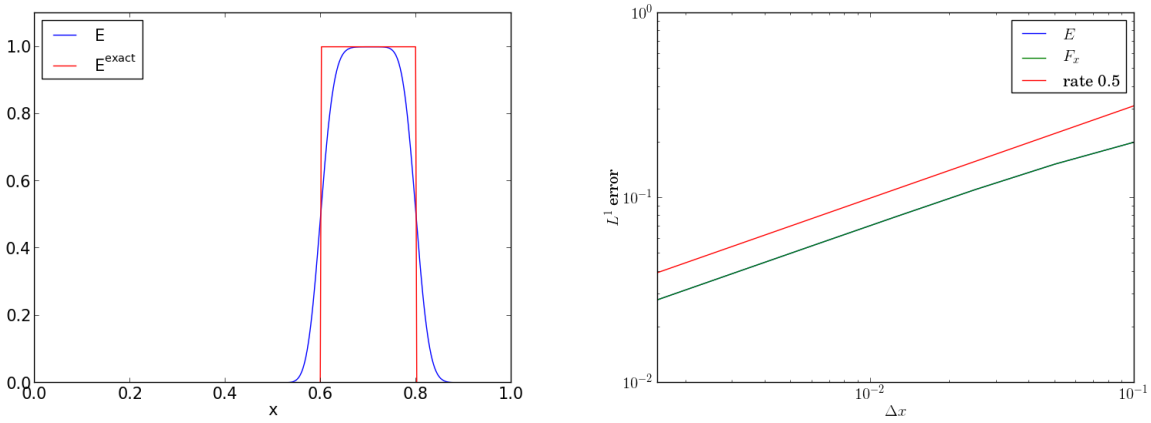

Figure 8: Numerical solution for $N_x = 400$ and $N_y = 1$ using the conical degenerate scheme at $t = 0.2$ (left) and L^1 error on E (right) with initial data (87).

6.2.2 Smooth solution

The initial data of this test case is smooth and writes :

$$(E, F_x, F_y)(0, \mathbf{x}) = \left(g(\mathbf{x}), \frac{g(\mathbf{x})}{\sqrt{2}}, \frac{g(\mathbf{x})}{\sqrt{2}} \right), \qquad g(\mathbf{x}) = 1 + e^{-\gamma ||\mathbf{x} - \mathbf{x_0}||^2} \tag{88}$$

with : $\mathbf{x_0} = (0.25, 0.25)$, $\gamma = 100$, and $\sigma = 0$, $\varepsilon = 1$. Here E and $\mathbf{F}$ are transported with a velocity $\mathbf{a} = (1,1)/\sqrt{2}$.

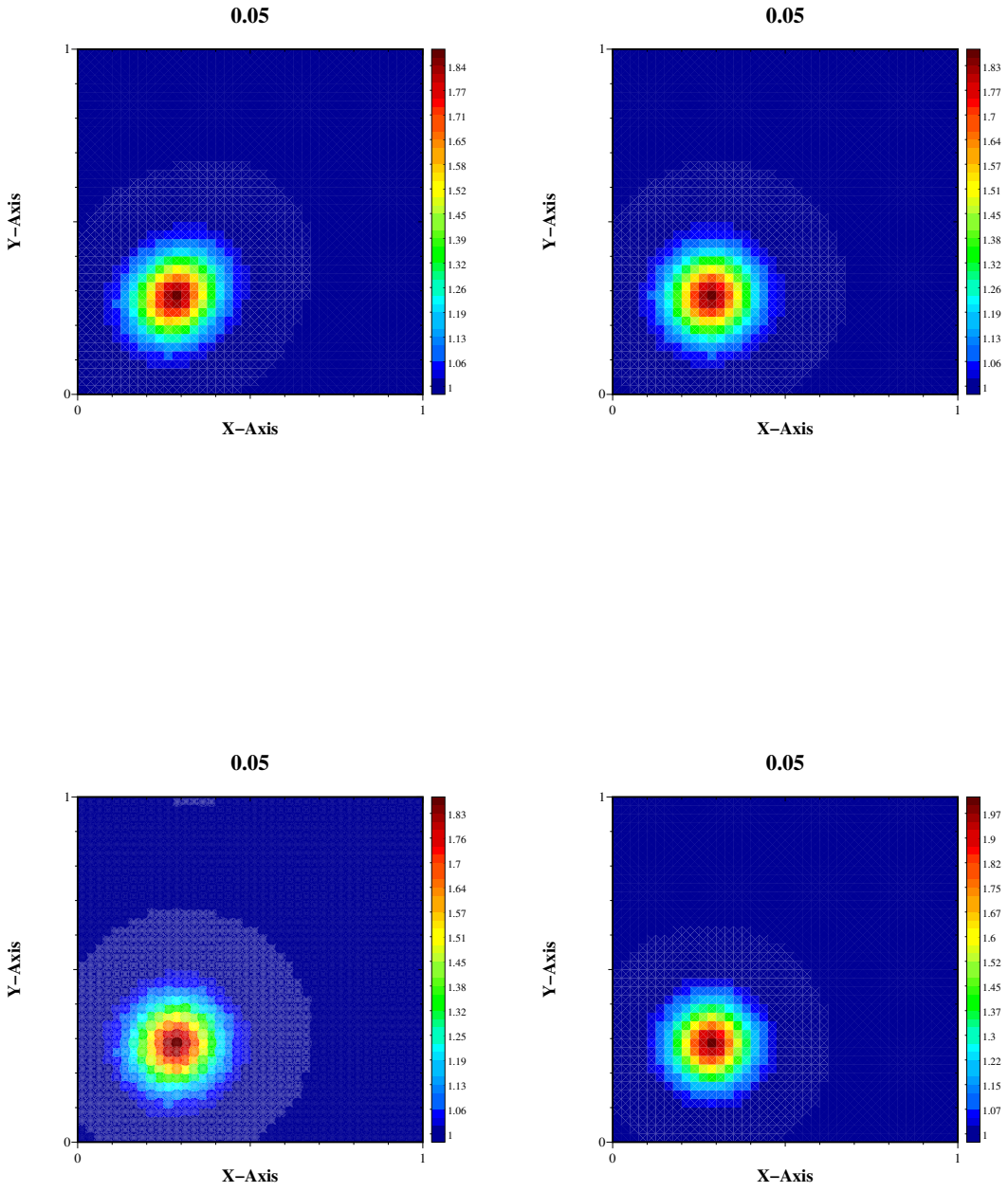

Figure 9: Numerical solution on a cartesian mesh ($N_x = N_y = 40$) with the polygonal scheme (41) (up left), conical degenerate (up right), conical parabolic (down left) and exact solution (down right) at times $t = 0.05$ with initial data (88).

Figure 9 shows that the conical solution is closer to the exact solution than the polygonal one. Figure 10 shows the L^1 error as a function of the space step $\Delta x = 1/N_x = \Delta y = 1/N_y$. A first order rate is recovered for both schemes. A second order rate is also recovered when using the reconstruction described in section 4.2.

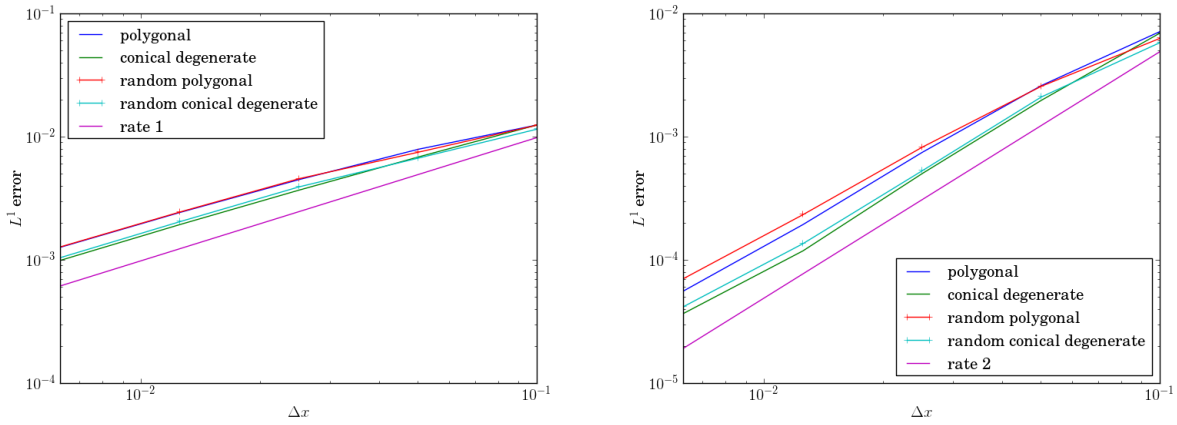

Figure 10: L^1 error on E with the first order scheme (40) (left), second order reconstruction (right) and with initial data (88).

6.2.3 Singular solution

This test case comes from [29]. The initial data is :

$$(E, F_x, F_y)(0, x, y) = \left(1, \frac{0.7 - x}{0.4}, \sqrt{E^2 - F_x^2}\right) \text{ if } x \in]0.3,\ 0.7[, \qquad 0, \text{ elsewhere,} \tag{89}$$

and $\sigma = 0$, $\varepsilon = 1$. The exact solution is defined for $t < 0.4$:

$$(E, F_x, F_y)(t, x, y) = \left(\frac{0.4}{0.4 - t}, 0.4\frac{0.7 - x}{(0.4 - t)^2}, \sqrt{E^2 - F_x^2}\right) \text{ if } x \in]0.3 + t,\ 0.7[, \qquad 0, \text{ elsewhere.}$$

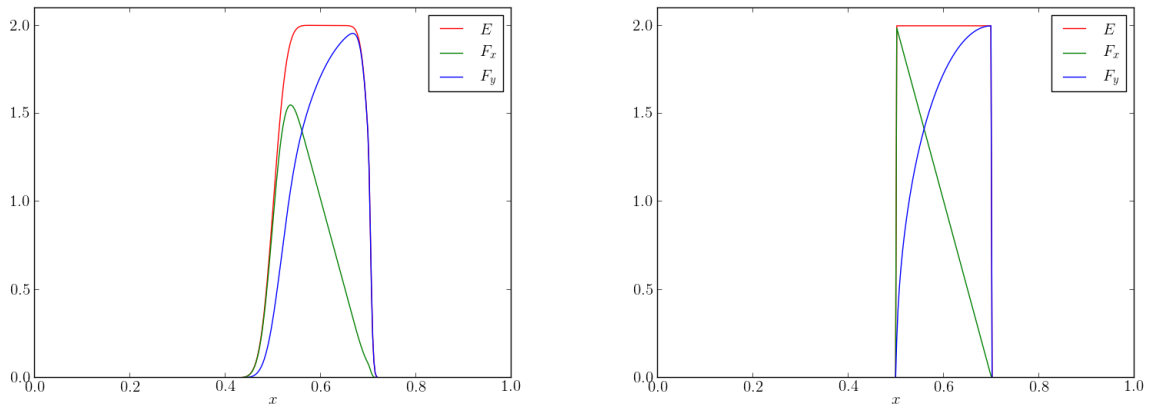

Figure 11: Numerical solution (left) computed with the conical degenerate scheme and exact solution (right) for $N_x = 400$ and $N_y = 1$ at $t = 0.2$ with initial data (89).

Figure 12 shows the L^1 error on E as a function of $\Delta x = 1/N_x$ computed with the conical degenerate scheme, N_y being set at 1. The results of the polygonal scheme are almost identical. The rate of convergence is 0.5, which is expected for a discontinuous solution.

For this test case, we had to choose $r_j^{\text{dof}} = r_j$ instead of $r_j^{\text{dof}} = r_{\text{dof}}$. Indeed the later lead to values of the $\mathbf{f}$ such that $\|\mathbf{f}\| > 4/3$, thus making the Eddington factor (2) ill-defined. We noticed that the choice $r_j^{\text{dof}} = r_j$ allowed to overcome that difficulty. One possible explanation is the following. The coefficient r_j^{dof} is the analogous of the acoustic impedance for the gas dynamic equations (12). It has been studied in [35] that choosing the value of the acoustic impedance at the centers of the cells (in our case, this corresponds to $r_j^{\text{dof}} = r_j$) makes the CFL condition much less restrictive than in the case where it is computed at the boundaries (here it corresponds to $r_j^{\text{dof}} = r_{\text{dof}}$).

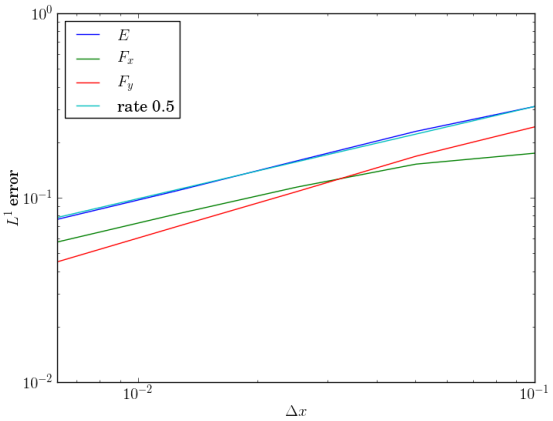

Figure 12: L^1 error on E using the conical degenerate scheme with initial data (89)

Remark 7. *We used the second order reconstruction 4.2 on this test case. We noticed that if the quantity* $\mathbf{F}$ *was directly reconstructed then the norm of the dimensionless flux* $\|\mathbf{f}\|$ *could be larger than* $4/3$*, thus making the Eddington factor (2) ill-defined. This is no longer the case when the quantity* $\mathbf{f}$ *is reconstructed.*

6.3 Diffusion regime

Here $\sigma = 1$, ε is chosen small enough and $\mathbf{F}(t = 0) = \mathbf{0}$. The conical scheme (40) is compared with the conical diffusion limit scheme (40) and the exact solution of the diffusion equation (the exact solution for $\mathbf{F}$ being $\mathbf{0}$). The final time is $T_f = 0.003$.

6.3.1 $1D$ test case

This test case comes from [29]. The initial data is :

$$E(0, x, y) = \begin{cases} 1 \text{ if } x \in]0.4, 0.6[, \\ 10^{-6} \text{ else.} \end{cases} \tag{90}$$

The exact solution is given by :

$$E(t, x, y) = \frac{1 - 10^{-6}}{2} \left(\text{erf}\left(\frac{x - 0.4}{\sqrt{4t\kappa}} \right) - \text{erf}\left(\frac{x - 0.6}{\sqrt{4t\kappa}} \right) \right) + 10^{-6}, \qquad \kappa = \frac{1}{3\sigma},$$

Figure 13 shows the numerical solution computed with the conical degenerate scheme with $\varepsilon = 0.015$ and $\sigma = 1$. The curve given in [29] is well recovered. The curves computed with smaller values of ε (10^{-3}, 10^{-4}) are quite similar.

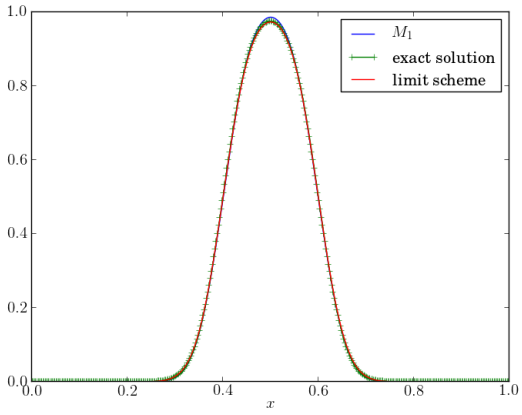

Figure 13: Numerical solution computed with the conical degenerate scheme, $N_x = 400$, $N_y = 1$, with $\varepsilon = 0.015$ and initial data (90).

Figure 14 shows the L^1 error on E as a function of $\Delta x = 1/N_x$, $N_y = 1$. The convergence rate is 1 (2 with the reconstruction step). We only display the results for the conical degenerate mesh since the polygonal scheme gives almost the same results in this test case.

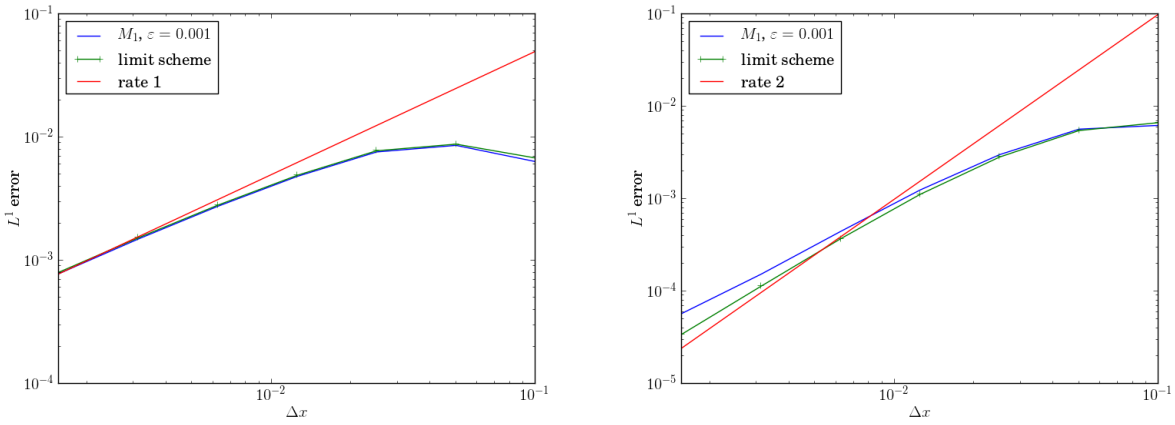

Figure 14: L^1 error on E using the scheme (40) (left), with the reconstruction step (right) with initial data (90).

6.3.2 Propagation of a Dirac mass in the diffusion limit

The diffusion limit schemes (42) and (43) are compared. The initial data is a Dirac mass and it is given by (86) and $\sigma = 1$. The exact solution is the fundamental solution to the diffusion equatin and it is given by :

$$E(t, \mathbf{x}) = \frac{1}{4\pi t\kappa} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x_0}\|^2}{4t\kappa}\right), \qquad \kappa = \frac{1}{3\sigma}, \qquad \mathbf{x_0} = (0.5, 0.5). \tag{91}$$

The final time is $t = 0.01$.

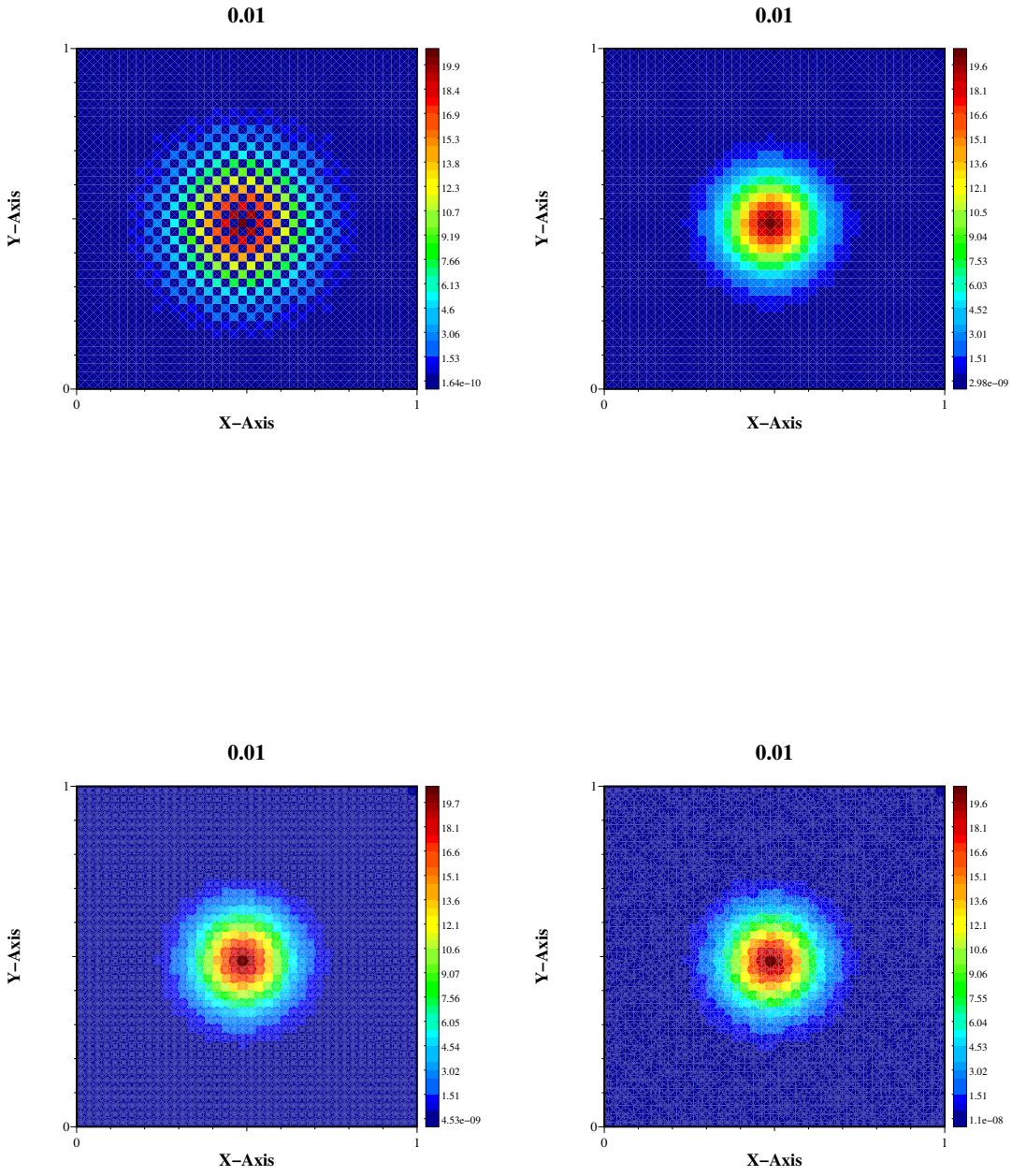

Figure 15: Solution computed on a cartesian mesh with $N_x = N_y = 41$ with the polygonal scheme (43) (up left), conical degenerate (up right), conical parabolic (down left) and conical random (the weights ω are chosen at random between 0 and 50) at $t = 0.01$ with initial data (86) and $\sigma = 1$.

Figure 15 (up left) shows a well known drawback of the node-based schemes (cf [16] and [27]). This phenomena is called *cross-stencil* : the information propagates alongside the diagonals lines. This drawback disappears when using the conical scheme due to the non-null contribution of the shoulder points. The energy can thus propagate in cells which share a common edge and the conical solution is much closer to the exact solution.

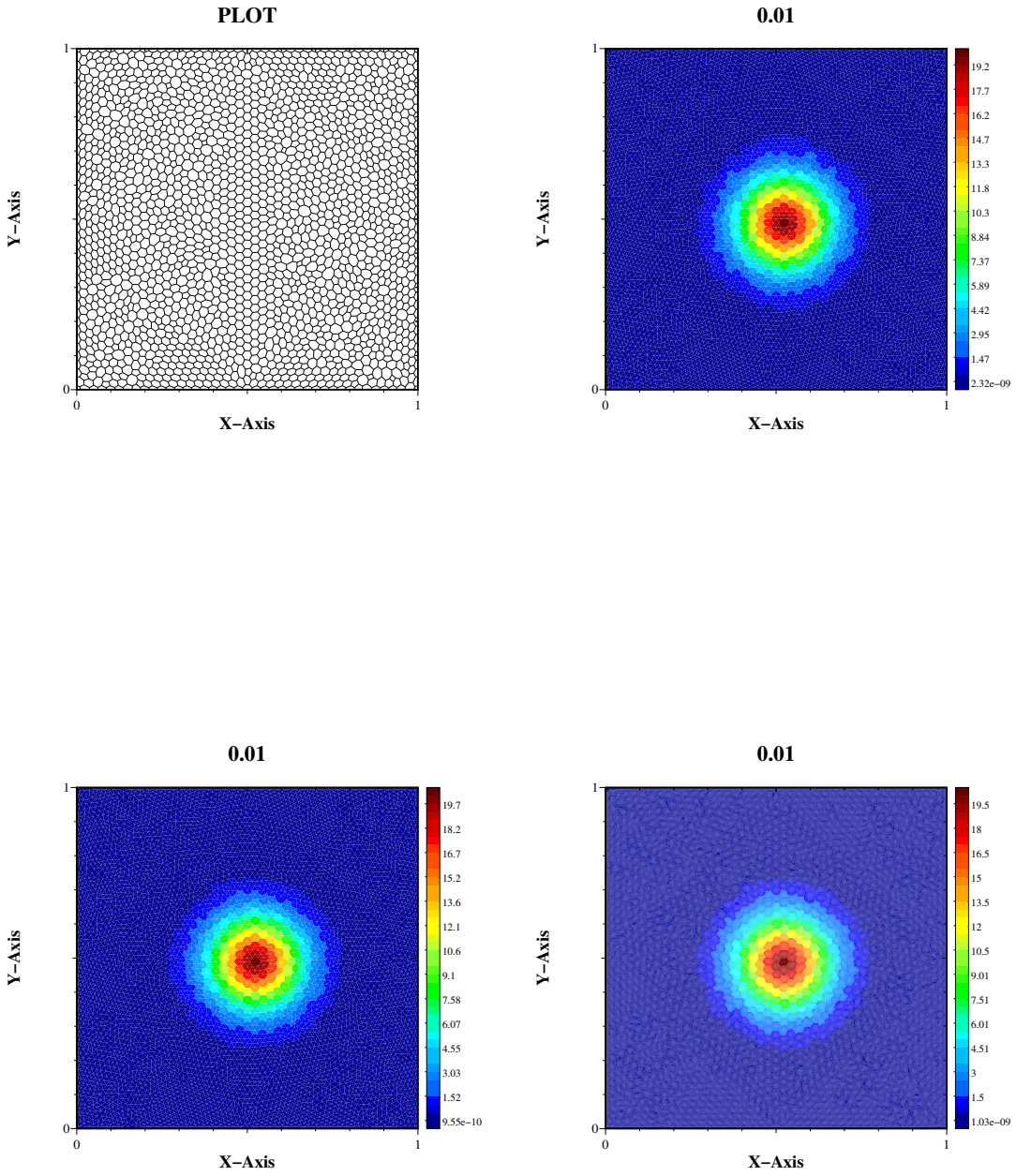

Figure 16: Voronoi type mesh (up left) and solution computed using the polygonal mesh (up right), conical degenerate (down left) and conical parabolic (down right) at $t = 0.01$ with initial data (86) and $\sigma = 1$.

As in [27], the results of the polygonal and conical schemes are quite similar on the Voronoi type mesh.

6.3.3 Fundamental solution of the diffusion equation

This test case comes from [16]. We chose $\sigma = 1$ and $\varepsilon = 0.0001$. The initial data is the fundamental solution of the diffusion equation (91) at time $t = t_0 = 0.01$. The exact solution is the fundamental solution of the diffusion equation at time $t = t_0 + T_f$ with $T_f = 0.003$.

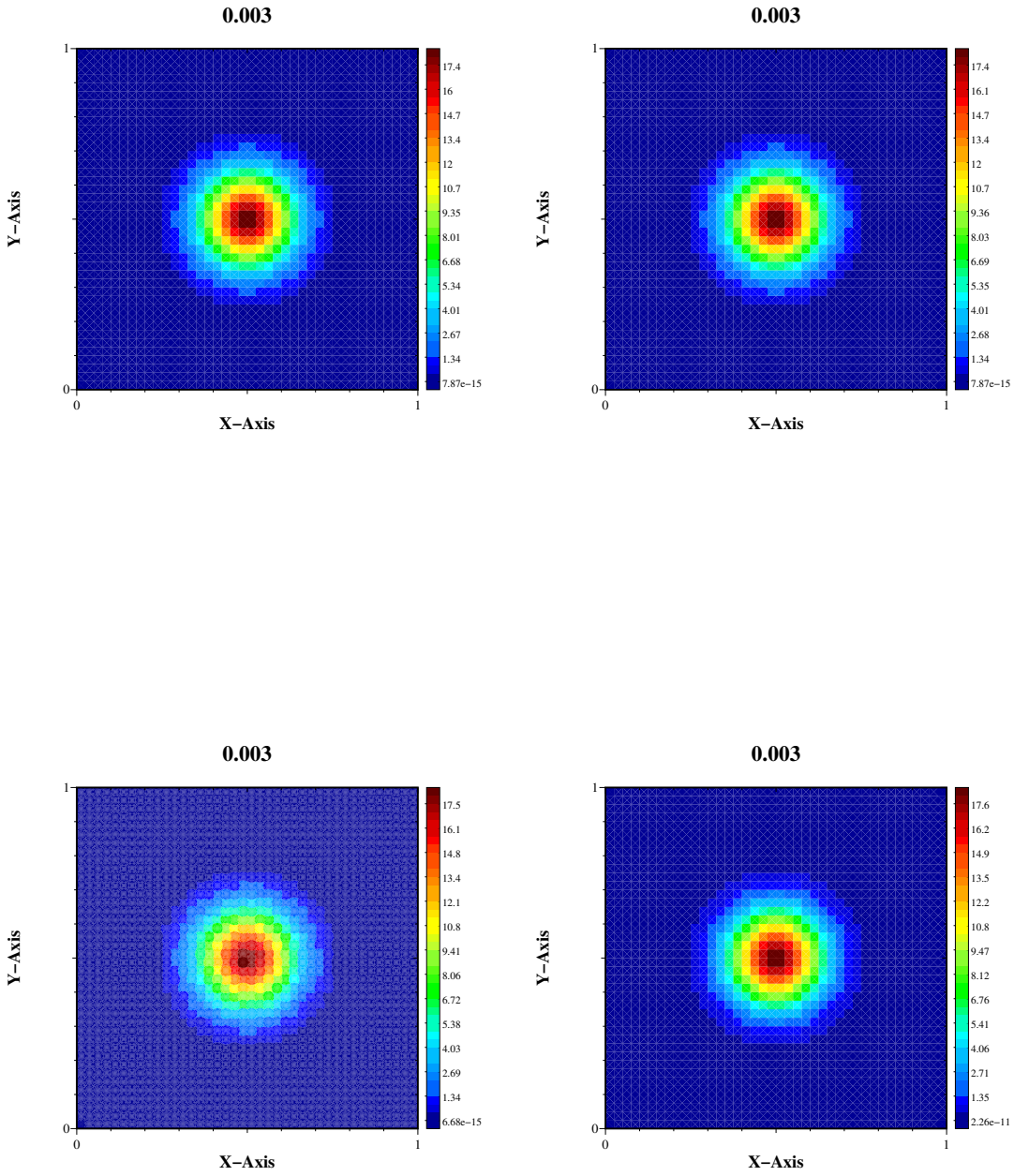

Figure 17: Numerical solution with the polygonal scheme (up left), conical degenerate (up right), conical parabolic (down left) and exact solution (down right) with $N_x = N_y = 40$ on a cartesian mesh with initial data (91) et $\varepsilon = 0.0001$.

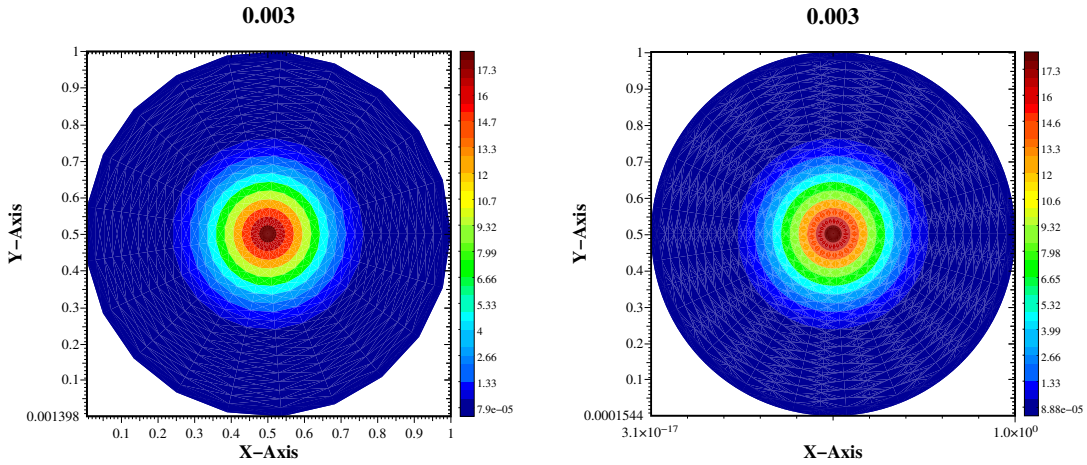

Figure 18: Numerical solution on a radial polygonal mesh (left) and on a conical one (right) at time $t = 0.003$ with initial data (91).

The results computed on a cartesian mesh are quite similar (Figure 17). However, we observe a significant difference on the Figure 18, the results of the conical scheme are much better than those of the polygonal one. Figure 19 show the L^1 error on E on cartesian meshes as a function of $\Delta x = 1/N_x = \Delta y = 1/N_y$. The convergence rate is 1 (2 with the reconstruction step).

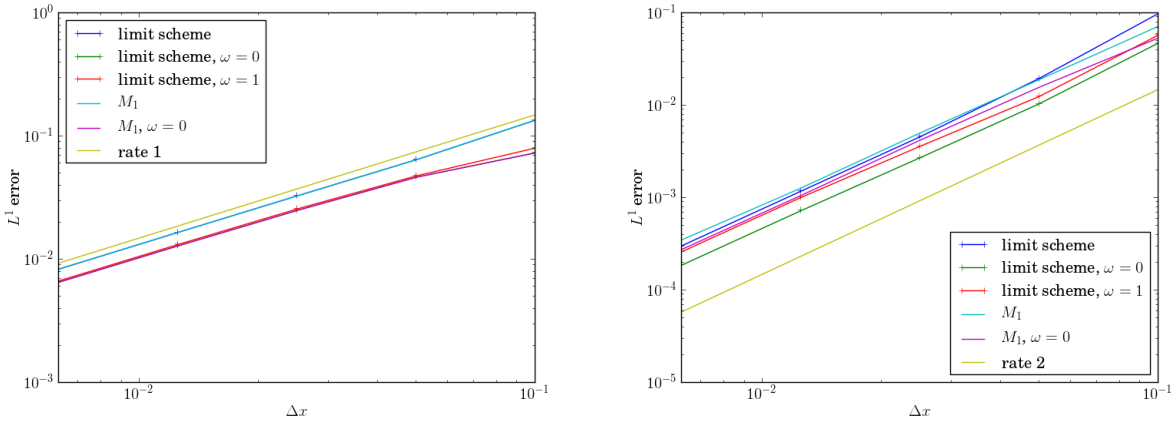

Figure 19: L^1 error on E using the scheme (40) (left), with the reconstruction step (right) with initial data (91).

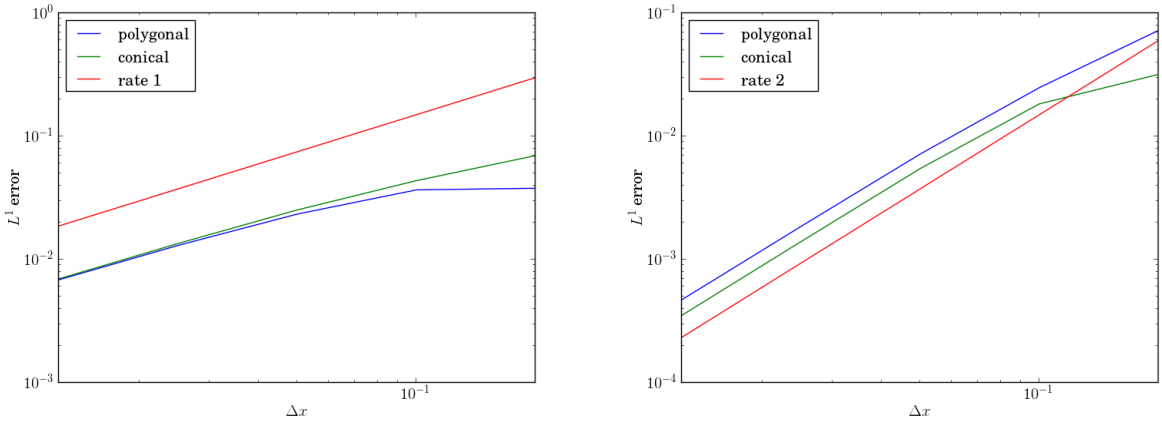

Figure 20: L^1 error on E as a function of $N_x = N_y$ on a radial mesh using the limit scheme (left) and the reconstruction step (right) with initial data (91) and homogeneous Dirichlet boundary condition.

For the mesh of Figure 21, the timestep is given by $\Delta t = (\Delta x)^2/100$. For the mesh of Figure 20, the timestep is $\Delta t = (\Delta x)^2/10000$.

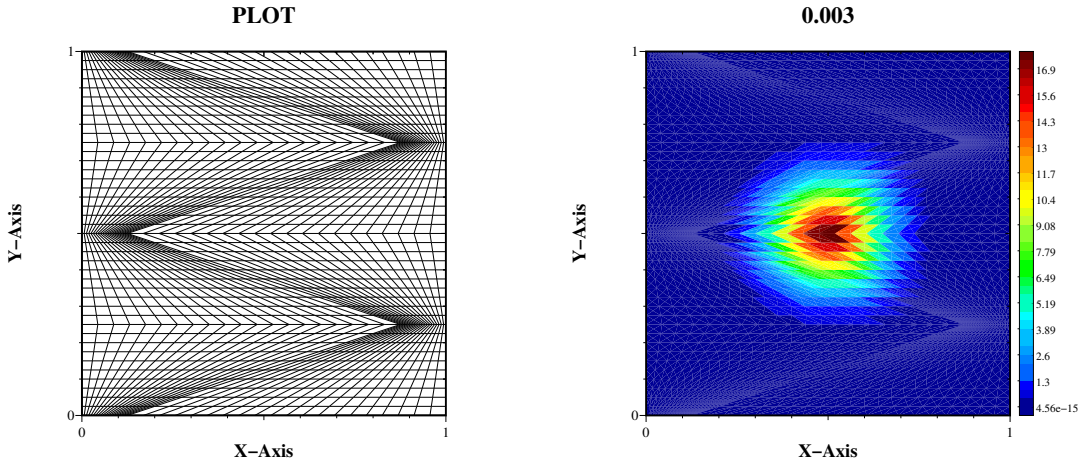

Figure 21: Kershaw type mesh of size 40×40 and solution computed with the conical degenerate scheme (40).

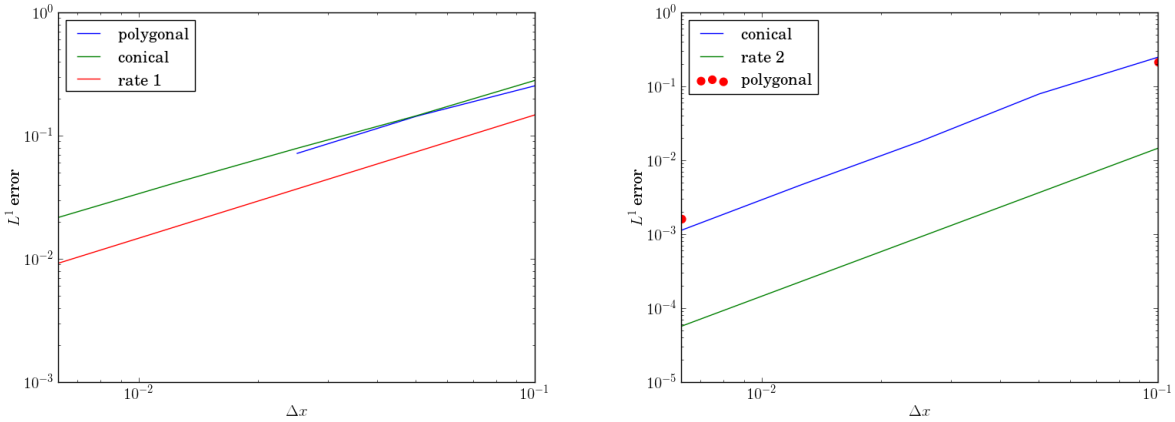

Figure 22: L^1 error on E on the mesh of Figure 21 using the scheme (40) (left), using the reconstruction step (right) with initial data (91).

Figure 22 displays the convergence analysis on Kershaw type meshes. The missing points on the curves for the polygonal scheme are due to instabilities : the scheme computed values of $\|\mathbf{f}\|$ larger than $4/3$, while the conical one did not. This illustrates the fact that the conical scheme is more stable the the polygonal one.

6.4 Comparison with the limit scheme of another model

In this section, we compare the limit schemes of M_1 model and P_1 model (92) in the diffusion limit. The later writes :

$$\begin{cases} \partial_t E + \dfrac{1}{\varepsilon}\text{div } \mathbf{F} = 0, \\ \partial_t \mathbf{F} + \dfrac{1}{\varepsilon}\nabla E = -\dfrac{\sigma}{\varepsilon^2}\mathbf{F}, \end{cases} \tag{92}$$

A conical scheme for P_1 model can be found in [27]. Both models admit the same diffusion limit, thus we compare here two schemes that discretise the same PDE. We set $\sigma = 1$ and we use the following Kershaw type mesh :

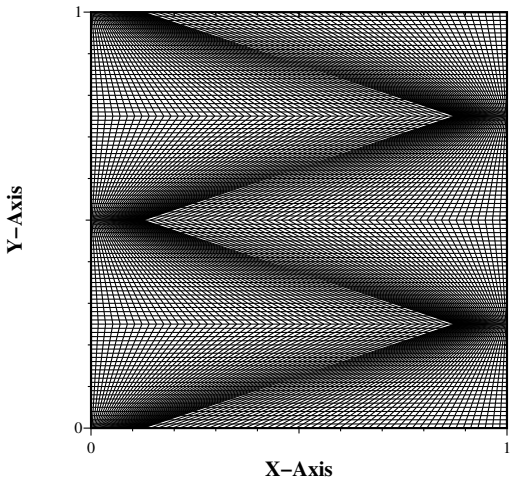

Figure 23: Kershaw type mesh of size 100×100.

For this test case, the timestep is given by $\Delta t = (\Delta x)^2/1000$.
Figure 24 shows the numerical solutions with a smooth initial data (91). The results of the two schemes are quite similar.

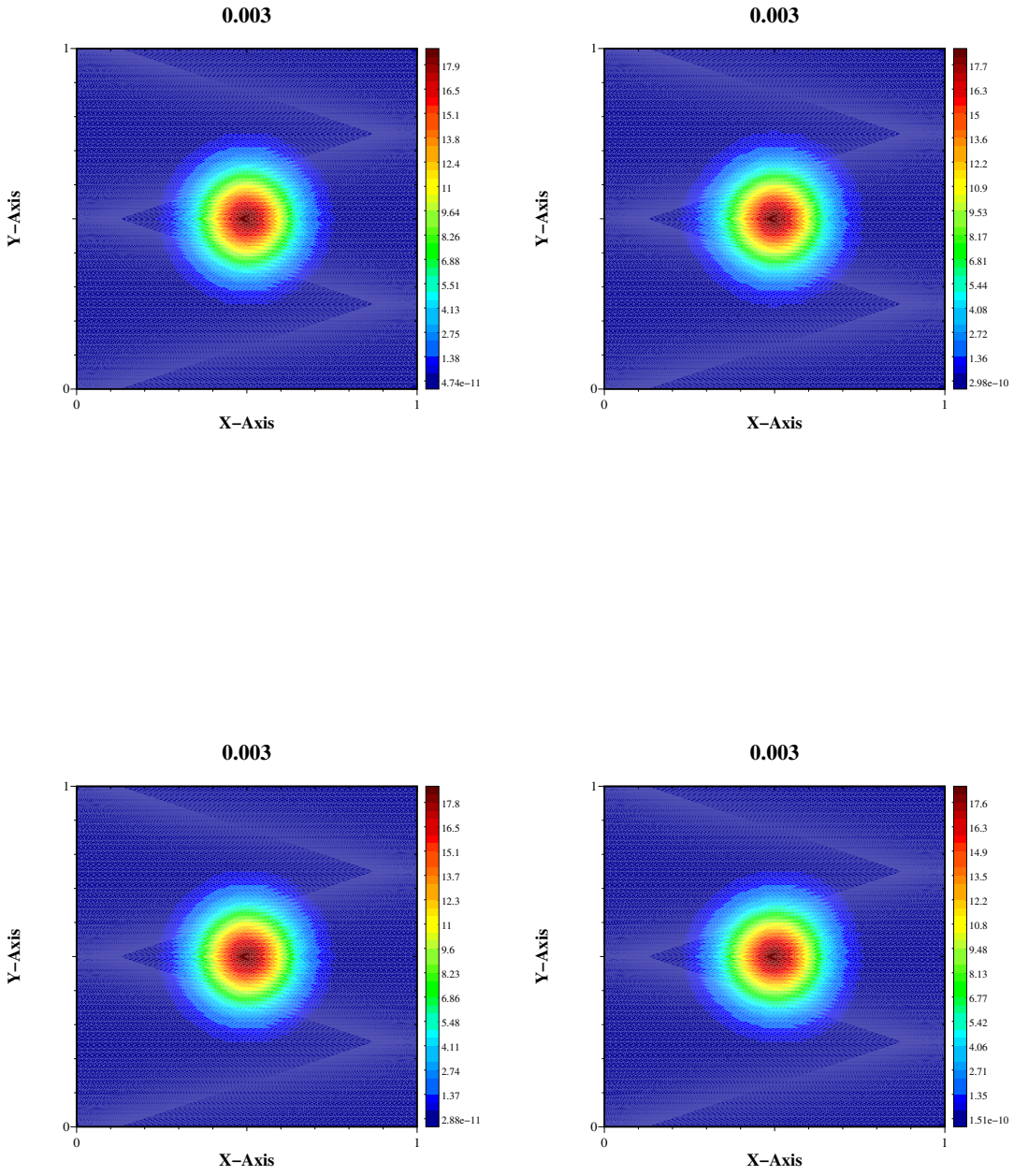

Figure 24: Solution computed with the polygonal P_1 scheme (up left), polygonal M_1 (up right), conical degenerate P_1 (down left) and conical degenerate M_1 (down right) with initial data (91).

Figure 25 shows the solution with a discontinuous initial data given by :

$$E(0,x,y) = \begin{cases} 1 \text{ if } x \ \in]0.4, 0.6[, \\ 10^{-12} \text{ else.} \end{cases} \tag{93}$$

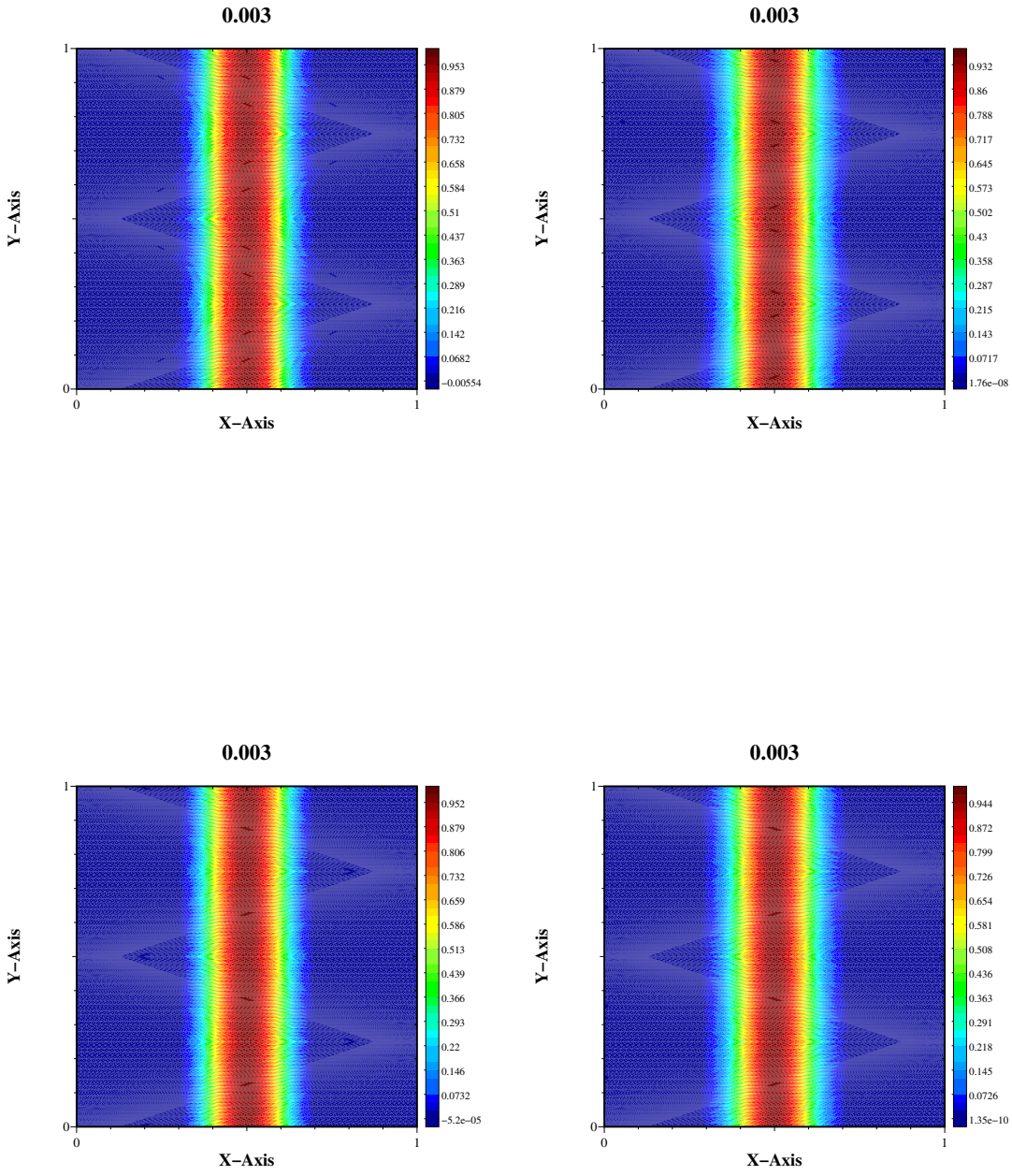

Figure 25: Solution computed with the polygonal P_1 scheme (up left), polygonal M_1 (up right), conical degenerate P_1 (down left) and conical degenerate M_1 (down right) with initial data (93).

One can notice that the P_1 limit scheme does not guarantee the positivity of the solution, while the M_1 limit scheme does.

7 Conclusion

In this work, we adapted a numerical scheme defined for the M_1 model on polygonal meshes to conical meshes. We numerically observed that the good properties of the original scheme were preserved (positive energy, limited flux and AP property). We also proposed a rigorous proof of the convergence of the scheme toward the diffusion limit scheme (this is the AP property). The test cases from [29] and [16] have been reproduced. We also highlighted some drawbacks of the polygonal scheme and we showed that they disappear when using the conical scheme. Moreover, we adapted a high order reconstruction procedure and obtained a second order convergence in space for different values of σ and ε (cf sections 6.2.2 and 6.3.3), contrary to [16]. Eventually, in most of the cases, we noticed a decreasing of the error between the numerical and exact solutions thanks to the use of the conical scheme regardless of σ, ε and the order of the space reconstruction step (cf sections 6.2.2 et 6.3.3).

8 Annex

8.1 Upwind scheme for the transport equation

We consider the following conservative linear transport equation :

$$\partial_t f + \mathrm{div}(f\mathbf{a}) = 0,$$

In order to make the algebra clearer, the unkown $f(t,\mathbf{x})$ is a scalar function. The velocity $\mathbf{a}$ only depends on the space variable $\mathbf{x}$. The upwind scheme is defined by (cf [28]) :

$$|\Omega_j|\partial_t f_j + \sum_{R_j^+} \langle \tilde{\mathbf{C}}_j^{\mathrm{dof}}, \mathbf{a}_{\mathrm{dof}} \rangle f_j^{\mathrm{dof}} + \sum_{R_j^-} \langle \tilde{\mathbf{C}}_j^{\mathrm{dof}}, \mathbf{a}_{\mathrm{dof}} \rangle f_{k(\mathrm{dof})} = 0, \tag{94}$$

where f_j^{dof} is the value of f at degree of freedom dof in cell j computed with some arbitrary reconstruction step (which can be of order one, that is to say : $f_j^{\mathrm{dof}} = f_j$), and :

$$R_j^+ = \{\mathrm{dof}, \langle \tilde{\mathbf{C}}_j^{\mathrm{dof}}, \mathbf{a}_{\mathrm{dof}} \rangle > 0\}, \qquad R_j^- = \{\mathrm{dof}, \langle \tilde{\mathbf{C}}_j^{\mathrm{dof}}, \mathbf{a}_{\mathrm{dof}} \rangle < 0\},$$

and for a given dof :

$$f_{k(\mathrm{dof})} = \frac{1}{\sum_{I_{\mathrm{dof}}^+} \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle} \sum_{I_{\mathrm{dof}}^+} \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle f_i^{\mathrm{dof}}, \tag{95}$$

with :

$$I_{\mathrm{dof}}^+ = \{i, \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle > 0\}, \qquad I_{\mathrm{dof}}^- = \{i, \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle < 0\}.$$

Remark 8. *In a 1D framework, the scheme (94) reduces to the classical 1D upwind scheme.*

Proposition 8.1. *The scheme is conservative :*

$$\partial_t \left(\sum_{j*} |\Omega_j| f_j \right) = 0 \tag{96}$$

Proof. Denoting by $\sum_{j*}$ the sum over all the cells of the mesh, and $\sum_{\mathrm{dof}*}$ the sum over all the degrees of freedom of the mesh, on can write, up to the boundary terms :

$$\begin{aligned}
\partial_t \left(\sum_{j*} |\Omega_j| f_j \right) &= \sum_{j*} \left[\sum_{R_j^+} \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_j^{\mathrm{dof}} \rangle f_j^{\mathrm{dof}} + \sum_{R_j^-} \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_j^{\mathrm{dof}} \rangle f_{k(\mathrm{dof})} \right] \\
&= \sum_{\mathrm{dof}*} \sum_{I_{\mathrm{dof}}^+} \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle f_i^{\mathrm{dof}} + \sum_{\mathrm{dof}*} \sum_{I_{\mathrm{dof}}^-} \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle f_{k(\mathrm{dof})} \\
&= \sum_{\mathrm{dof}*} \sum_{I_{\mathrm{dof}}^+} \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle f_i^{\mathrm{dof}} + \sum_{\mathrm{dof}*} f_{k(\mathrm{dof})} \langle \mathbf{a}_{\mathrm{dof}}, \sum_{I_{\mathrm{dof}}^-} \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle.
\end{aligned}$$

Moreover, according to equation (10) :

$$\sum_{I_{\mathrm{dof}}^+} \tilde{\mathbf{C}}_i^{\mathrm{dof}} + \sum_{I_{\mathrm{dof}}^-} \tilde{\mathbf{C}}_i^{\mathrm{dof}} = \sum_i \tilde{\mathbf{C}}_i^{\mathrm{dof}} = \mathbf{0},$$

thus, thanks to (95) :

$$f_{k(\mathrm{dof})} \langle \mathbf{a}_{\mathrm{dof}}, \sum_{I_{\mathrm{dof}}^-} \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle = -f_{k(\mathrm{dof})} \langle \mathbf{a}_{\mathrm{dof}}, \sum_{I_{\mathrm{dof}}^+} \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle = - \sum_{I_{\mathrm{dof}}^+} \langle \mathbf{a}_{\mathrm{dof}}, \tilde{\mathbf{C}}_i^{\mathrm{dof}} \rangle f_i^{\mathrm{dof}}. \tag{97}$$

Therefore, summing equation (97) over all degrees of freedom leads to equation (96).

□

References

[1] Thierry Goudon and Chunjin Lin. Analysis of the m1 model: Well-posedness and diffusion asymptotics. *Journal of Mathematical Analysis and Applications*, 402(2):579–593, 2013.

[2] Bruno Dubroca and Jean-Luc Feugeas. Étude théorique et numérique d'une hiérarchie de modèles aux moments pour le transfert radiatif. *C. R. Acad. Sci. Paris Sér. I Math.*, 329(10):915–920, 1999.

[3] Jean-François Coulombel, François Golse, and Thierry Goudon. Diffusion approximation and entropy-based moment closure for kinetic equations. *Asymptot. Anal.*, 45(1-2):1–39, 2005.

[4] Thierry Goudon and Chunjin Lin. Analysis of the $M1$ model: well-posedness and diffusion asymptotics. *J. Math. Anal. Appl.*, 402(2):579–593, 2013.

[5] C. Berthon, J. Dubois, and R. Turpault. Numerical approximation of the M_1-model. In *Mathematical models and numerical methods for radiative transfer*, volume 28 of *Panor. Synthèses*, pages 55–86. Soc. Math. France, Paris, 2009.

[6] Cory D. Hauck, C. David Levermore, and André L. Tits. Convex duality and entropy-based moment closures: characterizing degenerate densities. *SIAM J. Control Optim.*, 47(4):1977–2015, 2008.

[7] S. Guisset, J. G. Moreau, R. Nuter, S. Brull, E. d'Humières, B. Dubroca, and V. T. Tikhonchuk. Limits of the M_1 and M_2 angular moments models for kinetic plasma physics studies. *J. Phys. A*, 48(33):335501, 23, 2015.

[8] Shi Jin and C.David Levermore. Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. *Journal of Computational Physics*, 126(2):449–467, 1996.

[9] Christophe Buet and Bruno Despres. Asymptotic preserving and positive schemes for radiation hydrodynamics. *J. Comput. Phys.*, 215(2):717–740, 2006.

[10] Christophe Buet and Stéphane Cordier. Asymptotic preserving scheme and numerical methods for radiative hydrodynamic models. *C. R. Math. Acad. Sci. Paris*, 338(12):951–956, 2004.

[11] Christophe Berthon, Pierre Charrier, and Bruno Dubroca. An HLLC scheme to solve the M_1 model of radiative transfer in two space dimensions. *J. Sci. Comput.*, 31(3):347–389, 2007.

[12] C. Berthon, J. Dubois, B. Dubroca, T.-H. Nguyen-Bui, and R. Turpault. A free streaming contact preserving scheme for the M_1 model. *Adv. Appl. Math. Mech.*, 2(3):259–285, 2010.

[13] Edgar Olbrant, Cory D. Hauck, and Martin Frank. A realizability-preserving discontinuous Galerkin method for the M1 model of radiative transfer. *J. Comput. Phys.*, 231(17):5612–5639, 2012.

[14] Christophe Buet, Bruno Després, and Emmanuel Franck. An asymptotic preserving scheme with the maximum principle for the M_1 model on distorded meshes. *C. R. Math. Acad. Sci. Paris*, 350(11-12):633–638, 2012.

[15] Emmanuel Franck, Christophe Buet, and Bruno Després. Asymptotic preserving finite volumes discretization for non-linear moment model on unstructured meshes. In *Finite volumes for complex applications VI. Problems & perspectives. Volume 1, 2*, volume 4 of *Springer Proc. Math.*, pages 467–474. Springer, Heidelberg, 2011.

[16] Emmanuel Franck. *Construction et analyse numérique de schema asymptotic preserving sur maillages non structurés. Application au transport linéaire et aux systèmes de Friedrichs.* PhD thesis, Université Pierre et Marie Curie - Paris VI, 2012.

[17] Bruno Després and Constant Mazeran. Lagrangian gas dynamics in two dimensions and lagrangian systems. *Archive for Rational Mechanics and Analysis*, 178(3):327–372, 2005.

[18] P.H. Maire, R. Abgrall, J. Breil, and J. Ovadia. A cell-centered lagrangian scheme for 2d compressible flow problems. *Siam. J. Sci. Comput.*, 29(4):1781–1824, 2007.

[19] Prince Chidyagwai, Martin Frank, Florian Schneider, and Benjamin Seibold. A comparative study of limiting strategies in discontinuous Galerkin schemes for the M_1 model of radiation transport. *J. Comput. Appl. Math.*, 342:399–418, 2018.

[20] Graham Alldredge and Florian Schneider. A realizability-preserving discontinuous Galerkin scheme for entropy-based moment closures for linear kinetic equations in one space dimension. *J. Comput. Phys.*, 295:665–684, 2015.

[21] C. Kristopher Garrett, Cory Hauck, and Judith Hill. Optimization and large scale computation of an entropy-based moment closure. *J. Comput. Phys.*, 302:573–590, 2015.

[22] Sébastien Guisset. Angular moments models for rarefied gas dynamics. Numerical comparisons with kinetic and Navier-Stokes equations. *Kinet. Relat. Models*, 13(4):739–758, 2020.

[23] C. Chalons and S. Guisset. An antidiffusive HLL scheme for the electronic M_1 model in the diffusion limit. *Multiscale Model. Simul.*, 16(2):991–1016, 2018.

[24] Sébastien Guisset, Stéphane Brull, Bruno Dubroca, and Rodolphe Turpault. An admissible asymptotic-preserving numerical scheme for the electronic M_1 model in the diffusive limit. *Commun. Comput. Phys.*, 24(5):1326–1354, 2018.

[25] Sébastien Guisset, Stéphane Brull, Emmanuel D'Humières, and Bruno Dubroca. Asymptotic-preserving well-balanced scheme for the electronic M_1 model in the diffusive limit: particular cases. *ESAIM Math. Model. Numer. Anal.*, 51(5):1805–1826, 2017.

[26] Martin Frank, Cory D. Hauck, and Edgar Olbrant. Perturbed, entropy-based closure for radiative transfer. *Kinet. Relat. Models*, 6(3):557–587, 2013.

[27] Xavier Blanc, Vincent Delmas, and Philippe Hoch. Asymptotic preserving schemes on conical unstructured 2d meshes. *International Journal for Numerical Methods in Fluids*, 93(8):2763–2802, 2021.

[28] Aude Bernard-Champmartin, Philippe Hoch, and Nicolas Seguin. Stabilité locale et montée en ordre pour la reconstruction de quantités volumes finis sur maillages coniques non-structurés en dimension 2. preprint, https://hal.archives-ouvertes.fr/hal-02497832, March 2020.

[29] Christophe Buet and Bruno Després. A gas dynamics scheme for a two moments model of radiative transfer. working paper or preprint, November 2008.

[30] G. Carré, S. Del Pino, B. Després, and E. Labourasse. A cell-centered Lagrangian hydrodynamics scheme on general unstructured meshes in arbitrary dimension. *J. Comput. Phys.*, 228(14):5160–5183, 2009.

[31] Laurent Gosse and Giuseppe Toscani. An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations. *Comptes Rendus Mathematique*, 334(4):337–342, 2002.

[32] Christophe Buet, Bruno Després, and Emmanuel Franck. Design of asymptotic preserving finite volume schemes for the hyperbolic heat equation on unstructured meshes. *Numer. Math.*, 122(2):227–278, 2012.

[33] John K. Dukowicz and John W. Kodis. Accurate conservative remapping (rezoning) for arbitrary lagrangian-eulerian computations. *SIAM Journal on Scientific and Statistical Computing*, 8(3):305–321, 1987.

[34] Philippe Hoch and Emmanuel Labourasse. A frame invariant and maximum principle enforcing second-order extension for cell-centered ALE schemes based on local convex hull preservation. *Internat. J. Numer. Methods Fluids*, 76(12):1043–1063, 2014.

[35] Bruno Després. *Numerical Methods for Eulerian and Lagrangian Conservation Laws.* Birkhäuser, 01 2017.

[36] Florian Blachère, Christophe Chalons, and Rodolphe Turpault. Very high-order asymptotic-preserving schemes for hyperbolic systems of conservation laws with parabolic degeneracy on unstructured meshes. *Comput. Math. Appl.*, 87:41–49, 2021.

[37] F. Blachère and R. Turpault. An admissibility and asymptotic preserving scheme for systems of conservation laws with source term on 2D unstructured meshes with high-order MOOD reconstruction. *Comput. Methods Appl. Mech. Engrg.*, 317:836–867, 2017.

[38] Christophe Le Potier. A second order in space combination of methods verifying a maximum principle for the discretization of diffusion operators. *C. R. Math. Acad. Sci. Paris*, 358(1):89–96, 2020.

[39] Jay I. Frankel, Brian Vick, and M. Necati Ozisik. Flux formulation of hyperbolic heat conduction. *Journal of Applied Physics*, 58(9):3340–3345, 1985.

[40] Juan Antonio López Molina and Macarena Trujillo. Regularity of solutions of the anisotropic hyperbolic heat equation with nonregular heat sources and homogeneous boundary conditions. *Turkish J. Math.*, 41(3):461–482, 2017.

[41] Christophe Buet and Bruno Despres. Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 85, 03 2003.

[42] B. Boutin, E. Deriaz, P. Hoch, and P. Navaro. Extension of ALE methodology to unstructured conical meshes. In *ESAIM: Proc.*, volume 32, pages 31–55, 2011.

[43] A. Bernard-Champmartin, E. Deriaz, P. Hoch, G. Samba, and M. Schaefer. Extension of centered hydrodynamical schemes to unstructured deforming conical meshes : the case of circles. In *ESAIM : proc.*, volume 38, pages 135–162, 2012.

[44] P.H. Maire and B. Nkonga. Multi-scale Godunov type method for cell-centered discrete lagrangian hydrodynamics. *J. Comput. Phys*, 2009.

[45] Xavier Roynard. Extension du schéma vofire aux maillages à bords coniques. Technical report, CEA DAM-DIF, 2013.

[46] Dimitri Mihalas and Barbara Weibel Mihalas. *Foundations of radiation hydrodynamics*. Oxford University Press, New York, 1984.