Positive composite finite volume schemes for the diffusion equation on unstructured meshes

Xavier Blanc, Philippe Hoch, Clément Lasuen

Positive composite finite volume schemes for the diffusion equation on unstructured meshes

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

[1]Université Paris Cité, 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

December 13, 2022

Abstract

We present a finite volume scheme for the anisotropic diffusion equation. This scheme is obtained as a limit of an asymptotic preserving (AP) scheme for the M_1 model of radiative transfer. The latter was designed in [BDF12] and [Fra12] on polygonal meshes and [BHL21] on conical meshes. After having presented the construction of the scheme, we show that it writes as a convex combination of two consistent terms. The first one is second order consistent and may generate instabilities on unstructured meshes. The second one is first order consistent and more stable. It can be modified so as to reach a second order consistency using a reconstruction procedure. Moreover, we prove that the explicit time discretisation of our scheme preserves the positivity of the unknown under a CFL condition. Some numerical test cases are given in order to illustrate the good properties of the scheme.

Contents

1 Introduction

In this work, we study a finite volume scheme for the anisotropic diffusion equation (1) in two space dimensions:

$$\partial_t E - \text{div } (\kappa \nabla E) = \mathcal{S}. \tag{1}$$

The unknown is denoted by E. The diffusion tensor is κ and we assume that, for any $\mathbf{x} \in \Omega$ (Ω being the domain of computation), $\kappa(\mathbf{x})$ is a symmetric positive definite 2×2 matrix. The source term $\mathcal{S}$ is non-negative and depends on time and space. In order to be consistent with [BHL21], we define $\sigma = \kappa^{-1}$ and we write (1) under the following form:

$$\partial_t E - \text{div } \left(\sigma^{-1} \nabla E\right) = \mathcal{S}. \tag{2}$$

We also assume that there exists $\sigma_2 \geq \sigma_1 > 0$ such that:

$$\forall \mathbf{x} \in \Omega,\ \text{Sp}(\sigma(\mathbf{x})) \subset [\sigma_1, \sigma_2].$$

We focus on a generalization of the scheme that was developed in [Fra12] and [FBD11] on polygonal unstructured meshes and in [BHL21] for conical meshes (that is to say, meshes whom edges are curved). This scheme was developed to solve the isotropic diffusion equation and we want to extend it to the anisotropic case. We will describe the scheme only for polygonal unstructured meshes. Moreover, inspired by the conical scheme [BHL21] and by [Hoc22], we develop a *composite* scheme. We call it *composite* because the associated fluxes are neither purely nodal nor edge-based. They are computed using quantities which are defined on the *degrees of freedom*: the nodes and the middles of edges of the mesh (point of quadrature formula). The values of the unknown are located at the barycenters of the cells.

The question of defining an accurate finite volume scheme for the diffusion equation on deformed meshes is a long-standing problem. It is well-known that a standard two-point flux is consistent only on rectangular meshes. To our knowledge, the first attempt to design a consistent scheme is that of D. Kershaw [Ker81]. This scheme was not proved to be consistent on general meshes, and numerical tests indicate that it is convergent only when cells are parallelograms. This scheme does not satisfy the maximum principle, and an attempt to make it positive was proposed in [Per81]. Apart from this scheme, the diamond scheme was analyzed in [CVV99], and proved to be consistent. In such a strategy, one uses node values as auxiliary unknowns, allowing to compute consistent fluxes. These auxiliary unknowns are computed using interpolation. It is also possible to use a mixed finite element approach [RT83] and recast it as a finite volume method (see [AWY97]). Such a scheme is consistent, but not positive. Another strategy, called DDFV (Discrete Duality Finite Volume) was proposed by F. Hermeline in [Her98, Her00, Her03, Her07]. In this strategy, instead of computing the auxiliary (nodal) unknowns by interpolation, they are defined as a solution to a diffusion problem on a dual mesh. Several other methods were proposed, such as the mimetic finite difference method (see, for instance, [BBL09, LMS14]), or the SUSHI (Scheme Using Stabilization and Harmonic Interfaces) method, by R. Eymard, T. Gallouët and R. Herbin [Eym10]. In a second step, the mimetic finite difference method was extended as the virtual element method (VEM) [YSGN22, BaDVBC$^+$13]. Let us also mention the MPFA (Multi Point Flux Approximation) method proposed in [AEK$^+$07, BM07].

All the above schemes are convergent, but are not positive. This may be an important issue in applications, since the unknown may be a temperature or a concentration. A truncation strategy is in principle possible, but it breaks the conservation property of the scheme, which is also highly desirable. To address this problem, several strategies have been proposed. Most of them consist in using different consistent estimations of the fluxes and in combining them so that the matrix of the scheme becomes an M-matrix, thereby recovering positivity. Such a strategy was initially proposed in [BM05] and [LP09]. It makes the scheme nonlinear, even though the considered equation (1) is linear. Following these works, many similar strategies have been proposed. Let us cite [LMS14, YSGN22, SY16, SYY09, AN21, WPL$^+$22, NSL22, BL16], among others. Of course, we do not claim this list to be exhaustive.

In the present work, we propose a family of schemes that are naturally consistent, conservative and positive. Contrary to the above mentioned works, positivity is not enforced by modifying an existing non-positive

scheme. The starting point of our approach are schemes defined in [FBD11, BHL21] for the M1 model [DF99], which is a hyperbolic nonlinear model that satisfies a positivity principle and a diffusion limit. Such a scheme is positive by construction, and we study its diffusion limit, which is of course positive.

The present article is organised as follows. In Section 2, we present the semi-discrete version of our scheme. We first recall the M_1 limit diffusion scheme from [BHL21] in the case of conical degenerate (that is to say, polygonal) meshes. We generalize it and we write a two-parameter family of consistent schemes that are valid on unstructured polygonal meshes. Then we extend it to the anisotropic diffusion case. We also present a modification of this scheme so as to reach second order convergence. Then we focus on the time discretised version of the scheme. We prove that the explicit version is positivity-preserving on general unstructured meshes under a CFL condition that is presented in Section 3. Moreover, in Section 4 we study a particular case which has a less restrictive CFL condition. The last section is devoted to numerical test cases.

2 Numerical method

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

2.1 Composite normal vectors set on straight unstructured meshes and properties

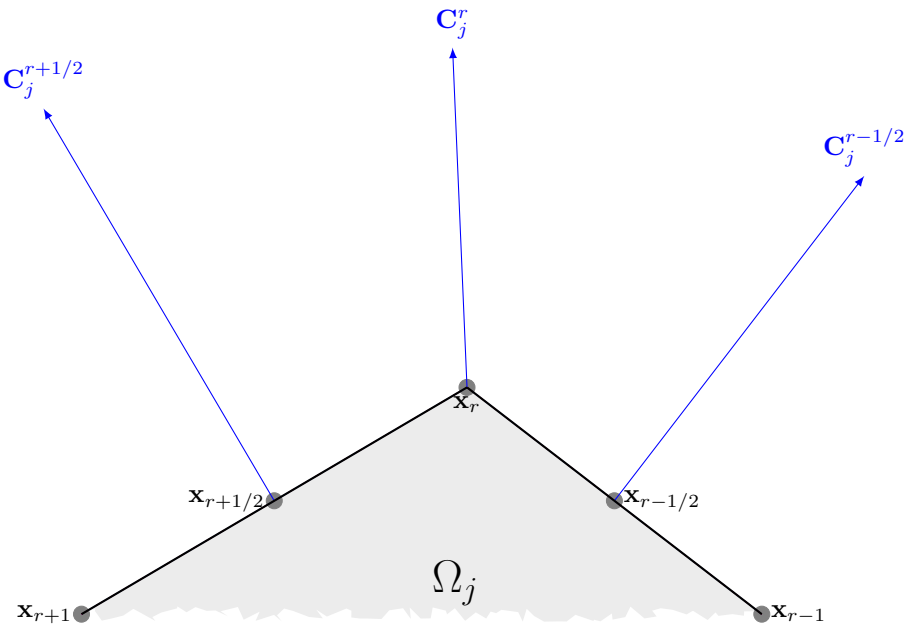

Figure 1: Normals at nodes, at edges : composite set

Let Ω_j be a cell of the mesh $\mathcal{T}$ paving the domain Ω. Let $\mathbf{x}_{r-1}$, $\mathbf{x}_r$ and $\mathbf{x}_{r+1}$ be 3 consecutive nodes of Ω_j. We define:

- the middle of the edge $[\mathbf{x}_r, \mathbf{x}_{r+1}]$: $\mathbf{x}_{r+1/2} = (\mathbf{x}_r + \mathbf{x}_{r+1})/2$,
- the normal to the edge $[\mathbf{x}_r, \mathbf{x}_{r+1}]$: $\mathbf{C}_j^{r+1/2} = (\mathbf{x}_{r+1} - \mathbf{x}_r)^\perp$,
- the normal to the node r:

$$\mathbf{C}_j^r = \frac{1}{2}(\mathbf{x}_{r+1} - \mathbf{x}_{r-1})^\perp = \frac{1}{2}\left(\mathbf{C}_j^{r+1/2} + \mathbf{C}_j^{r-1/2}\right), \tag{3}$$

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

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \qquad \boldsymbol{\xi}^\perp = \begin{pmatrix} -\xi_2 \\ \xi_1 \end{pmatrix}.$$

We present here some notations that will be used in the remainder of the paper. We define a *degree of freedom* as either a node or a middle of an edge. We also define:

- $(\mathbf{x}_r)_r$ the coordinates of the vertices of the cell j;
- $(\mathbf{x}_{r+1/2})_{r+1/2}$ the coordinates of the mid-edge points of the cell j;
- $\sum_{r\in\Omega_j} g_j^r$ the 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);
- $\sum_{r+1/2\in\Omega_j} g_j^{r+1/2}$ the sum over all the mid-edge points of the cell j of the quantity g;
- $N_{\text{dof}} = \sum_{i|\text{dof}\in\Omega_i} 1$ the number of cells that contains the given degree of freedom dof;
- $\sum_{i|\text{dof}\in\Omega_i} g_i^{\text{dof}}$ the sum, for a given degree of freedom, over all the cells that contains this degree of freedom;
- $\sum_{j\in\mathcal{T}} g_j$ the sum over all the cells of the mesh;
- $\sum_{r\in\mathcal{T}} g^r$ the sum over all the nodes of the mesh;
- $\sum_{r+1/2\in\mathcal{T}} g^{r+1/2}$ the sum over all the mid-edge points of the mesh;
- h the maximum length of edges of the mesh,
- $\langle\cdot,\cdot\rangle$ the inner product in $\mathbb{R}^2$.

Moreover, for any mid-edge point $r+1/2$, we denote by j and k the two cells that are separated by the edge containing $\mathbf{x}_{r+1/2}$, see Figure 2.

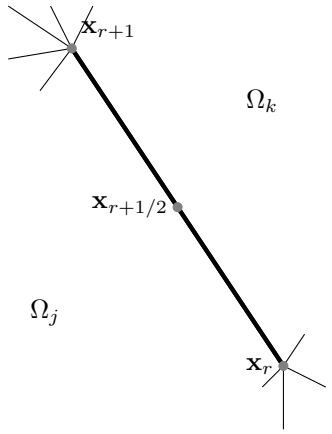

Figure 2: The neighbouring cells of mid-edge point $\mathbf{x}_{r+1/2}$

We have the following identity for any $\theta \in [0,1]$:

$$|\Omega_j| = \frac{1-\theta}{2} \sum_{r\in\Omega_j} \left\langle \mathbf{C}_j^r, \mathbf{x}_r - \mathbf{x}_j \right\rangle + \frac{\theta}{2} \sum_{r+1/2\in\Omega_j} \left\langle \mathbf{C}_j^{r+1/2}, \mathbf{x}_{r+1/2} - \mathbf{x}_j \right\rangle, \tag{4}$$

which is a consequence of Theorem 2.1 below. We also have the following:

- for any cell j:

$$\sum_{r\in\Omega_j} \mathbf{C}_j^r = \sum_{r+1/2\in\Omega_j} \mathbf{C}_j^{r+1/2} = \mathbf{0}, \tag{5}$$

- for any inner node r and any inner edge $r+1/2$:

$$\sum_{i|r\in\Omega_i} \mathbf{C}_i^r = \sum_{i|r+1/2\in\Omega_i} \mathbf{C}_i^{r+1/2} = \mathbf{0}. \tag{6}$$

Theorem 2.1. *Let $g \in \mathcal{C}^2(\mathbb{R}^2;\mathbb{R})$. Then, for all $\theta \in [0,1]$:*

$$\frac{1}{|\Omega_j|}\int_{\partial\Omega_j} g\mathbf{n} = \frac{1}{|\Omega_j|}\left[(1-\theta)\sum_{r\in\Omega_j} g(\mathbf{x}_r)\mathbf{C}_j^r + \theta \sum_{r+1/2\in\Omega_j} g(\mathbf{x}_{r+1/2})\mathbf{C}_j^{r+1/2}\right] + \mathcal{O}(h). \tag{7}$$

Moreover, the remainder in (7) vanishes when g is an affine function.

2.2 Isotropic M_1 diffusion limit scheme

Our scheme is based on the limit scheme for the M_1 model of the isotropic radiative transfer [BHL21] on a conical degenerate mesh (when the curvature of each edge vanishes). The M_1 model is a moment model for the radiative transfer equation. It depends on a small parameter $\varepsilon > 0$ which inverse accounts for the optical thickness of the medium. When ε tends to 0 (*ie* the medium is highly opaque), the model converges toward (2) and the scheme designed in [BHL21] converges toward the scheme (8), which is consistent with (2) (see Section 2.4 for the explanations and [BHL21] for some numerical examples). In this case, the diffusion coefficient σ is a positive scalar constant. The scheme writes:

$$|\Omega_j|\frac{d}{dt}E_j + \frac{3}{4}\left[\left(1-\frac{\pi}{4}\right)\sum_{r\in\Omega_j}\left\langle \mathbf{u}_r, \mathbf{C}_j^r\right\rangle E_j^r + \frac{\pi}{4}\sum_{r+1/2\in\Omega_j}\left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2}\right\rangle E_j^{r+1/2}\right] \tag{8}$$

$$+\frac{1}{4}\left[\left(1-\frac{\pi}{4}\right)\sum_{r\in\Omega_j}\left\langle E_j\mathbf{C}_j^r - E_r\beta_j^r\sigma\mathbf{u}_r, \mathbf{u}_r\right\rangle + \frac{\pi}{4}\sum_{r+1/2\in\Omega_j}\left\langle E_j\mathbf{C}_j^{r+1/2} - E_{r+1/2}\beta_j^{r+1/2}\sigma\mathbf{u}_{r+1/2}, \mathbf{u}_{r+1/2}\right\rangle\right]$$

$$= |\Omega_j|\mathcal{S}_j,$$

with:

$$E_{\text{dof}} = \frac{1}{N_{\text{dof}}}\sum_{i|\text{dof}\in\Omega_i} E_i, \tag{9}$$

$$E_j^{\text{dof}} = \begin{cases} E_j & \text{if } \langle \mathbf{u}_{\text{dof}}, \mathbf{C}_j^{\text{dof}}\rangle > 0, \\ \dfrac{1}{\sum_{i\in I_{\text{dof}}^+}\langle \mathbf{u}_{\text{dof}}, \mathbf{C}_i^{\text{dof}}\rangle}\displaystyle\sum_{i\in I_{\text{dof}}^+}\langle \mathbf{u}_{\text{dof}}, \mathbf{C}_i^{\text{dof}}\rangle E_i & \text{else,} \end{cases} \qquad I_{\text{dof}}^+ = \left\{i, \left\langle \mathbf{u}_{\text{dof}}, \mathbf{C}_i^{\text{dof}}\right\rangle > 0\right\}, \tag{10}$$

For any node r, $\mathbf{u}_r$ is defined by:

$$\beta_r\mathbf{u}_r = \frac{1}{\sigma E_r}\sum_{i|r\in\Omega_i} E_i\mathbf{C}_i^r, \tag{11}$$

with:

$$\beta_{\text{dof}} = \sum_{i|\text{dof}\in\Omega_i}\beta_i^{\text{dof}}, \qquad \beta_i^{\text{dof}} = \mathbf{C}_i^{\text{dof}}\otimes(\mathbf{x}_{\text{dof}} - \mathbf{x}_i). \tag{12}$$

and for any mid-edge point $r+1/2$:

$$\begin{cases} \langle \mathbf{u}_{r+1/2}, \mathbf{x}_k - \mathbf{x}_j \rangle = \dfrac{E_j - E_k}{\sigma E_{r+1/2}}, \\ \langle \mathbf{u}_{r+1/2}, (\mathbf{x}_k - \mathbf{x}_j)^\perp \rangle = \dfrac{1}{2} \langle \mathbf{u}_r + \mathbf{u}_{r+1}, (\mathbf{x}_k - \mathbf{x}_j)^\perp \rangle. \end{cases}$$

Equation (8) is not exactly identical to the scheme from [BHL21], we simplified some numerical coefficients (in particular, the diffusion coefficient wrote as $1/(3\sigma)$ and not as $1/\sigma$). As explained in [BHL21], the quantity $\mathbf{u}_{\text{dof}}$ is computed so as to make the scheme conservative. To this end, we impose the following relation around each dof:

$$\sum_{i|\text{dof}\in\Omega_i} E_i \mathbf{C}_i^{\text{dof}} - E_{\text{dof}} \beta_i^{\text{dof}} \sigma \mathbf{u}_{\text{dof}} = \mathbf{0}. \tag{13}$$

Equation (13) can be written as:

$$\beta_{\text{dof}} \mathbf{u}_{\text{dof}} = \frac{1}{\sigma E_{\text{dof}}} \sum_{i|\text{dof}\in\Omega_i} E_i \mathbf{C}_i^{\text{dof}}. \tag{14}$$

For a node r, Equation (14) reads as (11) and the matrix β_r is invertible (under some assumptions on the mesh, see Section 3.1). Thus $\mathbf{u}_r$ is well defined. However, for a midpoint $r+1/2$, the matrix $\beta_{r+1/2}$ is not invertible. Indeed, denoting by j and k the two cells that are separated by the edge containing $r+1/2$ (see Figure 2) and using (6), we have:

$$\mathbf{C}_j^{r+1/2} + \mathbf{C}_k^{r+1/2} = \mathbf{0}.$$

This leads to:

$$\beta_{r+1/2} = \mathbf{C}_j^{r+1/2} \otimes (\mathbf{x}_k - \mathbf{x}_j). \tag{15}$$

Therefore this matrix has rank 1 and it is not invertible. Using (15), Equation (14) can be simplified:

$$\langle \mathbf{u}_{r+1/2}, \mathbf{x}_k - \mathbf{x}_j \rangle = \frac{E_j - E_k}{\sigma E_{r+1/2}}. \tag{16}$$

Thus we see that $\mathbf{u}_{r+1/2}$ is only defined in one direction. In [BDH21] and [BHL21], the following formula is proposed so as to compute $\mathbf{u}_{r+1/2}$ in the orthogonal direction:

$$\left\langle \mathbf{u}_{r+1/2}, (\mathbf{x}_k - \mathbf{x}_j)^\perp \right\rangle = \frac{1}{2} \left\langle \mathbf{u}_r + \mathbf{u}_{r+1}, (\mathbf{x}_k - \mathbf{x}_j)^\perp \right\rangle. \tag{17}$$

Eventually, as explained in [BHL21], the explicit time discretisation of (8) preserves the positivity of the solution if, at any iteration n, the time step Δt satisfies:

$$\Delta t \leq C\sigma h^2 \min_{j\in\mathcal{T}} \left\{ \frac{E_j^n}{\sum_{i\in\mathcal{V}_j} E_i^n} \right\}.$$

2.3 A two-parameter family of consistent numerical schemes for the diffusion equation on unstructured polygonal meshes

The scheme (8) is a composite scheme as it uses unknowns located at the nodes and the mid-edges with a ponderation of $\pi/4$. As in [Fra12], we notice that it writes as a convex combination of pure advection term and a source term. The consistency of each term is proved in Section 2.4. We can generalize it for any $(\eta, \theta) \in [0,1]^2$ and obtain a two-parameter family of numerical schemes:

$$|\Omega_j| \frac{d}{dt} E_j + (1-\eta) \left[(1-\theta) \sum_{r\in\Omega_j} \left\langle \mathbf{u}_r, \mathbf{C}_j^r \right\rangle E_j^r + \theta \sum_{r+1/2\in\Omega_j} \left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2} \right\rangle E_j^{r+1/2} \right] \tag{18}$$

$$+\eta\left[(1-\theta)\sum_{r\in\Omega_j}\left\langle E_j\mathbf{C}_j^r-E_r\beta_j^r\sigma\mathbf{u}_r,\mathbf{u}_r\right\rangle+\theta\sum_{r+1/2\in\Omega_j}\left\langle E_j\mathbf{C}_j^{r+1/2}-E_{r+1/2}\beta_j^{r+1/2}\sigma\mathbf{u}_{r+1/2},\mathbf{u}_{r+1/2}\right\rangle\right]$$

$$=|\Omega_j|\mathcal{S}_j.$$

Remark 1. *We can notice the following particular cases:*

- $\theta=0$ *is the nodal scheme : it uses only the unknowns located at the nodes of the mesh,*
- $\theta=1$ *is not a purely edged-based scheme : it uses only the unknowns located at the edges of the mesh in order to compute the evolution of* E *but the quantities* $\mathbf{u}_r$ *are used as intermediate unknowns and appear in the computation of the tangential part of* $\mathbf{u}_{r+1/2}$.

Remark 2. *This decomposition of the scheme as the sum of a node-based contribution and an edge-based contribution is inspired from [Hoc22]. In this work, the author adapts some classical schemes (VFFC, Rusanov, Roe) to composite fluxes and generates new one-parameter families of numerical schemes.*

Remark 3. *The methodology we have developped here consists in starting from a model (here the* M_1 *model) that satisfies the diffusion limit, discretizing it on conical meshes, then take the particular case of polygonal mesh in the diffusion limit. This gives a two-paramter family of diffusion schemes. We could apply it to other models, such as the* P_1 *model. Using the paper [BDH21] instead of [BHL21], we would end up with a one-parameter family of diffusion schemes that are second-order consistent, but not positive.*

2.4 Interpretation of the scheme and consistency of the fluxes

In this section we give formal arguments indicating that the scheme (18) is consistent with the diffusion equation (2) for any $(\theta,\eta)\in[0,1]^2$. First we show that $\mathbf{u}_{\rm dof}$ is consistent with $-(\nabla E)_{\rm dof}/(\sigma E_{\rm dof})$.

Lemma 2.2. *The quantity* $\mathbf{u}_{dof}$ *is consistent with* $-(\nabla E)_{dof}/(\sigma E_{dof})$.

Proof. We use arguments from [BHL21]. We have:

$$E(\mathbf{x}_i)=E(\mathbf{x}_{\rm dof})+\langle\mathbf{x}_i-\mathbf{x}_{\rm dof},\nabla E(\mathbf{x}_{\rm dof})\rangle+O(h^2). \tag{19}$$

Multiplying (19) by $\mathbf{C}_i^{\rm dof}$ and summing the result over the cells around any inner *dof* leads to:

$$\sum_{i|{\rm dof}\in\Omega_i}E(\mathbf{x}_i)\mathbf{C}_i^{\rm dof}=E(\mathbf{x}_{\rm dof})\underbrace{\sum_{i|{\rm dof}\in\Omega_i}\mathbf{C}_i^{\rm dof}}_{=\mathbf{0}}-\beta_{\rm dof}\nabla E(\mathbf{x}_{\rm dof})+O(h^3), \tag{20}$$

where $\beta_{\rm dof}$ is defined by (12). Since the nodal matrix β_r is invertible (see [Fra12] for further details), we have, for any inner node r:

$$\frac{1}{\sigma E(\mathbf{x}_r)}\beta_r^{-1}\left(\sum_{i|r\in\Omega_i}E(\mathbf{x}_i)\mathbf{C}_i^r\right)=\frac{-1}{\sigma E(\mathbf{x}_r)}(\nabla E)(\mathbf{x}_r)+O(h).$$

Using (11), this proves that $\mathbf{u}_r$ is consistent with $-(\nabla E)_r/(\sigma E_r)$.
Besides, using arguments from Section 2.2, for any inner mid-point $r+1/2$, Equation (20) can be written as:

$$\left\langle\nabla E(\mathbf{x}_{r+1/2}),\mathbf{x}_k-\mathbf{x}_j\right\rangle=E(\mathbf{x}_k)-E(\mathbf{x}_j)+O(h^2).$$

This leads to:

$$\left\langle\frac{1}{\sigma E(\mathbf{x}_{r+1/2})}\nabla E(\mathbf{x}_{r+1/2}),\mathbf{x}_k-\mathbf{x}_j\right\rangle=-\frac{E(\mathbf{x}_j)-E(\mathbf{x}_k)}{\sigma E(\mathbf{x}_{r+1/2})}+O(h^2). \tag{21}$$

Moreover, one has:

$$\left\langle \frac{1}{\sigma E(\mathbf{x}_{r+1/2})} \nabla E(\mathbf{x}_{r+1/2}), (\mathbf{x}_k - \mathbf{x}_j)^\perp \right\rangle \tag{22}$$

$$= \frac{1}{2} \left\langle \frac{1}{\sigma E(\mathbf{x}_r)} \nabla E(\mathbf{x}_r) + \frac{1}{\sigma E(\mathbf{x}_{r+1})} \nabla E(\mathbf{x}_{r+1}), (\mathbf{x}_k - \mathbf{x}_j)^\perp \right\rangle + O(h^2).$$

Equations (16) (17) (21) and (22) prove that $\mathbf{u}_{r+1/2}$ is consistent with $-(\nabla E)_{r+1/2}/(\sigma E_{r+1/2})$

□

The scheme (18) is a convex combination of two terms within square brackets that can be interpreted as follows. The first one:

$$(1-\theta) \sum_{r \in \Omega_j} \left\langle \mathbf{u}_r, \mathbf{C}_j^r \right\rangle E_j^r + \theta \sum_{r+1/2 \in \Omega_j} \left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2} \right\rangle E_j^{r+1/2} \tag{23}$$

is an advection term. It corresponds to the discretisation of div $(E\mathbf{u})$ using an upwind scheme (see [BCHS20] and [BHL21] for further details). This scheme is consistent of order 1. Using Lemma 2.2, $\mathbf{u}$ is consistent with $-\nabla E/(\sigma E)$. This gives that (23) is an approximation of $-\text{div } \left(\sigma^{-1} \nabla E\right)$ that is first order consistent.

The second term within brackets in (8) reads as:

$$(1-\theta) \sum_{r \in \Omega_j} \left\langle E_j \mathbf{C}_j^r - E_r \beta_j^r \sigma \mathbf{u}_r, \mathbf{u}_r \right\rangle + \theta \sum_{r+1/2 \in \Omega_j} \left\langle E_j \mathbf{C}_j^{r+1/2} - E_{r+1/2} \beta_j^{r+1/2} \sigma \mathbf{u}_{r+1/2}, \mathbf{u}_{r+1/2} \right\rangle. \tag{24}$$

Using Theorem 2.1, it corresponds to writing:

$$\int_{\Omega_j} \text{div } (E\mathbf{u}) \approx (1-\theta) \sum_{r \in \Omega_j} \bar{E}_j^r \left\langle \mathbf{u}_r, \mathbf{C}_j^r \right\rangle + \theta \sum_{r+1/2 \in \Omega_j} \bar{E}_j^{r+1/2} \left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2} \right\rangle,$$

with:

$$\bar{E}_j^{\text{dof}} = E_j - \langle \mathbf{x}_{\text{dof}} - \mathbf{x}_j, E_{\text{dof}} \sigma \mathbf{u}_{\text{dof}} \rangle.$$

Since

$$E(\mathbf{x}_{\text{dof}}) = E(\mathbf{x}_j) + \langle \mathbf{x}_{\text{dof}} - \mathbf{x}_j, \nabla E(\mathbf{x}_{\text{dof}}) \rangle + O(h^2), \tag{25}$$

we infer that $\bar{E}_j^{\text{dof}}$ is a second order approximation of $E(\mathbf{x}_{\text{dof}})$. Multiplying by $\mathbf{C}_j^{\text{dof}}$ and using (12) gives:

$$\bar{E}_j^{\text{dof}} \mathbf{C}_j^{\text{dof}} = E_j \mathbf{C}_j^{\text{dof}} - E_{\text{dof}} \beta_j^{\text{dof}} \sigma \mathbf{u}_{\text{dof}}.$$

Therefore (24) is an approximation of $-\text{div } \left(\sigma^{-1} \nabla E\right)$ that is second order consistent.

2.5 Extension to the anisotropic case

In this section, we generalize the scheme (18) to the case of an anisotropic diffusion coefficient, that is to say when σ is a positive definite matrix that depends on the space coordinates. Thus we define $\sigma_{\text{dof}} = \sigma(\mathbf{x}_{\text{dof}})$ at any dof. The extension is straightforward and reads as:

$$|\Omega_j| \frac{d}{dt} E_j + (1-\eta) \left[(1-\theta) \sum_{r \in \Omega_j} \left\langle \mathbf{u}_r, \mathbf{C}_j^r \right\rangle E_j^r + \theta \sum_{r+1/2 \in \Omega_j} \left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2} \right\rangle E_j^{r+1/2} \right] \tag{26}$$

$$+\eta \left[(1-\theta) \sum_{r \in \Omega_j} \left\langle E_j \mathbf{C}_j^r - E_r \beta_j^r \sigma_r \mathbf{u}_r, \mathbf{u}_r \right\rangle + \theta \sum_{r+1/2 \in \Omega_j} \left\langle E_j \mathbf{C}_j^{r+1/2} - E_{r+1/2} \beta_j^{r+1/2} \sigma_{r+1/2} \mathbf{u}_{r+1/2}, \mathbf{u}_{r+1/2} \right\rangle \right]$$

$$= |\Omega_j| \mathcal{S}_j,$$

where:

$$\beta_{\text{dof}} \sigma_{\text{dof}} \ \mathbf{u}_{\text{dof}} = \frac{1}{E_{\text{dof}}} \sum_{i | \text{dof} \in \Omega_i} E_i \mathbf{C}_i^{\text{dof}}. \tag{27}$$

This writes, for any node r:

$$\beta_r \sigma_r \ \mathbf{u}_r = \frac{1}{E_r} \sum_{i | r \in \Omega_i} E_i \mathbf{C}_i^r, \tag{28}$$

and for any mid-edge point $r + 1/2$:

$$\langle \mathbf{u}_{r+1/2}, \sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j) \rangle = \frac{E_j - E_k}{E_{r+1/2}}. \tag{29}$$

A natural way of completing the definition of $\mathbf{u}_{r+1/2}$ then writes:

$$\left\langle \mathbf{u}_{r+1/2}, [\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)]^\perp \right\rangle = \frac{1}{2} \left\langle \mathbf{u}_r + \mathbf{u}_{r+1}, [\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)]^\perp \right\rangle. \tag{30}$$

However, this method is not optimal as it makes the scheme unstable. Collecting (29) and (30) we would end up with:

$$\begin{cases} \left\langle \mathbf{u}_{r+1/2}, \sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j) \right\rangle = \dfrac{E_j - E_k}{E_{r+1/2}}, \\ \left\langle \mathbf{u}_{r+1/2}, [\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)]^\perp \right\rangle = \dfrac{1}{2} \left\langle \mathbf{u}_r + \mathbf{u}_{r+1}, [\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)]^\perp \right\rangle. \end{cases} \tag{31}$$

We propose another way of computing $\mathbf{u}_{r+1/2}$ which is more stable. Our idea is to compute first $(\sigma \mathbf{u})_{r+1/2}$ and then deduce $\mathbf{u}_{r+1/2}$:

$$\begin{cases} \left\langle \sigma_{r+1/2} \mathbf{u}_{r+1/2}, \ \mathbf{x}_k - \mathbf{x}_j \right\rangle = \dfrac{E_j - E_k}{E_{r+1/2}}, \\ \left\langle \sigma_{r+1/2} \mathbf{u}_{r+1/2}, \ (\mathbf{x}_k - \mathbf{x}_j)^\perp \right\rangle = \dfrac{1}{2} \left\langle \sigma_r \mathbf{u}_r + \sigma_{r+1} \mathbf{u}_{r+1}, (\mathbf{x}_k - \mathbf{x}_j)^\perp \right\rangle. \end{cases} \tag{32}$$

The first line of system (32) is the same as (29) since $\sigma_{r+1/2}$ is symmetric, thus the scheme still remains locally conservative (*ie* Equation (27) is still satisfied). We will see from a theoretical point of view in Section 3 and from a practical point of view in Section 5 that this choice leads to a much more stable scheme. We can summarize (31) and (32) as:

$$\begin{cases} \left\langle P_{r+1/2} \mathbf{u}_{r+1/2}, \ P_{r+1/2}^{-1} \sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j) \right\rangle = \dfrac{E_j - E_k}{E_{r+1/2}}, \\ \left\langle P_{r+1/2} \mathbf{u}_{r+1/2}, \ \left[P_{r+1/2}^{-1} \sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j) \right]^\perp \right\rangle = \dfrac{1}{2} \left\langle P_r \mathbf{u}_r + P_{r+1} \mathbf{u}_{r+1}, \left[P_{r+1/2}^{-1} \sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j) \right]^\perp \right\rangle. \end{cases} \tag{33}$$

The matrix $P_{r+1/2}$ has to be symmetric and invertible. Choosing $P_{r+1/2} = I_2$ leads to (31) while choosing $P_{r+1/2} = \sigma_{r+1/2}$ leads to (32).

Remark 4. *Using the same arguments as in Section 2.4, the previous scheme is consistent for both choices $P_{dof} = I_2$ and $P_{dof} = \sigma_{dof}$.*

Remark 5. *If σ is isotropic and constant, ie $\sigma_{dof} = \bar{\sigma} I_2$ for some constant $\bar{\sigma} > 0$, then Systems (31), (32) and (33) are equivalent.*

Remark 6. *We can notice the following particular cases:*

- $(\eta, \theta) = (1/4, \pi/4)$ *is an extension of the conical degenerate scheme (8) from [BHL21] to an anisotropic diffusion coefficient,*
- $(\eta, \theta) = (1/4, 0)$ *is an extension of the nodal scheme from [Fra12] and [FBD11] to an anisotropic diffusion coefficient,*
- $\eta = 0$ *is studied below in Section 4.*

2.6 Second order reconstruction

Following the ideas of [BHL21], we briefly recall a reconstruction procedure so as to make our scheme second order accurate in space. We only modify the computation of the advection terms. We approximate the unknown in each cell using an affine function:

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

The exponent 1 stands for the degree of the approximation polynomial. Then the gradient of E is limited so as to ensure: $P_j^1(\mathbf{x}_{\text{dof}}) \geq 0$ for any *dof* of cell j, and we write :

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

where $\alpha_{j,E}$ is a scalar limiter (see [DK87]). The scheme now reads as:

$$|\Omega_j| \frac{d}{dt} E_j + (1-\eta) \left[(1-\theta) \sum_{r \in \Omega_j} \left\langle \mathbf{u}_r, \mathbf{C}_j^r \right\rangle \tilde{E}_j^r + \theta \sum_{r+1/2 \in \Omega_j} \left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2} \right\rangle \tilde{E}_j^{r+1/2} \right] \tag{34}$$

$$+\eta \left[(1-\theta) \sum_{r \in \Omega_j} \left\langle E_j \mathbf{C}_j^r - E_r \beta_j^r \sigma_r \mathbf{u}_r, \mathbf{u}_r \right\rangle + \theta \sum_{r+1/2 \in \Omega_j} \left\langle E_j \mathbf{C}_j^{r+1/2} - E_{r+1/2} \beta_j^{r+1/2} \sigma_{r+1/2} \mathbf{u}_{r+1/2}, \mathbf{u}_{r+1/2} \right\rangle \right]$$

$$= |\Omega_j| \mathcal{S}_j,$$

with:

$$\tilde{E}_j^{\text{dof}} = \begin{cases} P_j^1(\mathbf{x}_{\text{dof}}) & \text{if } \langle \mathbf{u}_{\text{dof}}, \mathbf{C}_j^{\text{dof}} \rangle > 0, \\ \dfrac{1}{\sum_{i \in I_{\text{dof}}^+} \langle \mathbf{u}_{\text{dof}}, \mathbf{C}_i^{\text{dof}} \rangle} \displaystyle\sum_{i \in I_{\text{dof}}^+} \langle \mathbf{u}_{\text{dof}}, \mathbf{C}_i^{\text{dof}} \rangle P_i^1(\mathbf{x}_{\text{dof}}) & \text{else.} \end{cases}$$

The computations of $\mathbf{u}_{\text{dof}}$ and E_{dof} are unchanged. The second term within brackets in (34) already being of order 2, thus (34) is second order consistent

2.7 Boundary conditions

The boundary conditions are imposed using the method described in [BHL21].

2.7.1 Periodic boundary conditions

In the case of periodic boundary conditions, we add some *ghost* cells on the outside of the mesh so as to make it periodic. We then define the unknown E on these new cells so as to make it periodic and we use this new geometric data to compute the β_j^{dof} on the boundary of the domain.

2.7.2 Dirichlet boundary condition

We implement Dirichlet boundary conditions as follows. Let dof be a degree of freedom where the solution is imposed at E_{boundary}. Then $E_{\text{dof}} = E_{\text{boundary}}$ and $\mathbf{u}_{\text{dof}}$ is given by:

$$\beta_r \sigma_r \ \mathbf{u}_r = \frac{1}{E_{\text{boundary}}} \sum_{i|r\in\Omega_i} (E_i - E_{\text{boundary}}) \mathbf{C}_i^r,$$

$$\begin{cases} \left\langle P_{r+1/2}\mathbf{u}_{r+1/2},\ P_{r+1/2}^{-1}\sigma_{r+1/2}(\mathbf{x}_{r+1/2} - \mathbf{x}_j) \right\rangle = \dfrac{E_j - E_{\text{boundary}}}{E_{\text{boundary}}}, \\ \left\langle P_{r+1/2}\mathbf{u}_{r+1/2},\ \left[P_{r+1/2}^{-1}\sigma_{r+1/2}(\mathbf{x}_{r+1/2} - \mathbf{x}_j)\right]^{\perp} \right\rangle = \dfrac{1}{2} \left\langle P_r\mathbf{u}_r + P_{r+1}\mathbf{u}_{r+1}, \left[P_{r+1/2}^{-1}\sigma_{r+1/2}(\mathbf{x}_{r+1/2} - \mathbf{x}_j)\right]^{\perp} \right\rangle, \end{cases}$$

and:

$$E_j^{\text{dof}} = \begin{cases} E_j & \text{if } \langle \mathbf{u}_{\text{dof}}, \mathbf{C}_j^{\text{dof}} \rangle > 0, \\ E_{\text{boundary}} & \text{else.} \end{cases}$$

3 Theoretical study of the scheme

In this section, we study the properties of the scheme (26). We first prove that it is conservative. Then we focus on the stability of the explicit scheme. We give a sufficient condition on the time step Δt so as to ensure the positivity of the unknown E at each iteration. We assume periodic boundary conditions.
In the following, the constant C is independent from the the characteristic length h of the mesh (defined in Section 3.1), from σ and the unknown E.

3.1 Assumptions on the mesh

We present here the assumptions on the regularity of the mesh. We denote by h the maximal length of the edges of the mesh ($h = \Delta x$ for a cartesian mesh). We assume that there exists a constant C_1 such that, for any dof and any cell j:

$$\frac{1}{C_1}h^2 \leq |\Omega_j| \leq C_1 h^2, \qquad \frac{1}{C_1}h \leq \|\mathbf{C}_j^{\text{dof}}\| \leq C_1 h, \qquad N_{\text{dof}} \leq C_1, \tag{35}$$

$$\forall \xi \in \mathbb{R}^2,\ \langle \beta_r \xi, \xi \rangle \geq \frac{1}{C_1} h^2 \|\xi\|^2, \tag{36}$$

and thus we have: $\|\beta_r^{-1}\| \leq C h^{-2}$. In addition, we assume that, for any cell j and any dof of j:

$$\frac{1}{C_1}h \leq \|\mathbf{x}_{\text{dof}} - \mathbf{x}_j\| \leq C_1 h, \tag{37}$$

and for any neighbouring cells i and j:

$$\frac{1}{C_1}h \leq \|\mathbf{x}_i - \mathbf{x}_j\| \leq C_1 h. \tag{38}$$

Proposition 3.1. *When the source term vanishes, the scheme (34) is conservative:*

$$\frac{d}{dt}\left(\sum_{j\in\mathcal{T}} |\Omega_j| E_j\right) = 0.$$

Proof. Using the definition of $\mathbf{u}_r$ (28) and $\mathbf{u}_{r+1/2}$ (33) the following properties are satisfied:

$$\sum_{j\in\mathcal{T}}\sum_{r\in\Omega_j}\left\langle E_j\mathbf{C}_j^r - E_r\beta_j^r\sigma_r\mathbf{u}_r\right\rangle = \sum_{r\in\mathcal{T}}\sum_{i|r\in\Omega_i}\left\langle E_i\mathbf{C}_i^r - E_r\beta_i^r\sigma_r\mathbf{u}_r\right\rangle = 0,$$

and:

$$\begin{aligned}&\sum_{j\in\mathcal{T}}\sum_{r+1/2\in\Omega_j}\left\langle E_j\mathbf{C}_j^{r+1/2} - E_{r+1/2}\beta_j^{r+1/2}\sigma_{r+1/2}\mathbf{u}_{r+1/2}\right\rangle\\ =&\sum_{r+1/2\in\mathcal{T}}\sum_{i|r+1/2\in\Omega_i}\left\langle E_i\mathbf{C}_i^{r+1/2} - E_{r+1/2}\beta_i^{r+1/2}\sigma_{r+1/2}\mathbf{u}_{r+1/2}\right\rangle = 0.\end{aligned}$$

Besides, the advection part of the scheme being conservative (cf [BHL21]), the result is proved. □

3.2 Computation of the CFL condition

In this section, we explain how to compute a CFL condition that ensures the positivity of the numerical solution when using an explicit time discretisation of (26). For clarity, we remove all the exponents for the iteration n. The explicit time discretisation of (26) writes:

$$\begin{aligned}&|\Omega_j|\frac{E_j^{n+1}-E_j}{\Delta t} + (1-\eta)\left[(1-\theta)\sum_{r\in\Omega_j}\left\langle\mathbf{u}_r,\mathbf{C}_j^r\right\rangle E_j^r + \theta\sum_{r+1/2\in\Omega_j}\left\langle\mathbf{u}_{r+1/2},\mathbf{C}_j^{r+1/2}\right\rangle E_j^{r+1/2}\right]\\ &+\eta\left[(1-\theta)\sum_{r\in\Omega_j}\left\langle E_j\mathbf{C}_j^r - E_r\beta_j^r\sigma_r\mathbf{u}_r,\mathbf{u}_r\right\rangle + \theta\sum_{r+1/2\in\Omega_j}\left\langle E_j\mathbf{C}_j^{r+1/2} - E_{r+1/2}\beta_j^{r+1/2}\sigma_{r+1/2}\mathbf{u}_{r+1/2},\mathbf{u}_{r+1/2}\right\rangle\right]\\ &= |\Omega_j|\mathcal{S}_j.\end{aligned}$$

We define:

$$\mathfrak{C}_{\mathbf{u}} = \max_{\text{dof}}\|\mathbf{u}_{\text{dof}}\|, \qquad \mathfrak{C}_{\sigma\mathbf{u}} = \max_{\text{dof}}\|\sigma_{\text{dof}}\mathbf{u}_{\text{dof}}\|. \tag{39}$$

We have, for every cell j:

$$\frac{1}{|\Omega_j|}\left|\sum_{r\in\Omega_j}\left\langle\mathbf{u}_r,\mathbf{C}_j^r\right\rangle E_j^r\right| + \frac{1}{|\Omega_j|}\left|\sum_{r+1/2\in\Omega_j}\left\langle\mathbf{u}_{r+1/2},\mathbf{C}_j^{r+1/2}\right\rangle E_j^{r+1/2}\right| \le C\frac{\mathfrak{C}_{\mathbf{u}}}{h}\sum_{i\in\mathcal{V}_j}E_i, \tag{40}$$

$$\frac{1}{|\Omega_j|}\left|\sum_{r\in\Omega_j}\left\langle E_j\mathbf{C}_j^r,\mathbf{u}_r\right\rangle\right| + \frac{1}{|\Omega_j|}\left|\sum_{r+1/2\in\Omega_j}\left\langle E_j\mathbf{C}_j^{r+1/2},\mathbf{u}_{r+1/2}\right\rangle\right| \le C\frac{\mathfrak{C}_{\mathbf{u}}}{h}\sum_{i\in\mathcal{V}_j}E_i, \tag{41}$$

and:

$$\frac{1}{|\Omega_j|}\left(\left|\sum_{r\in\Omega_j}\left\langle E_r\beta_j^r\sigma_r\mathbf{u}_r,\mathbf{u}_r\right\rangle\right| + \left|\sum_{r+1/2\in\Omega_j}\left\langle E_{r+1/2}\beta_j^{r+1/2}\sigma_{r+1/2}\mathbf{u}_{r+1/2},\mathbf{u}_{r+1/2}\right\rangle\right|\right) \le C\mathfrak{C}_{\mathbf{u}}\mathfrak{C}_{\sigma\mathbf{u}}\sum_{i\in\mathcal{V}_j}E_i. \tag{42}$$

Using equations (40), (41), (42) and the fact that $f \ge 0$ lead to:

$$E_j - \Delta t C\mathfrak{C}_{\mathbf{u}}\frac{1+h\mathfrak{C}_{\sigma\mathbf{u}}}{h}\sum_{i\in\mathcal{V}_j}E_i \le E_j^{n+1}.$$

Therefore, assuming that the energy is positive at iteration n, a sufficient condition to have $E^{n+1} > 0$, where E^{n+1} is computed with (26), reads as:

$$\Delta t \leq C \frac{h}{\mathfrak{C}_{\mathbf{u}}} \frac{1}{1 + h\mathfrak{C}_{\sigma\mathbf{u}}} \min_{j \in \mathcal{T}} \left\{ \frac{E_j}{\sum_{i \in \mathcal{V}_j} E_i} \right\}.$$

In the next sections, we estimate the quantities $\mathfrak{C}_{\mathbf{u}}$ and $\mathfrak{C}_{\sigma\mathbf{u}}$ in the cases $P_{\text{dof}} = I_2$ and $P_{\text{dof}} = \sigma_{\text{dof}}$ in (33).

Lemma 3.2. *Let r be a given node. Under Assumptions (35) and (36), the nodal quantity $\mathbf{u}_r$ defined in (28) satisfies:*

$$\|\mathbf{u}_r\| \leq C \frac{1}{\sigma_1 h}, \qquad \|\sigma_r \mathbf{u}_r\| \leq C \frac{1}{h}.$$

Lemma 3.3. *Under Assumptions (35), (36), (37) and (38), the constants $\mathfrak{C}_{\mathbf{u}}$ and $\mathfrak{C}_{\sigma\mathbf{u}}$ can be bounded as follows:*

- *if $P_{dof} = I_2$ then:*

$$\mathfrak{C}_{\mathbf{u}} \leq C \frac{\sigma_2}{\sigma_1^2 h}, \qquad \mathfrak{C}_{\sigma\mathbf{u}} \leq C \frac{\sigma_2^2}{\sigma_1^2 h},$$

- *if $P_{dof} = \sigma_{dof}$ then:*

$$\mathfrak{C}_{\mathbf{u}} \leq C \frac{1}{\sigma_1 h}, \qquad \mathfrak{C}_{\sigma\mathbf{u}} \leq C \frac{1}{h}.$$

Remark 7. *The way we prove Lemma 3.3 does not allow us to simplify the inequalities in terms of σ_1 and σ_2.*

Therefore we see that the constants $\mathfrak{C}_{\mathbf{u}}$ and $\mathfrak{C}_{\sigma\mathbf{u}}$ are larger in the case $P_{\text{dof}} = I_2$ than in the case $P_{\text{dof}} = \sigma_{\text{dof}}$. Thus the CFL condition is better in the second case and reads as:

$$\Delta t \leq C \sigma_1 h^2 \min_{j \in \mathcal{T}} \left\{ \frac{E_j}{\sum_{i \in \mathcal{V}_j} E_i} \right\}. \tag{43}$$

Remark 8. *If σ is isotropic and constant, ie $\sigma_{dof} = \bar{\sigma} I_2$ for some constant $\bar{\sigma} > 0$, then the CFL condition (43) is identical to the CFL for the diffusion limit scheme from [BHL21] which writes:*

$$\Delta t \leq C \bar{\sigma} h^2 \min_{j \in \mathcal{T}} \left\{ \frac{E_j}{\sum_{i \in \mathcal{V}_j} E_i} \right\}.$$

3.3 Proof of Lemma 3.2

One can easily show:

$$\left\| \frac{1}{E_r} \sum_{i | r \in \Omega_i} E_i \mathbf{C}_i^r \right\| \leq Ch.$$

Using $\|\beta_r^{-1}\| \leq Ch^{-2}$ and $\|\sigma_r^{-1}\| \leq C\sigma_1^{-1}$ gives the result.

3.4 Proof of Lemma 3.3

First case: $P_{\mathrm{dof}} = I_2$

Equation (31) can be written as:

$$B_{r+1/2}\mathbf{u}_{r+1/2} = \mathbf{y}_{r+1/2}, \qquad \mathbf{y}_{r+1/2} = \begin{pmatrix} \frac{E_j - E_k}{E_{r+1/2}} \\ \frac{1}{2}\left\langle \mathbf{u}_r + \mathbf{u}_{r+1}, \left[\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\right]^{\perp} \right\rangle \end{pmatrix}, \tag{44}$$

with:

$$B_{r+1/2} = \begin{pmatrix} \left[\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\right]^T \\ \left[\left[\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\right]^{\perp}\right]^T \end{pmatrix} = \begin{pmatrix} \left[\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\right]^{\perp} & \sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j) \end{pmatrix}. \tag{45}$$

We claim that the right hand side of (44) satisfies:

$$\|\mathbf{y}_{r+1/2}\| \leq C\frac{\sigma_2}{\sigma_1}.$$

Indeed, on the one hand we have $|(E_j - E_k)/E_{r+1/2}| \leq C$. On the other hand, defining:

$$R = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

and using Lemma 3.2, we have:

$$|\langle \mathbf{u}_r, [\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)]^{\perp}\rangle| = |\langle \sigma_r \mathbf{u}_r, \sigma_r^{-1} R\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\rangle| \leq C\|\sigma_r^{-1} R\sigma_{r+1/2}\| \leq C\frac{\sigma_2}{\sigma_1}.$$

We easily prove:

$$\|\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\| \geq C\sigma_1 h. \tag{46}$$

Since the matrix $B_{r+1/2}/\|\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\|$ in (45) is a rotation matrix, one has:

$$B_{r+1/2}^{-1} = \frac{1}{\|\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\|^2} B_{r+1/2}^T, \qquad \text{hence : } \|B_{r+1/2}^{-1}\| \leq C\frac{1}{\sigma_1 h}.$$

Moreover, writing:

$$\sigma_{r+1/2}B_{r+1/2}^{-1} = \frac{1}{\|\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j)\|^2} \begin{pmatrix} \sigma_{r+1/2}^2(\mathbf{x}_k - \mathbf{x}_j) & \sigma_{r+1/2} R\sigma_{r+1/2}(\mathbf{x}_k - \mathbf{x}_j) \end{pmatrix},$$

leads to:

$$\|\sigma_{r+1/2}B_{r+1/2}^{-1}\| \leq \frac{\sigma_2}{\sigma_1 h},$$

and this inequality is optimal (we can not find an inequality that does not involve σ_1 and σ_2). This implies:

$$\|\mathbf{u}_{r+1/2}\| \leq C\frac{\sigma_2}{\sigma_1^2 h}, \qquad \|\sigma_{r+1/2}\mathbf{u}_{r+1/2}\| \leq C\frac{\sigma_2^2}{\sigma_1^2 h}.$$

This gives the desired result.

Second case: $P_{\text{dof}} = \sigma_{\text{dof}}$

Equation (44) can be written as:

$$\tilde{B}_{r+1/2}\sigma_{r+1/2}\mathbf{u}_{r+1/2} = \tilde{\mathbf{y}}_{r+1/2},$$

with:

$$\tilde{B}_{r+1/2} = \begin{pmatrix} [\mathbf{x}_k - \mathbf{x}_j]^T \\ \left[(\mathbf{x}_k - \mathbf{x}_j)^\perp\right]^T \end{pmatrix}, \qquad \tilde{\mathbf{y}}_{r+1/2} = \begin{pmatrix} \frac{E_j - E_k}{E_{r+1/2}} \\ \frac{1}{2}\left\langle \sigma_r \mathbf{u}_r + \sigma_{r+1}\mathbf{u}_{r+1}, (\mathbf{x}_k - \mathbf{x}_j)^\perp \right\rangle \end{pmatrix}.$$

Using Lemma 3.2, one can easily show that $\|\tilde{\mathbf{y}}_{r+1/2}\| \leq C$. Using $\|\tilde{B}^{-1}_{r+1/2}\| \leq C/h$ gives the desired result.

4 Upwind advection scheme

In this section, we focus on the scheme obtained by choosing $\eta = 0$ in (26). We show that this scheme has a much less restrictive positivity preserving condition than in the case $\eta \neq 0$. This scheme corresponds to the discretisation of the heat equation using an upwind scheme:

$$|\Omega_j|\frac{d}{dt}E_j + (1-\theta)\sum_{r\in\Omega_j}\left\langle \mathbf{u}_r, \mathbf{C}_j^r\right\rangle E_j^r + \theta\sum_{r+1/2\in\Omega_j}\left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2}\right\rangle E_j^{r+1/2} = |\Omega_j|\mathcal{S}_j. \tag{47}$$

We define:

$$R_j^+ = \{r, \langle \mathbf{C}_j^r, \mathbf{u}_r\rangle > 0\}, \qquad R_j^- = \{r, \langle \mathbf{C}_j^r, \mathbf{u}_r\rangle \leq 0\},$$

$$\tilde{R}_j^+ = \{r+1/2, \langle \mathbf{C}_j^{r+1/2}, \mathbf{u}_{r+1/2}\rangle > 0\}, \qquad \tilde{R}_j^- = \{r+1/2, \langle \mathbf{C}_j^{r+1/2}, \mathbf{u}_{r+1/2}\rangle \leq 0\},$$

and:

$$E_{k(r)} = \frac{1}{\sum_{i\in I_r^+}\langle \mathbf{u}_r, \mathbf{C}_i^r\rangle}\sum_{i\in I_r^+}\langle \mathbf{u}_r, \mathbf{C}_i^r\rangle E_i. \tag{48}$$

Similarly to (48), we define $k(r+1/2)$ as the index of the unique cell containing the edge $r+1/2$ such that $\langle \mathbf{u}_{r+1/2}, \mathbf{C}_i^{r+1/2}\rangle > 0$. Equation (47) writes:

$$|\Omega_j|\frac{d}{dt}E_j + \left[(1-\theta)\sum_{r\in R_j^+}\left\langle \mathbf{u}_r, \mathbf{C}_j^r\right\rangle + \theta\sum_{r+1/2\in\tilde{R}_j^+}\left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2}\right\rangle\right]E_j \tag{49}$$
$$+(1-\theta)\sum_{r\in R_j^-}\left\langle \mathbf{u}_r, \mathbf{C}_j^r\right\rangle E_{k(r)} + \theta\sum_{r+1/2\in\tilde{R}_j^-}\left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2}\right\rangle E_{k(r+1/2)} = |\Omega_j|\mathcal{S}_j.$$

4.1 Explicit time discretisation

The explicit time discretisation of (49) reads as:

$$E_j^{n+1} = E_j\left(1 - \frac{\Delta t}{|\Omega_j|}\left[(1-\theta)\sum_{r\in R_j^+}\left\langle \mathbf{u}_r, \mathbf{C}_j^r\right\rangle + \theta\sum_{r+1/2\in\tilde{R}_j^+}\left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2}\right\rangle\right]\right) \tag{50}$$
$$-\frac{\Delta t}{|\Omega_j|}\left[(1-\theta)\sum_{r\in R_j^-}\left\langle \mathbf{u}_r, \mathbf{C}_j^r\right\rangle E_{k(r)} + \theta\sum_{r+1/2\in\tilde{R}_j^-}\left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2}\right\rangle E_{k(r+1/2)}\right] + \Delta t\mathcal{S}_j.$$

One can notice:

$$-\frac{\Delta t}{|\Omega_j|}\left[(1-\theta)\sum_{r\in R_j^-}\left\langle \mathbf{u}_r,\mathbf{C}_j^r\right\rangle E_{k(r)}+\theta\sum_{r+1/2\in\tilde{R}_j^-}\left\langle \mathbf{u}_{r+1/2},\mathbf{C}_j^{r+1/2}\right\rangle E_{k(r+1/2)}\right]\geq 0. \tag{51}$$

Therefore, reminding that $f\geq 0$, a natural CFL condition writes:

$$\frac{\Delta t}{|\Omega_j|}\left[(1-\theta)\sum_{r\in R_j^+}\left\langle \mathbf{u}_r,\mathbf{C}_j^r\right\rangle+\theta\sum_{r+1/2\in\tilde{R}_j^+}\left\langle \mathbf{u}_{r+1/2},\mathbf{C}_j^{r+1/2}\right\rangle\right]\leq 1. \tag{52}$$

This criterion depends on time but it can be simplified, see Lemma 4.1.

Lemma 4.1. *Under Assumptions (35), (36), (37) and (38), a sufficient condition so as to satisfy Equation (52) reads as:*

- *if* $P_{dof}=I_2$ *then:*

$$\Delta t\leq C\frac{\sigma_1^2}{\sigma_2}h^2,$$

- *if* $P_{dof}=\sigma_{dof}$ *then:*

$$\Delta t\leq C\sigma_1 h^2.$$

Proof. Using (51) and:

$$\frac{\Delta t}{|\Omega_j|}\left|(1-\theta)\sum_{r\in R_j^+}\left\langle \mathbf{u}_r,\mathbf{C}_j^r\right\rangle+\theta\sum_{r+1/2\in\tilde{R}_j^+}\left\langle \mathbf{u}_{r+1/2},\mathbf{C}_j^{r+1/2}\right\rangle\right|\leq C\mathfrak{C}_{\mathbf{u}}\frac{\Delta t}{h^2},$$

where $\mathfrak{C}_{\mathbf{u}}$ is defined in (39), gives the desired result.

□

4.2 Implicit time discretisation

The implicit version writes:

$$\frac{E^{n+1}-E^n}{\Delta t}+B(\mathbf{u}^{n+1})E^{n+1}=\mathcal{S}, \tag{53}$$

where, for all cells j and l:

$$\begin{aligned}(B(\mathbf{u}))_{jl}=1_{j=l}\frac{1}{|\Omega_j|}&\left[(1-\theta)\sum_{r\in R_j^+}\left\langle \mathbf{u}_r,\mathbf{C}_j^r\right\rangle+\theta\sum_{r+1/2\in\tilde{R}_j^+}\left\langle \mathbf{u}_{r+1/2},\mathbf{C}_j^{r+1/2}\right\rangle\right]\\&+\frac{1-\theta}{|\Omega_j|}\sum_{r\in\Omega_j\cap\Omega_l}1_{r\in R_j^-}\left\langle \mathbf{u}_r,\mathbf{C}_j^r\right\rangle 1_{l\in I_r^+}\frac{\langle \mathbf{u}_r,\mathbf{C}_l^r\rangle}{\sum_{i\in I_r^+}\langle \mathbf{u}_r,\mathbf{C}_i^r\rangle}\\&+\frac{\theta}{|\Omega_j|}\sum_{r+1/2\in\Omega_j\cap\Omega_l}1_{r+1/2\in\tilde{R}_j^+}\left\langle \mathbf{u}_{r+1/2},\mathbf{C}_j^{r+1/2}\right\rangle.\end{aligned}$$

System (53) is solved using a fixed-point iteration:

$$\frac{E^{k+1}-E^n}{\Delta t}+B(\mathbf{u}^k)E^{k+1}=\mathcal{S}. \tag{54}$$

Lemma 4.2. *Under Assumptions (35), (36), (37) and (38), the implicit scheme (53) preserves the positivity for any time step Δt.*

Proof. We prove that $I + \Delta t B(\mathbf{u})$ is the transpose of an $M-$matrix, thus it is invertible and its inverse has non-negative coefficients. First, we have:

$$\forall \mathbf{u},\ \forall j,\ (B(\mathbf{u}))_{jj} \geq 0, \qquad \forall l \neq j,\ (B(\mathbf{u}))_{lj} \leq 0.$$

and:

$$\forall j, \sum_l (B(\mathbf{u}))_{lj} = 0 \tag{55}$$

Indeed, the scheme being conservative (see [BHL21]) for the proof):

$$\forall \mathbf{u},\ \forall E,\ \langle \mathbf{1}, B(\mathbf{u})E \rangle = 0, \qquad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

hence:

$$\forall \mathbf{u},\ B(\mathbf{u})^T \mathbf{1} = 0,$$

which gives (55). Therefore $I + \Delta t B(\mathbf{u})$ is the transpose of a strict $M-$matrix, thus its inverse has non negative coefficients. In addition, each line of $(I + \Delta t B(\mathbf{u}))^{-1}$ contains at least one positive coefficient. This gives that, if $E_j^k > 0$ for any cell j, then $E_j^{k+1} > 0$. □

We give now a sufficient condition on the time step Δt so as to ensure the convergence of the fixed-point iteration (54).

Lemma 4.3. *Under Assumptions (35), (36), (37) and (38), and assuming that there exists $\delta > 0$ such that:*

$$\forall k, E_{dof}^k \geq \delta > 0, \tag{56}$$

and if:

$$\Delta t \leq C \sigma_1 h^2 \frac{\delta}{\|E^n + \mathcal{S}\|},$$

then (53) admits a unique solution and the sequence $(E^k)_{k \in \mathbb{N}}$ converges toward it.

Proof. Equation (53) can be written as:

$$f_n(E^{n+1}) = E^{n+1}, \qquad f_n(E) = \left[I + \Delta t B(\mathbf{u}(E))\right]^{-1} (E^n + \mathcal{S}).$$

The fixed point iteration (54) writes:

$$f_n(E^{k+1}) = E^k.$$

We prove here that if Δt is small enough, then f_n is a contraction mapping. This property ensures the convergence of the sequence $(E^k)_{k \in \mathbb{N}}$.
Since $\mathbf{u}$ and B are uniformly bounded with respect to E, we can use ideas from [BL16]. This gives that f_n is a contraction mapping if:

$$C \Delta t \left\| \nabla_E B(\mathbf{u}) \right\| \leq 1.$$

Using the same arguments as in Section 3, it can be proven that there exists a constant such that:

$$\left\| \frac{\partial \mathbf{u}_{\text{dof}}}{\partial E_j} \right\| \leq C \frac{1}{\delta} C_\sigma \frac{1}{\sigma_1 h},$$

with $C_\sigma = \sigma_2/\sigma_1$ if $P_{\text{dof}} = I_2$ and $C_\sigma = 1$ if $P_{\text{dof}} = \sigma_{\text{dof}}$. This leads to:

$$\|\nabla_E B(\mathbf{u})\| \leq C\frac{1}{\delta}C_\sigma\frac{1}{\sigma_1 h^2}.$$

The condition on Δt eventually writes:

$$\Delta t \leq C\delta\frac{1}{C_\sigma}\sigma_1 h^2\frac{1}{\|E^n + \mathcal{S}\|}.$$

We can notice that, as in the explicit case, the choice $P_{\text{dof}} = \sigma_{\text{dof}}$ leads to a less restrictive constraint.

□

Remark 9. *If we change the definition of E_{dof} in (9) and replace it by:*

$$E_{dof} = h^\gamma + \frac{1}{N_{dof}}\sum_{i|\, dof\in\Omega_i} E_i,$$

then we do not need Assumption (56) and the condition on Δt becomes:

$$\Delta t \leq C\frac{1}{C_\sigma}\sigma_1 h^{2+\gamma}\frac{1}{\|E^n + \mathcal{S}\|}.$$

5 Numerical results

In this section, we present some numerical test cases so as to illustrate the good properties of our scheme. We use an explicit time discretisation. For the test cases of Sections 5.1 and 5.3, we define the analytical solution E and compute the source term $\mathbb{S}$ so as to satisfy Equation (2). In some of the test cases, we use random meshes. They are generated by randomly moving the nodes of a cartesian grid. We denote by N_x the number of cells in the x direction and N_y the number of cells in the y direction.
Moreover, it is well known that the purely nodal scheme ($\theta = 0$ in (26)) may exhibit some *cross-stencil* propagation. This issue is corrected using the composite scheme ($\theta > 0$ in (26)). We do not give here any illustration of this property, examples can be found in [BHL21].

5.1 $1D$ test case

For $\mathbf{x} = (x, y)$, the diffusion coefficient is given by:

$$\sigma(x) = \alpha\exp\left(-\sin(2\pi x)\right), \qquad \alpha = 4\pi^2 e, \qquad E(x,t) = \exp\left(t - \sin(2\pi x)\right). \tag{57}$$

Thus (2) becomes:

$$\partial_t E - \frac{1}{\alpha}\partial_x\left[\exp\left(\sin(2\pi x)\right)\partial_x E\right] = \mathbb{S},$$

with:

$$\mathcal{S}(t,\mathbf{x}) = \mathcal{S}(t,x) = e^t\left[\exp\left(-\sin(2\pi x)\right) - e^{-1}\sin(2\pi x)\right].$$

We use cartesian meshes with N_x cells in the x direction and $N_y = 1$ cell in the y direction. The time step is given by $\Delta t = 0.1(\Delta x)^2$. Periodic boundary conditions are imposed. The computational domain is $\Omega = [0,1]^2$. The final time is $t = 0.003$, we choose $\eta = 1/4$, $\theta = 1$ and $P_{\text{dof}} = I_2$. The initial condition is given by $E(t = 0)$ in (57). Figure 3 shows the L^1 error as a function of the space step $h = \Delta x = 1/N_x$ and $N_y = 1$. As it is a $1D$ test case, the results do not significantly vary with θ and we plot the error for $\theta = 1$. We can see that the scheme is first order convergent for any $(\theta, \eta) \in [0,1]^2$ and even second order convergent for $\eta = 1$ and any θ. Moreover, the reconstruction procedure of Section 2.6 makes the scheme second order convergent for any $(\theta, \eta) \in [0,1]^2$.

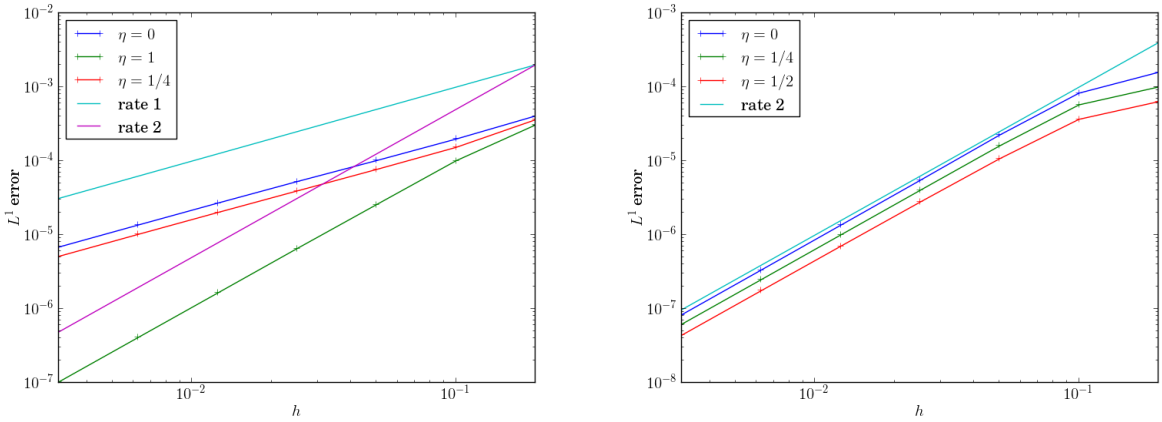

Figure 3: L^1 error at $t = 0.003$, with $\theta = 1$, $P_{\text{dof}} = I_2$ and initial condition given in (57). The right hand side picture is computed using the reconstruction procedure of Section 2.6.

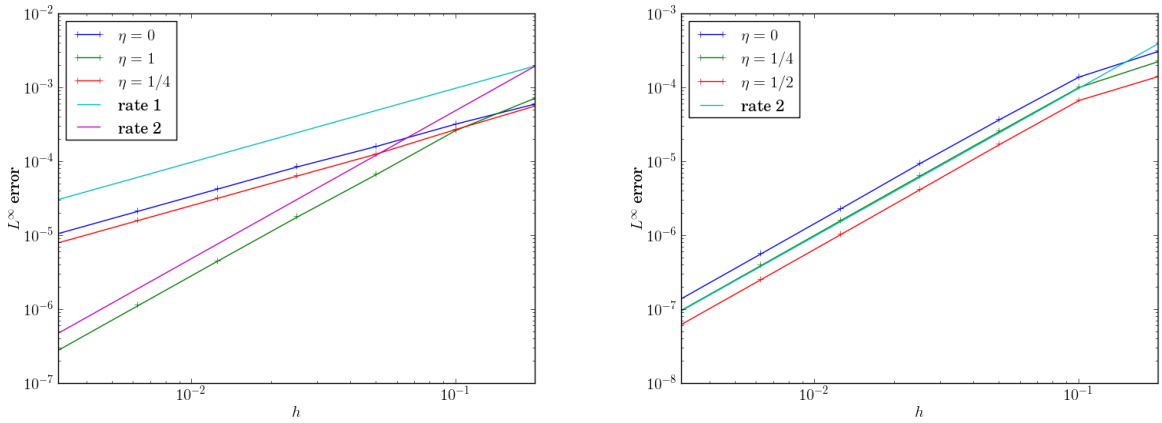

Figure 4: L^∞ error at $t = 0.003$, with $\theta = 1$ and $P_{\text{dof}} = I_2$ and initial condition given in (57). The right hand side picture is computed using the reconstruction procedure of Section 2.6

5.2 Isotropic $2D$ test case

This test case is borrowed from [BHL21]. We set $\sigma = 3I_2$ (which comes down to choosing a scalar diffusion coefficient equal to 3) and $\mathcal{S} = 0$, thus (2) reads as:

$$\partial_t E - \frac{1}{3}\Delta E = 0. \tag{58}$$

Here the two schemes (31) and (32) give exactly the same result. The exact solution of (58) satisfying $E(t = 0) = \delta_{\mathbf{x_0}}$ for a given $\mathbf{x_0}$ is:

$$E(t, \mathbf{x}) = \frac{3}{4\pi t} \exp\left(-3\frac{\|\mathbf{x} - \mathbf{x_0}\|^2}{4t}\right). \tag{59}$$

The initial data is $E(t = t_0)$ and the exact solution is $E(t = t_0 + t_f)$ with $t_0 = 0.01$ and $t_f = 0.001$. We choose $\mathbf{x_0} = (0.5, 0.5)$. The computational domain is $\Omega = [0, 1]^2$. The boundary conditions do not affect the result since the solution is almost 0 on the boundary.

Figures 5 and 6 show the solution on a triangle mesh and on Voronoi type mesh respectively.

For Kershaw type meshes (see Figure 9) and non-convex type meshes (see Figure 13), the time step is given by $\Delta t = (\Delta x)^2/100$.
Figure 7 shows the L^1 error on cartesian meshes for different values of θ and η. For this type of mesh, the time step is given by $\Delta t = (\Delta x)^2/10$. Figure 8 shows the L^1 and L^∞) errors with the reconstruction procedure for $\eta = 1/2$.

Figure 10 (resp 11) shows the L^1 (resp L^∞) error on Kershaw type meshes (see Figure 9).

Figure 7 shows that the scheme is first order convergent for any $(\theta, \eta) \in [0, 1]^2$ and even second order convergent for $\eta = 1$ and any θ. However, one can notice some missing points in Figures 10 and 11. They are due to instabilities of the scheme on Kershaw type meshes, which are highly deformed. Figure 12 shows the L^1 and L^∞ errors with the reconstruction procedure for $\eta = 0$ on Kershaw type meshes. We see that the scheme is more stable for small values of η. Moreover, one can notice that the reconstruction procedure of Section 2.6 allows to reach a second order convergence for $\eta < 1$ and any θ.

Figures 14, 15 and 16 that the scheme is second order convergent in both L^1 and L^∞ norm even on highly deformed meshes with non convex cells 13.

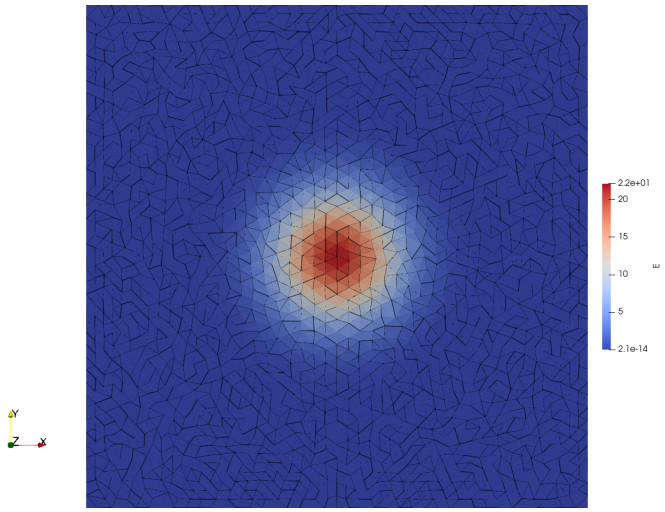

Figure 5: Numerical solution at time $t = 0.001$ with $\eta = 1$ and $\theta = \pi/4$ on a triangle mesh.

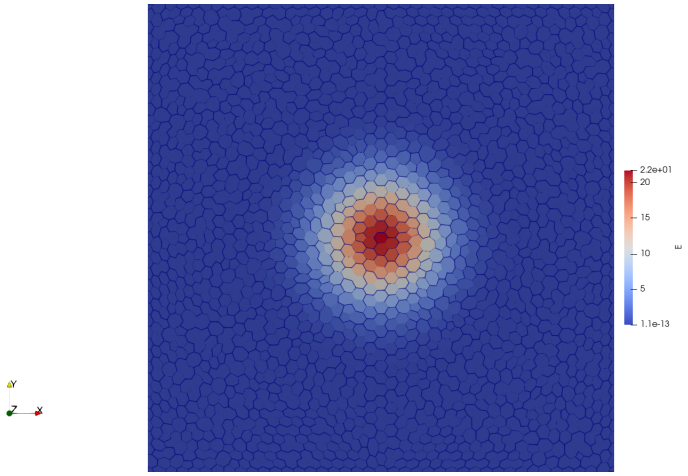

Figure 6: Numerical solution at time $t = 0.001$ with $\eta = 1$ and $\theta = \pi/4$ on a Voronoi type mesh.

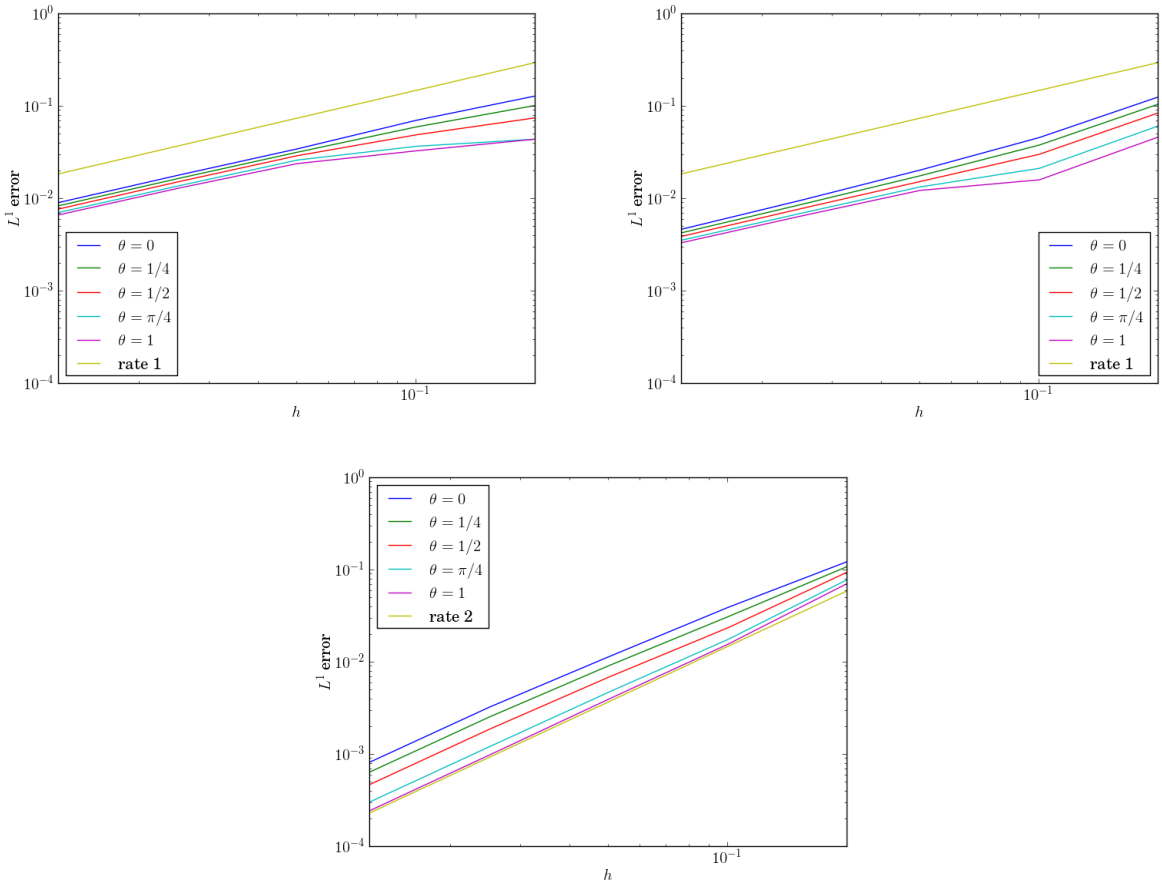

Figure 7: L^1 error on cartesian meshes for different values of θ with $\eta = 0$ (up left), $\eta = 1/2$ (up right) and $\eta = 1$ (down).

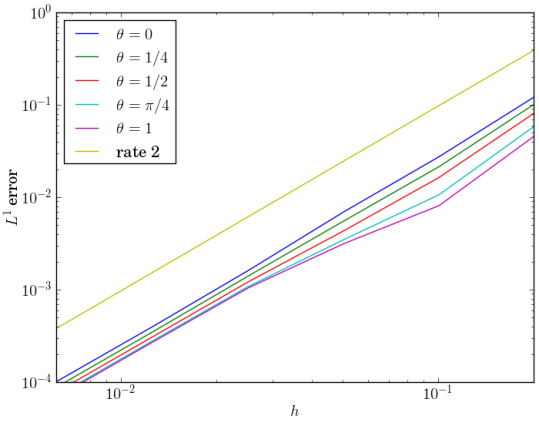

Figure 8: L^1 error on cartesian meshes for different values of θ with $\eta = 1/2$ and the reconstruction procedure.

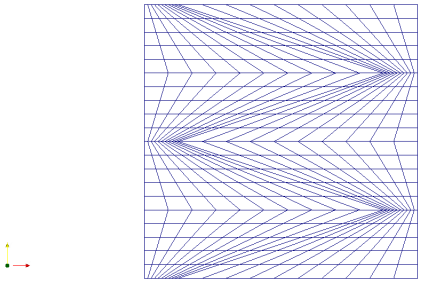

Figure 9: Kershaw type mesh of size 20×20.

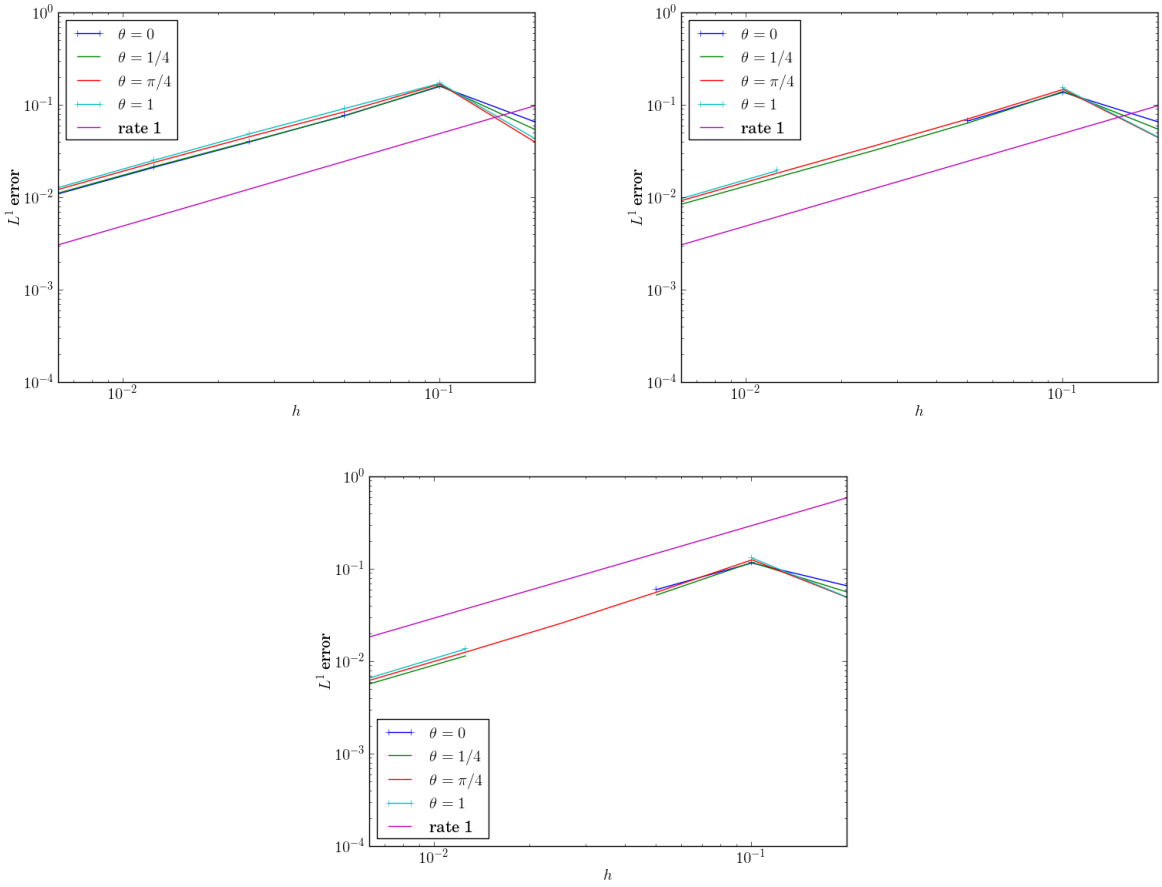

Figure 10: L^1 error on Kershaw type meshes for different values of θ with $\eta = 0$ (up left), $\eta = 1/4$ (up right) and $\eta = 1/2$ (down).

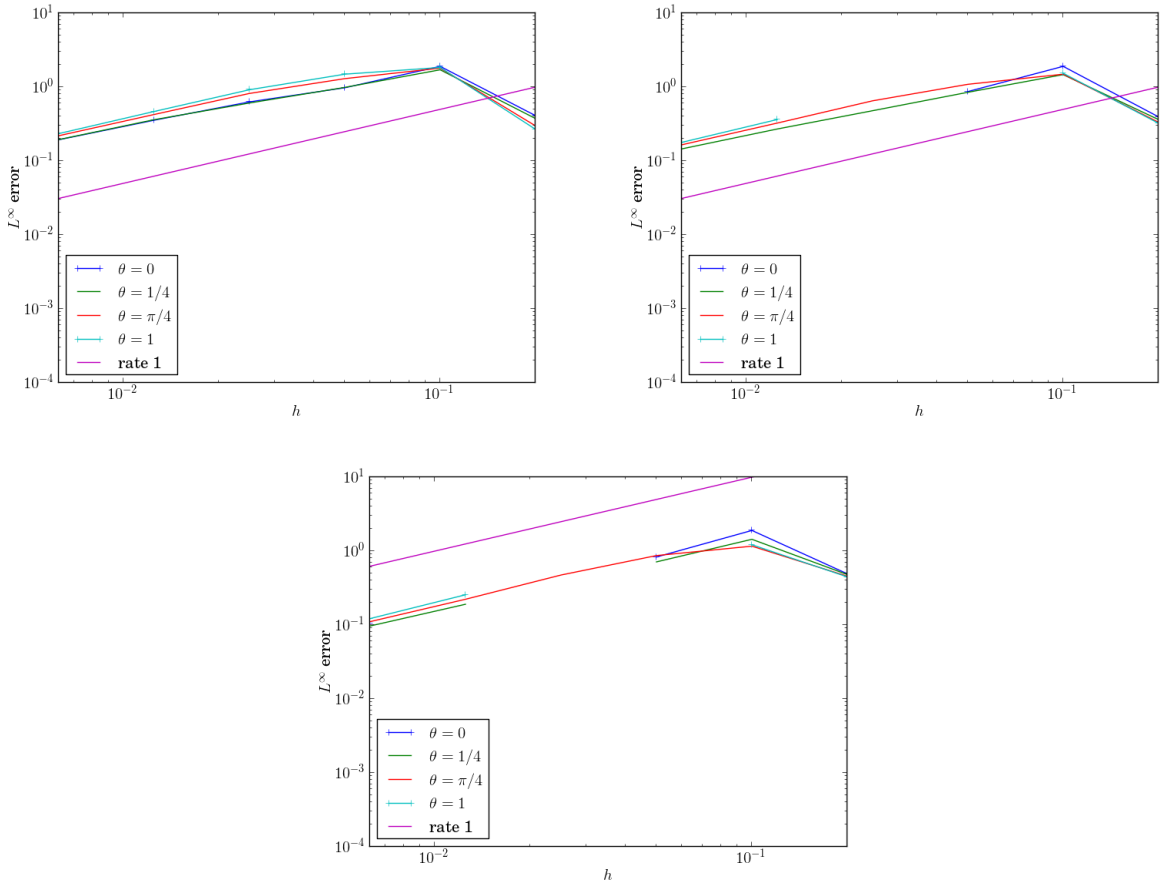

Figure 11: L^∞ error on Kershaw type meshes for different values of θ with $\eta = 0$ (up left), $\eta = 1/4$ (up right) and $\eta = 1/2$ (down).

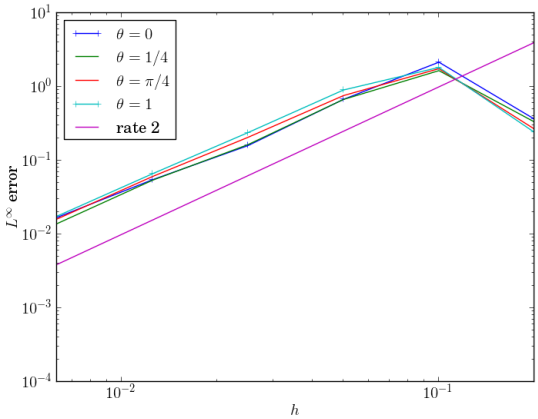

Figure 12: L^1 (left) and L^∞ (right) errors on Kershaw type meshes for different values of θ with $\eta = 0$ and the reconstruction procedure.

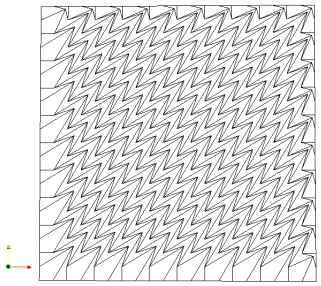

Figure 13: non-convex type mesh

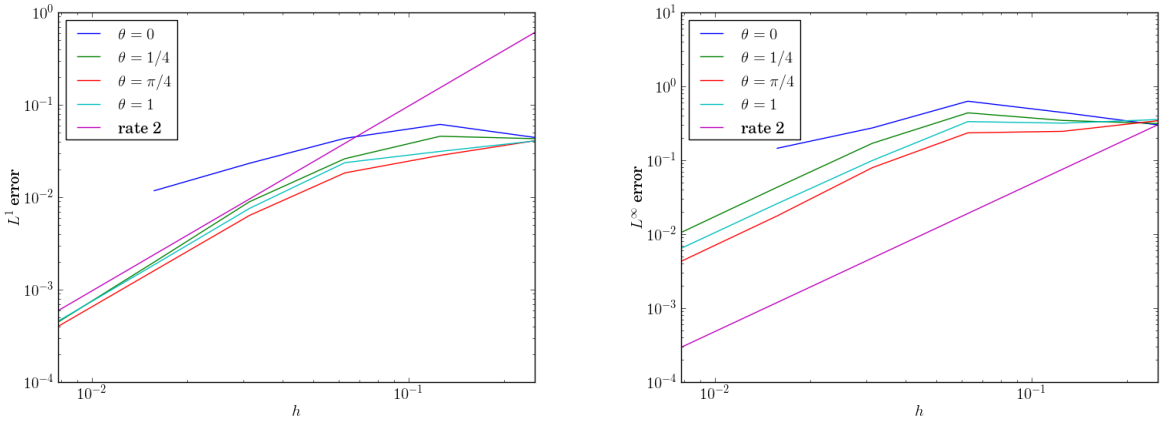

Figure 14: L^1 (left) and L^∞ errors on non-convex type meshes with $\eta = 1$.

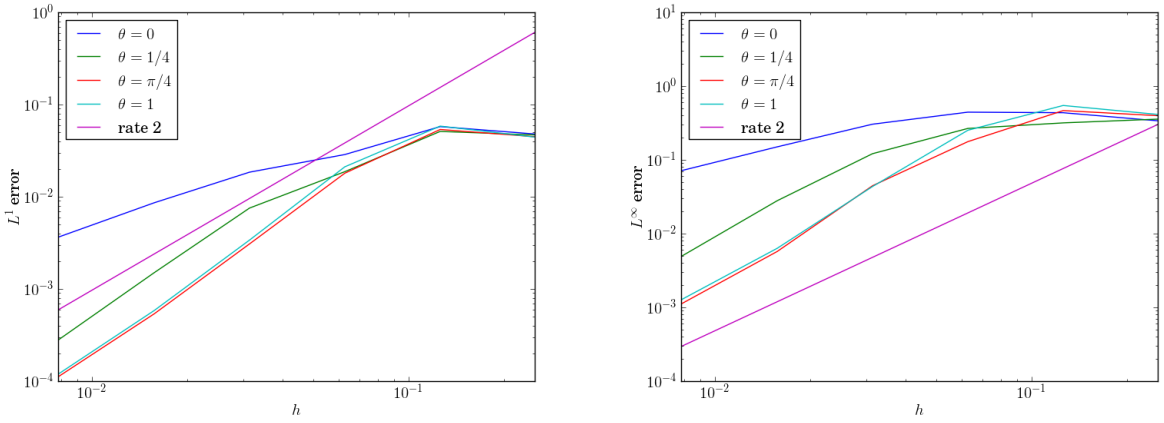

Figure 15: L^1 (left) and L^∞ errors on non-convex type meshes with $\eta = 0$ and the reconstruction procedure.

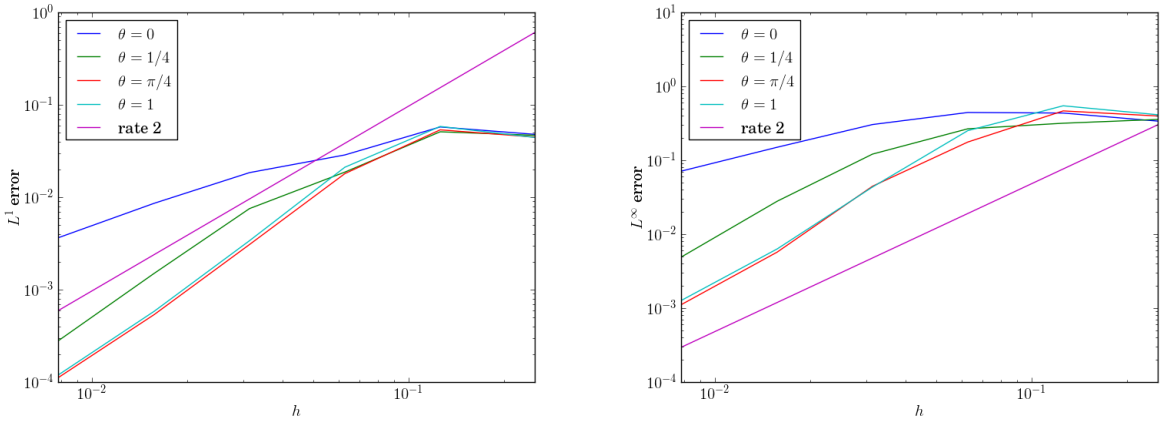

Figure 16: L^1 (left) and L^∞ errors on non-convex type meshes with $\eta = 1/2$ and the reconstruction procedure.

5.3 Stationary analytical solution

This test case comes from [LP20] and [CCP13]. The computational domain is $\Omega = [0, 0.5]^2$. The time step is given by $\Delta t = 0.01(\Delta x)^2$. The final time is $t_f = 0.001$. The solution reads as:

$$E(t, \mathbf{x}) = 1 + \sin(\pi x)\sin(\pi y).$$

The diffusion coefficient is given by:

$$\kappa(x, y) = \sigma^{-1}(x, y) = \frac{1}{x^2 + y^2}\begin{pmatrix} y^2 + \alpha x^2 & -(1-\alpha)xy \\ -(1-\alpha)xy & x^2 + \alpha y^2 \end{pmatrix}, \qquad \alpha = 10^{-6}.$$

Its eigenvalues are 1 and α. The source term reads as:

$$\mathcal{S}(x, y) = -\left\langle \nabla E(x, y), \mathrm{div}\kappa(x, y)\right\rangle - \mathrm{Tr}\left(\kappa(x, y)H(x, y)\right),$$

with:

$$\nabla E(x, y) = \pi\begin{pmatrix}\sin(\pi x)\cos(\pi y) \\ \cos(\pi x)\sin(\pi y)\end{pmatrix}, \qquad \mathrm{div}\kappa(x, y) = \frac{1}{x^2 + y^2}\left[(3\alpha - 1)I_2 - 2\kappa(x, y)\right]\begin{pmatrix} x \\ y\end{pmatrix}$$

and:

$$H(x, y) = \pi^2\begin{pmatrix} -\sin(\pi x)\sin(\pi y) & \cos(\pi x)\cos(\pi y) \\ \cos(\pi x)\cos(\pi y) & \sin(\pi x)\sin(\pi y)\end{pmatrix}.$$

We use random meshes, see Figure 17 for an example. Dirichlet boundary conditions are imposed : the numerical solution is set to be equal to the exact solution on the boundary of the domain. Figure 18 and 20 (resp 19 and 21) show the L^1 (resp L^∞) error for different values of η and θ and for the two possibles formulas for P_{dof}. The space step is $h = \Delta x = 1/N_x = \Delta y = 1/N_y$. We can see some missing points on the curve $P_{\mathrm{dof}} = I_2$. This is due to instabilities of the scheme with this choice of P_{dof}. The scheme is more stable when choosing $P_{\mathrm{dof}} = \sigma_{\mathrm{dof}}$. We can see that the scheme is first order convergent for any $(\theta, \eta) \in [0, 1]^2$ and even second order convergent for $\eta = 1$ and any θ.

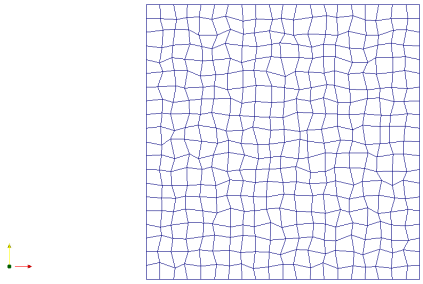

Figure 17: Random mesh of size 20×20.

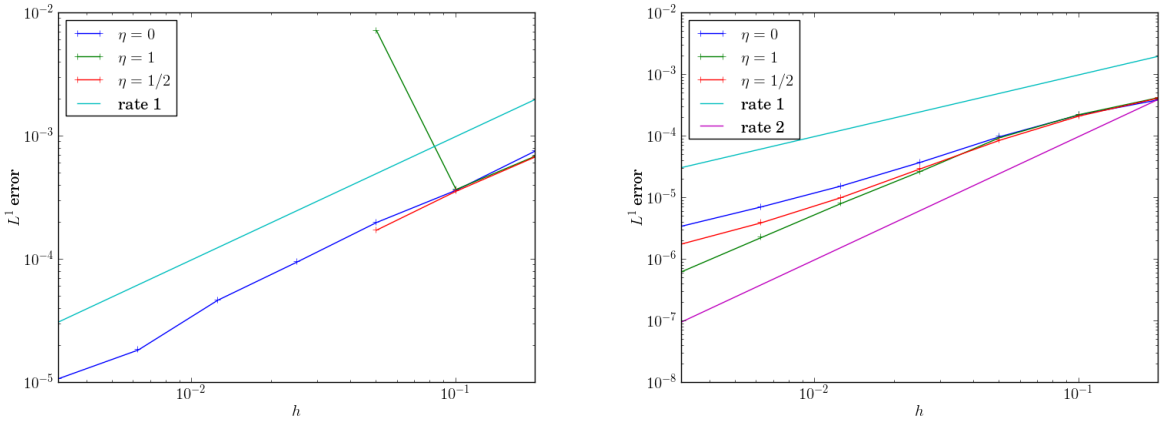

Figure 18: L^1 error on random meshes with $\theta = \pi/4$ for different values of η with $P_{\text{dof}} = I_2$ (left) and $P_{\text{dof}} = \sigma_{\text{dof}}$ (right).

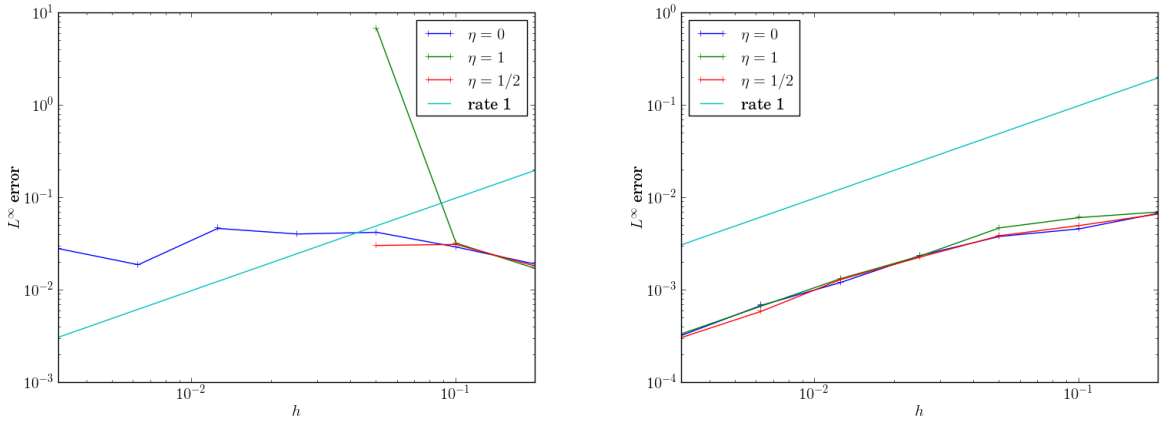

Figure 19: L^∞ error on random meshes with $\theta = \pi/4$ for different values of η with $P_{\text{dof}} = I_2$ (left) and $P_{\text{dof}} = \sigma_{\text{dof}}$ (right).

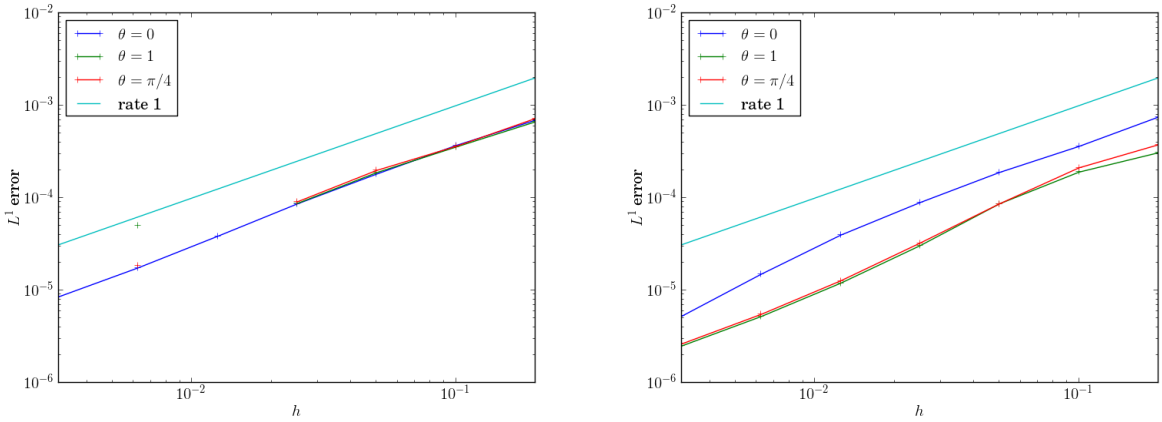

Figure 20: L^1 error on random meshes with $\eta = 1/4$ for different values of θ with $P_{\text{dof}} = I_2$ (left) and $P_{\text{dof}} = \sigma_{\text{dof}}$ (right).

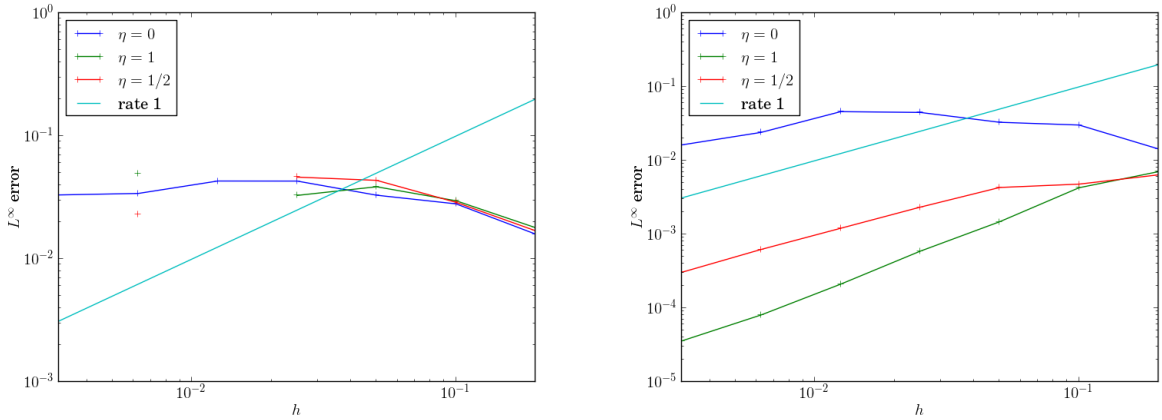

Figure 21: L^∞ error on random meshes with $\eta = 1/4$ for different values of θ with $P_{\text{dof}} = I_2$ (left) and $P_{\text{dof}} = \sigma_{\text{dof}}$ (right).

6 Conclusion

In this work, we propose an extension of a diffusion scheme to the anisotropic diffusion equation. In addition, we develop a two-parameter family of consistent schemes which are positivity-preserving under a CFL condition. As in [BDH21] we have shown that the flaws of the classical nodal-bases scheme ($\theta = 0$) are corrected by the composite scheme ($\theta > 0$). Moreover, we emphasize that the edge-based version of our scheme ($\theta = 1$) is consistent on unstructured meshes. A relevant perspective to this work would be to modify our method so as to make it third order consistent.

7 Appendix

7.1 Link with a classical cartesian grid solver

Choosing $\eta = \theta = 1$ in (26) with a scalar diffusion coefficient $\sigma(\mathbf{x})$ allows to recover the classical 5-points flux scheme on a cartesian grid ($N_x = N_y$). Indeed, one has:

$$\mathbf{C}_j^{r+1/2} = \mathbf{x}_k - \mathbf{x}_j, \qquad |\Omega_j| = h^2 = (\Delta x)^2 \qquad \left\langle \mathbf{u}_{r+1/2}, \mathbf{C}_j^{r+1/2} \right\rangle = \left\langle \mathbf{u}_{r+1/2}, \mathbf{x}_k - \mathbf{x}_j \right\rangle = 2\frac{E_j - E_k}{\sigma_{r+1/2}(E_j + E_k)},$$

and $\left\langle \mathbf{u}_{r+1/2}, (\mathbf{x}_k - \mathbf{x}_j)^{\perp} \right\rangle$ is not involved. Using:

$$\beta_j^{r+1/2}\mathbf{u}_{r+1/2} = \frac{1}{2}\left\langle \mathbf{u}_{r+1/2}, \mathbf{x}_k - \mathbf{x}_j \right\rangle \mathbf{C}_j^{r+1/2} = \frac{1}{2}\frac{E_j - E_k}{\sigma_{r+1/2}E_{r+1/2}}(\mathbf{x}_k - \mathbf{x}_j) = \frac{E_j - E_k}{\sigma_{r+1/2}(E_j + E_k)}(\mathbf{x}_k - \mathbf{x}_j),$$

we have:

$$\left\langle E_j\mathbf{C}_j^{r+1/2} - E_{r+1/2}\beta_j^{r+1/2}\sigma_{r+1/2}\mathbf{u}_{r+1/2}, \mathbf{u}_{r+1/2} \right\rangle = 2[E_j - \frac{1}{2}(E_j - E_k)]\frac{E_j - E_k}{\sigma_{r+1/2}(E_j + E_k)} = \frac{E_j - E_k}{\sigma_{r+1/2}}.$$

Eventually, the scheme (26) now reads as:

$$\frac{d}{dt}E_j + \sum_{r+1/2\in\Omega_j} \frac{E_j - E_k}{\sigma_{r+1/2}(\Delta x)^2} = \mathcal{S}_j. \tag{60}$$

Using the notations of Figure 22, Equation (60) writes:

$$\frac{d}{dt}E_j + \sum_{l=1}^{4} \frac{E_j - E_{k_l}}{\sigma_{k_l}(\Delta x)^2} = \mathcal{S}_j,$$

which is exactly the classical 5-points flux scheme.

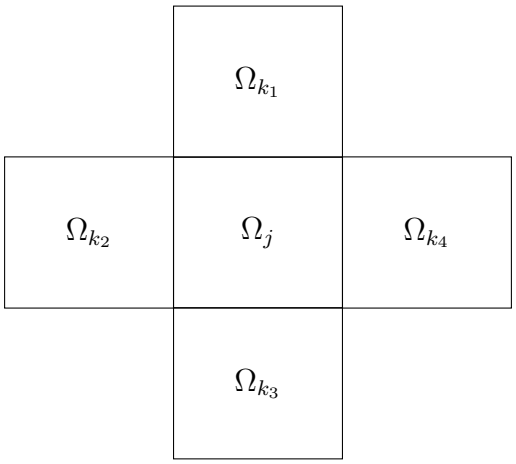

Figure 22: On cartesian grid, the typical five points scheme is recovered choosing $\eta = 1$ and $\theta = 1$.

7.2 Proof of Theorem 2.1

Let $j \in \mathcal{T}$, we show the following equality: let $\boldsymbol{\xi} \in \mathbb{R}^2$ and $\theta \in [0, 1]$,

$$\int_{\partial\Omega_j} \langle \boldsymbol{\xi}, \mathbf{x}\rangle \mathbf{n} d\mathbf{x} = (1-\theta) \sum_{r\in\Omega_j} \langle \boldsymbol{\xi}, \mathbf{x}_r\rangle \mathbf{C}_j^r + \theta \sum_{r+1/2\in\Omega_j} \langle \boldsymbol{\xi}, \mathbf{x}_{r+1/2}\rangle \mathbf{C}_j^{r+1/2}. \tag{61}$$

On the one hand, we have

$$\int_{\partial\Omega_j} \langle \boldsymbol{\xi}, \mathbf{x}\rangle \mathbf{n} d\mathbf{x} = \sum_{r+1/2\in\Omega_j} \left(\int_{x_r}^{x_{r+1}} \langle \boldsymbol{\xi}, \mathbf{x}\rangle \mathbf{n} d\mathbf{x} \right).$$

On each edge, the outward unit normal vector $\mathbf{n}$ is constant and it is given by $\mathbf{n}_j^{r+1/2} = \mathbf{C}_j^{r+1/2}/\|\mathbf{C}_j^{dof}\|$. Therefore we have:

$$\sum_{r+1/2\in\Omega_j}\left(\int_{x_r}^{x_{r+1}}\langle\boldsymbol{\xi},\mathbf{x}\rangle\mathbf{n}d\mathbf{x}\right) = \sum_{r+1/2\in\Omega_j}\left\langle\boldsymbol{\xi},\ \underbrace{\int_{\mathbf{x}_r}^{\mathbf{x}_{r+1}}\mathbf{x}d\mathbf{x}}_{=\|\mathbf{x}_{r+1}-\mathbf{x}_r\|\mathbf{x}_{r+1/2}}\right\rangle\mathbf{n}_j^{r+1/2} = \sum_{r+1/2\in\Omega_j}\left\langle\boldsymbol{\xi},\mathbf{x}_{r+1/2}\right\rangle\mathbf{C}_j^{r+1/2}.$$

Moreover, we have:

$$\sum_{r+1/2\in\Omega_j}\left\langle\boldsymbol{\xi},\mathbf{x}_{r+1/2}\right\rangle\mathbf{C}_j^{r+1/2} = \sum_{r\in\Omega_j}\langle\boldsymbol{\xi},\mathbf{x}_r\rangle\,\mathbf{C}_j^r. \tag{62}$$

Indeed, using (3) leads to:

$$\begin{aligned}\sum_{r\in\Omega_j}\langle\boldsymbol{\xi},\mathbf{x}_r\rangle\,\mathbf{C}_j^r &= \frac{1}{2}\sum_{r\in\Omega_j}\langle\boldsymbol{\xi},\mathbf{x}_r\rangle\,\mathbf{C}_j^{r-1/2} + \frac{1}{2}\sum_{r\in\Omega_j}\langle\boldsymbol{\xi},\mathbf{x}_r\rangle\,\mathbf{C}_j^{r+1/2}\\ &= \frac{1}{2}\sum_{r+1/2\in\Omega_j}\langle\boldsymbol{\xi},\mathbf{x}_{r+1}\rangle\,\mathbf{C}_j^{r+1/2} + \frac{1}{2}\sum_{r+1/2\in\Omega_j}\langle\boldsymbol{\xi},\mathbf{x}_r\rangle\,\mathbf{C}_j^{r+1/2},\end{aligned}$$

which gives (62). This proves (61) and gives the result (7).

References

[AEK+07] I. Aavatsmark, G.T. Eigestad, R.A. Klausen, M.F. Wheeler, and I. Yotov. Convergence of a symmetric MPFA method on quadrilateral grids. *Comput. Geosci.*, 11(4):333–345, 2007.

[AN21] Ashwani Assam and Ganesh Natarajan. A novel least squares finite volume scheme for discontinuous diffusion on unstructured meshes. *Comput. Math. Appl.*, 96:120–130, 2021.

[AWY97] Todd Arbogast, Mary F. Wheeler, and Ivan Yotov. Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences. *SIAM J. Numer. Anal.*, 34(2):828–852, 1997.

[BaDVBC+13] L. Beirão Da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L.D. Marini, and A. Russo. Basic principles of virtual element methods. *Mathematical Models and Methods in Applied Sciences*, 23(1):199 – 214, 2013.

[BBL09] Franco Brezzi, Annalisa Buffa, and Konstantin Lipnikov. Mimetic finite differences for elliptic problems. *ESAIM, Math. Model. Numer. Anal.*, 43(2):277–295, 2009.

[BCHS20] 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.

[BDF12] 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.

[BDH21] 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.

[BHL21] Xavier Blanc, Philippe Hoch, and Clément Lasuen. An asymptotic preserving scheme for the M1 model on conical meshes. working paper or preprint, 2021.

[BL16] Xavier Blanc and Emmanuel Labourasse. A positive scheme for diffusion problems on deformed meshes. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 96(6):660–680, 2016.

[BM05] Enrico Bertolazzi and Gianmarco Manzini. A second-order maximum principle preserving finite volume method for steady convection-diffusion problems. *SIAM J. Numer. Anal.*, 43(5):2172–2199, 2005.

[BM07] Jérôme Breil and Pierre-Henri Maire. A cell-centered diffusion scheme on two-dimensional unstructured meshes. *J. Comput. Phys.*, 224(2):785–823, 2007.

[CCP13] Clément Cancès, Mathieu Cathala, and Christophe Potier. Monotone corrections for generic cell-centered finite volume approximations of anisotropic diffusion equations. *Numerische Mathematik*, 125, 11 2013.

[CVV99] Yves Coudière, Jean-Paul Vila, and Philippe Villedieu. Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem. *M2AN Math. Model. Numer. Anal.*, 33(3):493–516, 1999.

[DF99] 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.

[DK87] 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.

[Eym10] Eymard, R. and Gallouët, T. and Herbin, R. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: A scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 30(4):1009–1043, 2010.

[FBD11] 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.

[Fra12] 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.

[Her98] F. Hermeline. A finite volume method for second-order elliptic equations. (Une méthode de volumes finis pour les équations elliptiques du second ordre.). *C. R. Acad. Sci. Paris, Ser. I*, 1998.

[Her00] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, 160(2):481–499, 2000.

[Her03] F. Hermeline. Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. *Comput. Methods Appl. Mech. Eng.*, 192(16-18):1939–1959, 2003.

[Her07] F. Hermeline. Approximation of 2D and 3D diffusion operators with variable full tensor coefficients on arbitrary meshes. *Comput. Methods Appl. Mech. Eng.*, 196(21-24):2497–2526, 2007.

[Hoc22] Philippe Hoch. Nodal extension of Approximate Riemann Solvers and nonlinear high order reconstruction for finite volume method on unstructured polygonal and conical meshes: the homogeneous case. working paper or preprint, February 2022.

[Ker81] David S. Kershaw. Differencing of the diffusion equation in Lagrangian hydrodynamic codes. *J. Comput. Phys.*, 39:375–395, 1981.

[LMS14] Konstantin Lipnikov, Gianmarco Manzini, and Mikhail Shashkov. Mimetic finite difference method. *Journal of Computational Physics*, 257, Part B(0):1163 – 1227, 2014. Physics-compatible numerical methods.

[LP09] Christophe Le Potier. A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators. *Int. J. Finite Vol.*, 6(2):20, 2009.

[LP20] 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.

[NSL22] Cunyun Nie, Shi Shu, and Menghuan Liu. A novel monotone finite volume element scheme for diffusion equations. *J. Comput. Appl. Math.*, 414:Paper No. 114458, 20, 2022.

[Per81] G. J. Pert. Physical constraints in numerical calculations of diffusion. *J. Comput. Phys.*, 42(1):20–52, 1981.

[RT83] P.-A. Raviart and J.-M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983.

[SY16] Zhiqiang Sheng and Guangwei Yuan. A new nonlinear finite volume scheme preserving positivity for diffusion equations. *J. Comput. Phys.*, 315:182–193, 2016.

[SYY09] Zhiqiang Sheng, Jingyan Yue, and Guangwei Yuan. Monotone finite volume schemes of nonequilibrium radiation diffusion equations on distorted meshes. *SIAM J. Sci. Comput.*, 31(4):2915–2934, 2009.

[WPL+22] Xiaoxin Wu, Kejia Pan, Jin Li, Yunlong Yu, Guangwei Yuan, and Zhengyong Ren. A robust, interpolation-free and monotone finite volume scheme for diffusion equations on arbitrary quadrilateral meshes. *Internat. J. Numer. Methods Engrg.*, 123(16):3631–3657, 2022.

[YSGN22] Di Yang, Meihua Sheng, Zhiming Gao, and Guoxi Ni. The VEM-based positivity-preserving conservative scheme for radiation diffusion problems on generalized polyhedral meshes. *Comput. & Fluids*, 239:Paper No. 105356, 20, 2022.