High-order monotone finite-volume schemes for 1D elliptic problems

Xavier Blanc[1], Francois Hermeline[2,3], Emmanuel Labourasse[2,3], and Julie Patela[1,2]

[1]Université de Paris, Sorbonne Université, CNRS, Laboratoire Jacques-Louis Lions, F-75013 Paris, France;
[2]CEA, DAM, DIF, F-91297 Arpajon, France;
[3]Université Paris-Saclay, CEA DAM DIF, Laboratoire en Informatique Haute Performance pour le Calcul et la simulation, 91297 Arpajon, France;

November 10, 2021

Abstract

When solving numerically an elliptic problem, it is important in most applications that the scheme used preserves the positivity of the solution. When using finite volume schemes on deformed mesh, the question has been solved rather recently. Such schemes are usually (at most) second order convergent, and nonlinear. On the other hand, many high-order schemes have been proposed, that do not ensure positivity of the solution. In this paper we propose a very high-order *monotone* (that is, positivity preserving) numerical method for elliptic problems in 1D. We prove that this method converges to an arbitrary order and is indeed monotone. We also show how to handle discontinuous sources or diffusion coefficients, while keeping the order of convergence. We assess the new scheme, on several test problems, with arbitrary (regular, distorted, random) meshes.

Contents

Introduction

In this paper we are interested in the resolution of the following elliptic problem with mixed boundary conditions

$$\begin{cases} -\operatorname{div}(\kappa\nabla\bar{u}) + \alpha\bar{u} = f & \text{in} \quad \Omega, \\ \beta\bar{u} + \gamma\kappa\nabla\bar{u}\cdot\mathbf{n} = g & \text{on} \quad \partial\Omega, \end{cases} \tag{1}$$

where Ω is a bounded open domain of $\mathbb{R}^d$ and $\mathbf{n} \in \mathbb{R}^d$ the external unit normal vector, with d the dimension. The data are such that $f \in L^2(\Omega)$, $g \in H^{1/2}(\partial\Omega)$, $\alpha \in \mathbb{R}^+\backslash\{0\}$, and $\kappa \in L^\infty(\Omega)$. The diffusion coefficient κ satisfies the ellipticity condition

$$\forall x \in \Omega, \qquad \kappa(x) \geq \kappa_0 > 0. \tag{2}$$

Besides, β and γ are functions such that

$$\forall x \in \partial\Omega, \qquad \beta(x) \geq 0, \qquad \gamma(x) \geq 0$$

and they do not vanish at the same point. Under the above conditions, one can prove (see [10]) that system (1) has a unique solution in $H^1(\Omega)$. This solution satisfies a positivity principle, i.e. if $f \geq 0$ and $g \geq 0$, then $\bar{u} \geq 0$. For linear problems considered in this work, this property is equivalent to a maximum principle on $\bar{u}$, which can be stated as follows: if the data f_1, f_2 and g_1, g_2 are such that $f_1 \leq f_2$ and $g_1 \leq g_2$, then the associated solutions to (1), that we denote by $\bar{u}_1$ and $\bar{u}_2$ respectively, satisfy $\bar{u}_1 \leq \bar{u}_2$ almost everywhere in Ω.

Because system (1) is intended to model, for instance, concentration diffusion and thermal conduction, preservation of the positivity principle at the discrete level is highly desirable. The standard finite volume two-point flux approximation (TPFA, see for example [12]) is positivity preserving (one also says montone) but is unfortunately inconsistent on deformed meshes, in dimension $d \geq 2$. For this reason, a great deal of work has been devoted to the design of positivity preserving schemes on general (namely non-orthogonal) meshes over the past two decades. While elliptic problems are often solved using a Finite Element discretization, all the works we know of on monotone methods deal with Finite Volume schemes. The Finite Volume framework is well suited to achieve montonicity because it allows for an easy manipulation of the fluxes. The first works we know of are those of Le Potier [15] and Bertolazzi and Manzini [2]. In such methods, one uses a manipulation of the fluxes that leads to introduce a dependence on the discrete solution in the coefficients of the fluxes, making the scheme non-linear, although (1) is linear. Thus, mononicity is in general not equivalent to the maximum principle. In general, one also introduces secondary unknowns (for instance vertex-located or edge-located unknows) in addition to the primarly (cell-located) unknowns. Among others, important contributions to this field are [3, 17, 25], which propose efficient numerical schemes preserving the positivity of the primary unknowns. In [20], the requirement of positive secondary unknowns is relaxed. In [4], a non-linear solver based on an iterative resolution of two problems is described, the primary unknowns of one problem being the secondary unknowns of the other one. The works [26, 18] explain how to build monotone schemes without relying on secondary unknowns. In [16, 19, 21], maximum principle preserving schemes are proposed. Some concepts and proofs about the existence of solutions for these types of scheme can be found in [6, 9]. Recent

advances in this field are [24, 23]. All the works mentioned above concern 2D or 3D low-order (that is at most of order 2) numerical methods.

We are interested in designing high-order positive scheme (that is at least of order 3). We start, in the present paper, with the 1D case. Thus, for now on, the system we study is the 1D version of (1), that is,

$$\begin{cases} -\dfrac{d}{dx}\left(\kappa\dfrac{d\bar{u}}{dx}\right)+\alpha\bar{u}=f & \text{in} \quad \Omega, \\ \beta\bar{u}+\gamma\kappa\dfrac{d\bar{u}}{dn}=g & \text{on} \quad \partial\Omega, \end{cases} \tag{3}$$

when $\Omega =]0,1[$.

Although this setting is very specific, we believe it can be seen as a first step to tackle the question in higher dimension. Let us be more precise about the 1D setting: in such a case, the TPFA scheme is actually consistent (and monotone), contrary to dimensions $d \geq 2$. Thus, the relevant question here is to design a high-order scheme that satisfies the positivity principle. Of course, as one may expect, a naive extension to higher orders of the TPFA scheme gives non-positive schemes. In particular, none of the existing [1, 5, 7, 8] arbitrary high order methods for the problem (1) is monotone. In [6] it is shown how to use Le Potier's trick [16] to obtain monotone 1D schemes of order greater than 2. But this method uses a finite difference discretization on Cartesian meshes, that seems hard to extend to general meshes even in 1D. In the present paper we propose a new numerical method that has the following properties:

- it has a provable arbitrarily high order of accuracy;
- it is monotone;
- it is conservative, and
- it operates on general 1D meshes.

The organization of the paper is as follows. In Section 1 we design a high-order Finite-Volume method by integrating the k-th order Taylor expansion of the unknown. The high-order derivatives of this series are approximated using to a polynomial reconstruction of the solution while the degrees of freedom are the integral mean values of the solution on the cells. The monotone behavior of the scheme is enforced using the trick described in [14], which leads to a non-linear resolution. A symmetric version of the scheme is also proposed, allowing to obtain a Local Maximum Preserving (LMP) structure (see for instance to [9] for a definition) for the fluxes. In Section 2, we prove the properties of the method: conservation, consistency of the fluxes at order k, monotonicity (or the LMP structure for the symmetric version) and convergence of the scheme at order k. Section 3 describes some implementation details required by the method as the fixed-point iterations in order to solve the non-linearity. Finally in Section 4 we verify the properties previously stated on 1D test problems, showing that the method is indeed monotone and of order k.

In all the article, C will denote an unspecified strictly positive constant independent of the mesh size.

1 High-order finite volume scheme

Consider a mesh whose cells are numbered from 1 to n. The center of cell i is denoted by x_i and its two vertices are $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$. The length of cell i is h_i and the length between the centers x_i and x_{i+1} is $h_{i+\frac{1}{2}}$, see Fig. 1. Without loss of generality, we will suppose that

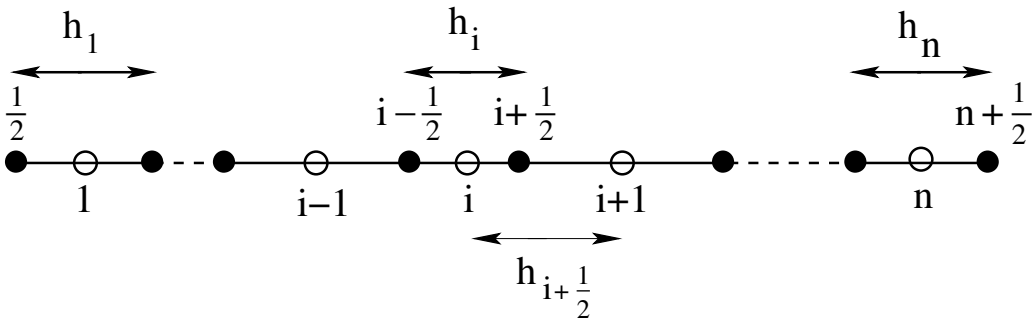
$$x_i < x_{i+1}, \forall i \in [\![1, n-1]\!]. \tag{4}$$

We will also assume a regularity condition on the mesh: there exists C such that

$$\frac{\max\limits_{1\leq i\leq n}(h_i)}{\min\limits_{1\leq i\leq n}(h_i)} < C. \tag{5}$$

We define $h = \max\limits_{1\leq i\leq n}(h_i)$ and $\mathbf{u} = (u_i)_{1\leq i\leq n}$. The notation $\mathbf{u} > \mathbf{0}$ (resp. $\mathbf{u} \geq \mathbf{0}$) means that

$$u_i > 0, \ (\text{resp. } u_i \geq 0) \ \ \forall i \in [\![1, n]\!].$$



Let us introduce some notations for the norms we are going to use. We first define the L^p norm, $p \in [1, +\infty[$

$$\begin{aligned} \|\cdot\|_{L^p} : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \mathbf{u} &\longmapsto \left(\sum_{i=1}^{n} h_i |u_i|^p\right)^{1/p} \end{aligned} \tag{6}$$

and the L^∞ norm

$$\begin{aligned} \|\cdot\|_{L^\infty} : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \mathbf{u} &\longmapsto \max_{1\le i\le n} |u_i|. \end{aligned} \tag{7}$$

Finally the H^1 norm

$$\begin{aligned} \|\cdot\|_{H^1} : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \mathbf{u} &\longmapsto \sqrt{\sum_{i=1}^{n-1} \frac{(u_{i+1}-u_i)^2}{h_{i+\frac12}} + \sum_{i=1}^{n} h_i |u_i|^2}. \end{aligned} \tag{8}$$

1.1 Finite volume formulation

From now on we note $\kappa_{i+\frac12} = \kappa(x_{i+\frac12})$ and $\bar{\mathbf{u}} \in \mathbb{R}^n$ the vector defined by

$$\bar{u}_i = \frac{1}{h_i} \int_{x_{i-\frac12}}^{x_{i+\frac12}} \bar{u}(x)dx.$$

Let $\bar{u} \in \mathcal{C}^k(\Omega)$. The first step to design a finite volume scheme consists in integrating (3) on cell i

$$-\left[\kappa_{i+\frac12}\left(\frac{d\bar{u}}{dx}\right)_{i+\frac12} - \kappa_{i-\frac12}\left(\frac{d\bar{u}}{dx}\right)_{i-\frac12}\right] + \alpha h_i \bar{u}_i = h_i f_i.$$

Thus we need to define the fluxes

$$\bar{\mathcal{F}}_{i+\frac12} = \kappa_{i+\frac12}\left(\frac{d\bar{u}}{dx}\right)_{i+\frac12} \quad \text{and} \quad \bar{\mathcal{F}}_{i-\frac12} = \kappa_{i-\frac12}\left(\frac{d\bar{u}}{dx}\right)_{i-\frac12}.$$

First of all, the Taylor expansion at order k in the neighborhood of $x_{i+\frac12}$ gives

$$\forall x, \quad \bar{u}(x) = \bar{u}(x_{i+\frac12}) + \sum_{\ell=1}^{k} \frac{(x - x_{i+\frac12})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac12}) + \mathcal{O}\left((x - x_{i+\frac12})^{k+1}\right). \tag{9}$$

In order to have mean values as degrees of freedom we integrate (9) from $x_{i+\frac12}$ to $x_{i+\frac32}$ and divide by h_{i+1}

$$\frac{1}{h_{i+1}} \int_{x_{i+\frac12}}^{x_{i+\frac32}} \bar{u}(x)dx = \bar{u}(x_{i+\frac12}) + \frac{1}{h_{i+1}} \sum_{\ell=1}^{k} \int_{x_{i+\frac12}}^{x_{i+\frac32}} \frac{(x - x_{i+\frac12})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac12})dx + \mathcal{O}\left(h_{i+1}^{k+1}\right),$$

that is to say

$$\bar{u}_{i+1} = \bar{u}(x_{i+\frac{1}{2}}) + \frac{1}{h_{i+1}} \sum_{\ell=1}^{k} \left[\frac{(x - x_{i+\frac{1}{2}})^{\ell+1}}{(\ell+1)!} \right]_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h_{i+1}^{k+1}\right),$$

namely

$$\bar{u}_{i+1} = \bar{u}(x_{i+\frac{1}{2}}) + \sum_{\ell=1}^{k} \frac{h_{i+1}^\ell}{(\ell+1)!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h_{i+1}^{k+1}\right).$$

In a similar way, by integrating (9) from $x_{i-\frac{1}{2}}$ to $x_{i+\frac{1}{2}}$ we obtain

$$\bar{u}_{i} = \bar{u}(x_{i+\frac{1}{2}}) + \sum_{\ell=1}^{k} \frac{(-1)^\ell h_{i}^\ell}{(\ell+1)!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h_{i}^{k+1}\right).$$

The difference between these last two equalities gives

$$\bar{u}_{i+1} - \bar{u}_i = h_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + \sum_{\ell=2}^{k} \frac{h_{i+1}^\ell - (-1)^\ell h_i^\ell}{(\ell+1)!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h^{k+1}\right),$$

from which we obtain

$$\frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) = \frac{1}{h_{i+\frac{1}{2}}} \big(\bar{u}_{i+1} - \bar{u}_i - \sum_{\ell=2}^{k} \frac{h_{i+1}^\ell + (-1)^{\ell+1} h_i^\ell}{(\ell+1)!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}})\big) + \mathcal{O}\left(h^{k}\right). \tag{10}$$

Let $\mathbf{u} = (u_i)_{1 \le i \le n}$ be the numerical solution. By mimicking the expression of the exact flux (10) the numerical flux is defined by

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) \right), \tag{11}$$

with

$$r_{i+\frac{1}{2}}(\mathbf{u}) = -\frac{1}{h_{i+\frac{1}{2}}} \sum_{\ell=2}^{k} \frac{h_{i+1}^\ell + (-1)^{\ell+1} h_i^\ell}{(\ell+1)!} \frac{d^\ell P}{dx^\ell}(x_{i+\frac{1}{2}})(x_{i+\frac{1}{2}}), \tag{12}$$

where P is an interpolation polynomial of u as we will see in the next section.

1.2 High-order reconstruction by interpolation

In the calculation of the flux, it is necessary to evaluate the derivatives of u in $x_{i+\frac{1}{2}}$. In this method, the neighboring cells of $x_{i+\frac{1}{2}}$ are used in order to compute the polynomial reconstruction of the solution by considering that the average of the polynomial in a cell is equal to the average of the solution in this cell.

For a polynomial of degree k, there are $k+1$ coefficients to calculate, so $k+1$ neighboring cells of $x_{i+\frac{1}{2}}$ will be necessary. When it is possible, the stencil will be centered in $x_{i+\frac{1}{2}}$, but the closer $x_{i+\frac{1}{2}}$ is to the boundary, the more the stencil will be shifted in order to stay in the interior of Ω.

The notation $u_0, ..., u_k$ denotes the $k+1$ values of $\mathbf{u}$ used for the calculation. Let us denote by $\mathcal{S}_{i+\frac{1}{2}} = \{x_0, ..., x_k\}$ the stencil of the node $x_{i+\frac{1}{2}}$. The polynomial will be of this form

$$P(x) = a_k(u_0, ..., u_k)x^k + ... + a_0(u_0, ..., u_k).$$

The coefficients of the polynomial $P(x)$ are approximated by

$$\frac{1}{x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} P(x)dx = u_j, \qquad \forall j \in \mathcal{S}_{i+\frac{1}{2}}.$$

This leads to the following system

$$\underbrace{\begin{pmatrix} 1 & \frac{1}{x_{0+\frac{1}{2}}-x_{0-\frac{1}{2}}}\int_{x_{0-\frac{1}{2}}}^{x_{0+\frac{1}{2}}} x & \dots & \frac{1}{x_{0+\frac{1}{2}}-x_{0-\frac{1}{2}}}\int_{x_{0-\frac{1}{2}}}^{x_{0+\frac{1}{2}}} x^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{1}{x_{k+\frac{1}{2}}-x_{k-\frac{1}{2}}}\int_{x_{k-\frac{1}{2}}}^{x_{k+\frac{1}{2}}} x & \dots & \frac{1}{x_{k+\frac{1}{2}}-x_{k-\frac{1}{2}}}\int_{x_{k-\frac{1}{2}}}^{x_{k+\frac{1}{2}}} x^k \end{pmatrix}}_{=:M_k} \underbrace{\begin{pmatrix} a_0 \\ \vdots \\ a_k \end{pmatrix}}_{=:\mathbf{a}} = \begin{pmatrix} u_0 \\ \vdots \\ u_k \end{pmatrix}.$$

The matrix M_k can be rewritten

$$M_k = \begin{pmatrix} 1 & \frac{x_{0+\frac{1}{2}}^2-x_{0-\frac{1}{2}}^2}{2(x_{0+\frac{1}{2}}-x_{0-\frac{1}{2}})} & \dots & \frac{x_{0+\frac{1}{2}}^{k+1}-x_{0-\frac{1}{2}}^{k+1}}{(k+1)(x_{0+\frac{1}{2}}-x_{0-\frac{1}{2}})} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{x_{k+\frac{1}{2}}^2-x_{k-\frac{1}{2}}^2}{2(x_{k+\frac{1}{2}}-x_{k-\frac{1}{2}})} & \dots & \frac{x_{k+\frac{1}{2}}^{k+1}-x_{k-\frac{1}{2}}^{k+1}}{(k+1)(x_{k+\frac{1}{2}}-x_{k-\frac{1}{2}})} \end{pmatrix}. \tag{13}$$

Proposition 1. *Let $\{x_i\}_{1\leq i\leq n}$ be a mesh satisfying (4). Let $k \in \mathbb{N}^*$. The matrix M_k defined by (13) is invertible.*

Proof. $M_k\mathbf{a} = \mathbf{0}$ means that the integral of the polynomial $P(x)$ vanishes over $k+1$ distinct intervals. Therefore, this polynomial of degree k has at least $k+1$ roots. It is therefore zero, and all the coefficients $a_j, j \in [\![0,k]\!]$, vanish. Thus, this implies that $\mathbf{a} = \mathbf{0}$, so M_k is invertible. □

The exact derivatives can then be approximated by

$$\frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) \approx \frac{d^\ell P}{dx^\ell}(x_{i+\frac{1}{2}}), \forall \ell \in [\![2,k]\!].$$

Remark 1. *A polynomial P is calculated for each node $x_{i+\frac{1}{2}}$. So, the polynomial $P = P_{i+\frac{1}{2}}$ can be different for each node but in order to simplify the notation, we will denote it P.*

1.3 A method to obtain monotonicity

A method borrowed from [14] can be used to make the scheme monotone. The flux (11) can be rewritten as follows

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}}\left(\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}} + r^+_{i+\frac{1}{2}}(\mathbf{u}) - r^-_{i+\frac{1}{2}}(\mathbf{u})\right),$$

with

$$r^+_{i+\frac{1}{2}}(\mathbf{u}) = \frac{|r_{i+\frac{1}{2}}(\mathbf{u})| + r_{i+\frac{1}{2}}(\mathbf{u})}{2} \geq 0 \quad \text{and} \quad r^-_{i+\frac{1}{2}}(\mathbf{u}) = \frac{|r_{i+\frac{1}{2}}(\mathbf{u})| - r_{i+\frac{1}{2}}(\mathbf{u})}{2} \geq 0.$$

Let us assume that $\mathbf{u} > \mathbf{0}$, the flux then reads as

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}}\left[\left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}}\right)u_{i+1} - \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u})}{u_i}\right)u_i\right], \tag{14}$$

and the coefficients of u_i, u_{i+1} are positive.

1.4 Symmetric version

In order to make the scheme symmetric, a coefficient $s_{i+\frac{1}{2}}$ depending on $\mathbf{u}$ is introduced in the flux, so that $\mathcal{F}_{i+\frac{1}{2}}$ can be written as

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left[\left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) u_{i+1} - \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i \right]. \tag{15}$$

To have a symmetric scheme the coefficients of u_i et u_{i+1} must be equal

$$\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} = \frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i},$$

which leads to

$$s_{i+\frac{1}{2}}(\mathbf{u}) = \frac{u_i r^+_{i+\frac{1}{2}}(\mathbf{u}) - u_{i+1} r^-_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1} - u_i}.$$

To preserve positivity, it is necessary to impose

$$\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} = \frac{1}{h_{i+\frac{1}{2}}} + \frac{r_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1} - u_i} \geq 0,$$

that is to say

$$\frac{\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1} - u_i} \geq 0. \tag{16}$$

In other words, $u_{i+1} - u_i$ and $\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})$, defined by (11), must have the same sign which seems natural because if $\dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) \geq 0$ (resp. ≤ 0), then $\bar{u}$ is locally non-decreasing (resp. non-increasing) hence $\bar{u}_{i+1} \geq \bar{u}_i$ (resp. $\bar{u}_{i+1} \leq \bar{u}_i$).

When this condition is *not* satisfied, we replace the numerical flux (14) by the first order approximation

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} \right). \tag{17}$$

1.5 Boundary conditions

In this section we detail how we take into account the boundary conditions.

1.5.1 Dirichlet boundary condition

Consider problem (3) with $\beta = 1$, $\gamma = 0$ and consider first the right boundary of the domain. The adaptation to the left boundary is straightforward. The k-th order Taylor expansion in the neighborhood of $x_{n+\frac{1}{2}}$ gives

$$\forall x, \quad \bar{u}(x) = \bar{u}(x_{n+\frac{1}{2}}) + \sum_{\ell=1}^{k} (-1)^\ell \frac{(x - x_{n+\frac{1}{2}})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{n+\frac{1}{2}}) + \mathcal{O}\left((x - x_{n+\frac{1}{2}})^{k+1} \right).$$

Here again, we integrate this expression in order to use mean values. This gives

$$\frac{1}{h_n} \int_{x_{n-\frac{1}{2}}}^{x_{n+\frac{1}{2}}} \bar{u}(x) dx = \bar{u}(x_{n+\frac{1}{2}}) + \frac{1}{h_n} \sum_{\ell=1}^{k} \int_{x_{n-\frac{1}{2}}}^{x_{n+\frac{1}{2}}} (-1)^\ell \frac{(x - x_{n+\frac{1}{2}})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{n+\frac{1}{2}}) dx + \mathcal{O}\left(h_n^{k+1} \right),$$

that is to say

$$\bar{u}_n = \bar{u}(x_{n+\frac{1}{2}}) + \frac{1}{h_n}\sum_{\ell=1}^{k}(-1)^\ell \left[\frac{(x-x_{n+\frac{1}{2}})^{\ell+1}}{(\ell+1)!}\right]_{x_{n-\frac{1}{2}}}^{x_{n+\frac{1}{2}}} \frac{d^\ell \bar{u}}{dx^\ell}(x_{n+\frac{1}{2}}) + \mathcal{O}\left(h_n^{k+1}\right),$$

from which we obtain

$$\frac{d\bar{u}}{dx}(x_{n+\frac{1}{2}}) = \frac{2}{h_n}(\bar{u}(x_{n+\frac{1}{2}}) - \bar{u}_n) + 2\sum_{\ell=2}^{k}\frac{h_n^{\ell-1}}{(\ell+1)!}\frac{d^\ell \bar{u}}{dx^\ell}(x_{n+\frac{1}{2}}) + \mathcal{O}\left(h_n^{k+1}\right).$$

The numerical flux is obtained by approximating the derivatives of $\bar{u}$ at $x_{n+\frac{1}{2}}$ using a polynomial reconstruction of the solution

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}}\left(\frac{2}{h_n}(u_{n+\frac{1}{2}} - u_n) + r_{n+\frac{1}{2}}(\mathbf{u})\right).$$

The trick of Section 1.3 can be applied to ensure monotonicity, that is in the non-symmetric version

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}}\left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}}\right)u_{n+\frac{1}{2}} - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right)u_n\right],$$

and, in the symmetric version

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}}\left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}}\right)u_{n+\frac{1}{2}} - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right)u_n\right], \tag{18}$$

with

$$s_{n+\frac{1}{2}}(\mathbf{u}) = \frac{u_n r^+_{n+\frac{1}{2}}(\mathbf{u}) - u_{n+\frac{1}{2}} r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}} - u_n}.$$

In order to preserve positivity, a condition similar to (16) must be satisfied for the symmetric version of the scheme

$$\frac{\frac{2}{h_n}(u_{n+\frac{1}{2}} - u_n) + r_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}} - u_n} \geq 0,$$

that is to say that $u_{n+\frac{1}{2}} - u_n$ and $\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u})$ must have the same sign. As above, this condition seems natural because if $\dfrac{d\bar{u}}{dx}(x_{n+\frac{1}{2}}) \geq 0$ (resp. ≤ 0), then $\bar{u}$ is locally increasing (resp. decreasing) so $\bar{u}_{n+\frac{1}{2}} \geq \bar{u}_n$ (resp. $\bar{u}_{n+\frac{1}{2}} \leq \bar{u}_n$).

Applying the boundary condition, (18) becomes

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}}\left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})}\right)g(x_{n+\frac{1}{2}}) - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right)u_n\right]. \tag{19}$$

For the left boundary we obtain similarly

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}}\left[\left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1}\right)u_1 - \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})}\right)g(x_{\frac{1}{2}})\right]. \tag{20}$$

1.5.2 Neumann boundary condition

Consider problem (3) with $\beta = 0$, $\gamma = 1$. For the left ($i = 1$) boundary cell, the flux is

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}} \left.\frac{d\bar{u}}{dn}\right|_{\frac{1}{2}} = -\kappa_{\frac{1}{2}} \left.\frac{d\bar{u}}{dx}\right|_{\frac{1}{2}} = g(x_{\frac{1}{2}}) \tag{21}$$

while for the right ($i = n$) boundary cell, the flux is

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left.\frac{d\bar{u}}{dn}\right|_{n+\frac{1}{2}} = \kappa_{n+\frac{1}{2}} \left.\frac{d\bar{u}}{dx}\right|_{n+\frac{1}{2}} = g(x_{n+\frac{1}{2}}). \tag{22}$$

1.5.3 Mixed boundary condition

Consider finally problem (3) with $\beta(x) > 0, \gamma(x) > 0, \forall x \in \partial\Omega$. In this case we have for $i = 0$ or $i = n$

$$\bar{u}(x_{i+\frac{1}{2}}) = \frac{1}{\beta(x_{i+\frac{1}{2}})} \left(g(x_{i+\frac{1}{2}}) - \gamma(x_{i+\frac{1}{2}})\kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dn}(x_{i+\frac{1}{2}}) \right). \tag{23}$$

Consider first the right boundary of the domain. The adaptation for the left boundary is straightforward. We use the same method as for Dirichlet boundary condition in section 1.5.1. Replacing $u_{n+\frac{1}{2}}$ by its expression given by (23) in (18) yields

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \frac{\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right) g(x_{n+\frac{1}{2}}) - \beta(x_{n+\frac{1}{2}})\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) u_n}{\beta(x_{n+\frac{1}{2}}) + \gamma(x_{n+\frac{1}{2}})\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)}. \tag{24}$$

For the left boundary ($i = 0$) we obtain similarly

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \frac{\beta(x_{\frac{1}{2}})\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) u_1 - \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right) g(x_{\frac{1}{2}})}{\beta(x_{\frac{1}{2}}) + \gamma(x_{\frac{1}{2}})\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)}. \tag{25}$$

Remark 2. *In the expression of the fluxes* (25) *and* (24)*, if we take $\beta = 0$, $\gamma = 1$, we obtain the same fluxes as* (21) *and* (22)*. Likewise, if we take $\beta = 1, \gamma = 0$, we obtain the same flux as* (20) *and* (19)*.*

1.6 Summary of the method and matrix form

The scheme reads as

$$-(\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) + \alpha h_i u_i = h_i f_i, \tag{26}$$

that is, using (15),

$$\begin{aligned} &- \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) u_{i+1} + \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i \\ &+ \kappa_{i-\frac{1}{2}} \left(\frac{1}{h_{i-\frac{1}{2}}} + \frac{r^+_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i - \kappa_{i-\frac{1}{2}} \left(\frac{1}{h_{i-\frac{1}{2}}} + \frac{r^-_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_{i-1}} \right) u_{i-1} + \alpha h_i u_i = h_i f_i. \end{aligned}$$

With a more compact notation, we write this as $A\mathbf{u} = A(\mathbf{u})\mathbf{u} = \mathbf{b}$, with

$$b_i = h_i f_i \quad \forall i \neq \{1, n\},$$

$$A_{ij} = \begin{cases} -\kappa_{i+\frac{1}{2}} \left(\dfrac{1}{h_{i+\frac{1}{2}}} + \dfrac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) & \text{if } j = i+1, \forall i \neq n, \\ \kappa_{i+\frac{1}{2}} \left(\dfrac{1}{h_{i+\frac{1}{2}}} + \dfrac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) + \kappa_{i-\frac{1}{2}} \left(\dfrac{1}{h_{i-\frac{1}{2}}} + \dfrac{r^+_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_i} \right) + \alpha h_i & \text{if } j = i, \forall i \neq 1, n, \\ -\kappa_{i-\frac{1}{2}} \left(\dfrac{1}{h_{i-\frac{1}{2}}} + \dfrac{r^-_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_{i-1}} \right) & \text{if } j = i-1, \forall i \neq 1, \\ 0 & \text{else.} \end{cases} \tag{27}$$

The expression of the boundary terms depends on the type of boundary conditions. First, in the case of a Dirichlet boundary condition, we have

$$b_1 = h_1 f_1 + \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})} \right) g(x_{\frac{1}{2}}), \tag{28}$$

$$\begin{cases} A_{1,1} = \kappa_{\frac{3}{2}} \left(\dfrac{1}{h_{\frac{3}{2}}} + \dfrac{r^-_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \kappa_{\frac{1}{2}} \left(\dfrac{2}{h_1} + \dfrac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1, \\ A_{1,2} = -\kappa_{\frac{3}{2}} \left(\dfrac{1}{h_{\frac{3}{2}}} + \dfrac{r^+_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_2} \right), \end{cases} \tag{29}$$

and

$$b_n = h_n f_n + \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})} \right) g(x_{n+\frac{1}{2}}), \tag{30}$$

$$\begin{cases} A_{n,n} = \kappa_{n+\frac{1}{2}} \left(\dfrac{2}{h_n} + \dfrac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \kappa_{n-\frac{1}{2}} \left(\dfrac{1}{h_{n-\frac{1}{2}}} + \dfrac{r^+_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n, \\ A_{n,n-1} = -\kappa_{n-\frac{1}{2}} \left(\dfrac{1}{h_{n-\frac{1}{2}}} + \dfrac{r^-_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_{n-1}} \right). \end{cases} \tag{31}$$

Next, in the case of a Neumann boundary condition, we have

$$b_1 = h_1 f_1 + g(x_{\frac{1}{2}}), \tag{32}$$

$$\begin{cases} A_{1,1} = \kappa_{\frac{3}{2}} \left(\dfrac{1}{h_{\frac{3}{2}}} + \dfrac{r^-_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1, \\ A_{1,2} = -\kappa_{\frac{3}{2}} \left(\dfrac{1}{h_{\frac{3}{2}}} + \dfrac{r^+_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_2} \right), \end{cases} \tag{33}$$

and

$$b_n = h_n f_n + g(x_{n+\frac{1}{2}}), \tag{34}$$

$$\begin{cases} A_{n,n} = \kappa_{n-\frac{1}{2}} \left(\dfrac{1}{h_{n-\frac{1}{2}}} + \dfrac{r^+_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n, \\ A_{n,n-1} = -\kappa_{n-\frac{1}{2}} \left(\dfrac{1}{h_{n-\frac{1}{2}}} + \dfrac{r^-_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_{n-1}} \right). \end{cases} \tag{35}$$

Finally, in the case of a mixed boundary condition, we have

$$b_1 = h_1 f_1 + \frac{\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)}{\beta(x_{\frac{1}{2}}) + \gamma(x_{\frac{1}{2}})\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)} g(x_{\frac{1}{2}}), \tag{36}$$

$$\begin{cases} A_{1,1} = \kappa_{\frac{3}{2}} \left(\dfrac{1}{h_{\frac{3}{2}}} + \dfrac{r^-_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \dfrac{\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right)}{1 + \frac{\gamma(x_{\frac{1}{2}})\kappa_{\frac{1}{2}}}{\beta(x_{\frac{1}{2}})} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)} + \alpha h_1, \\ A_{1,2} = -\kappa_{\frac{3}{2}} \left(\dfrac{1}{h_{\frac{3}{2}}} + \dfrac{r^+_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_2} \right), \end{cases} \tag{37}$$

and

$$b_n = h_n f_n + \frac{\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)}{\beta(x_{n+\frac{1}{2}}) + \gamma(x_{n+\frac{1}{2}})\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)} g(x_{n+\frac{1}{2}}), \tag{38}$$

$$\begin{cases} A_{n,n} = \kappa_{n-\frac{1}{2}} \left(\dfrac{1}{h_{n-\frac{1}{2}}} + \dfrac{r^+_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \dfrac{\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right)}{1 + \frac{\gamma(x_{n+\frac{1}{2}})\kappa_{n+\frac{1}{2}}}{\beta(x_{n+\frac{1}{2}})} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)} + \alpha h_n, \\ A_{n,n-1} = -\kappa_{n-\frac{1}{2}} \left(\dfrac{1}{h_{n-\frac{1}{2}}} + \dfrac{r^-_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_{n-1}} \right). \end{cases} \tag{39}$$

The matrix has been written for the symmetric version of the scheme. For the non-symmetric version, the matrix is the same with $s_{i+\frac{1}{2}}(\mathbf{u}) = s_{i-\frac{1}{2}}(\mathbf{u}) = 0, \forall i \in [\![1, n]\!]$.

Remark 3. *Assuming that $f \geq 0$ and $g \geq 0$, and that $\mathbf{u} > \mathbf{0}$, the right hand side $\mathbf{b}$ has all its components nonnegative, for any type of boundary conditions.*

2 Properties

2.1 Conservation

Proposition 2. *Assume that $\mathbf{u} > \mathbf{0}$ and consider homogeneous Neumann boundary conditions, then the scheme defined by* (26) *is conservative. Indeed it satisfies the equality*

$$\alpha \sum_{i=1}^{n} h_i u_i = \sum_{i=1}^{n} h_i f_i,$$

that is to say

$$\sum_{i=1}^{n} (-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) = 0.$$

Proof. The sum is telescopic so only the boundary terms remain. The homogeneous Neumann boundary condition means that the boundary terms are zero, which leads to

$$\sum_{i=1}^{n} (-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) = 0,$$

that is to say

$$\alpha \sum_{i=1}^{n} h_i u_i = \sum_{i=1}^{n} h_i f_i.$$

The scheme is conservative. □

2.2 Monotonicity and Local Maximum Principle (LMP) structure

Definition 1. *A matrix $A = (a_{ij})$ is an M-matrix if it satisfies the following inequalities*

$$\forall i \neq j, a_{ij} \leq 0,$$

and

$$\forall i, \sum_{j=1}^{n} a_{i,j} \geq 0. \tag{40}$$

Moreover, if (40) is strict for all $i \in [\![1, n]\!]$, we say that A is a strict M-matrix.

2.2.1 Non-symmetric version: monotonicity

Proposition 3. *Assume that $\mathbf{u} > \mathbf{0}$, the matrix A defined by (27) and (28) through (31), or (32) through (35), or (36) through (39) depending on the boundary conditions, with $s_{i+\frac{1}{2}} = 0$, is such that A^t is a strict M-matrix.*

Remark 4. *In the following proof we have considered Dirichlet boundary conditions, but the result also holds with other boundary conditions. For mixed boundary conditions, the sum of the first and the last column have also two positive terms. For Neumann boundary conditions, the sum of the first and the last column are also positive but the first term vanishes, that is to say $\sum_i A_{i,1} = \alpha h_1 > 0$ and $\sum_i A_{i,n} = \alpha h_n > 0$.*

Proof of Proposition 3. The matrix satisfies

$$\forall i \neq j, A_{ij} \leq 0 \quad \text{and} \quad \forall j, \sum_{i=1}^{n} A_{i,j} > 0.$$

Indeed, for the first column there are only two elements in the sum

$$\sum_i A_{i,1} = A_{1,1} + A_{2,1},$$

which leads to

$$\sum_i A_{i,1} = \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) - \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1,$$

that is to say

$$\sum_i A_{i,1} = \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1 > 0.$$

And for the last column,

$$\sum_i A_{i,n} = A_{n-1,n} + A_{n,n},$$

which leads to

$$\sum_i A_{i,n} = -\kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n,$$

that is to say

$$\sum_i A_{i,n} = \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n > 0.$$

Besides, for other columns

$$\sum_i A_{i,j} = A_{j-1,j} + A_{j,j} + A_{j+1,j},$$

which leads to

$$\begin{aligned} \sum_i A_{i,j} = -\kappa_{(j-1)+\frac{1}{2}} \left(\frac{1}{h_{(j-1)+\frac{1}{2}}} + \frac{r^+_{(j-1)+\frac{1}{2}}(\mathbf{u})}{u_{(j-1)+1}} \right) + \kappa_{j+\frac{1}{2}} \left(\frac{1}{h_{j+\frac{1}{2}}} + \frac{r^-_{j+\frac{1}{2}}(\mathbf{u})}{u_j} \right) + \alpha h_j \\ + \kappa_{j-\frac{1}{2}} \left(\frac{1}{h_{j-\frac{1}{2}}} + \frac{r^+_{j-\frac{1}{2}}(\mathbf{u})}{u_j} \right) - \kappa_{(j+1)-\frac{1}{2}} \left(\frac{1}{h_{(j+1)-\frac{1}{2}}} + \frac{r^-_{(j+1)-\frac{1}{2}}(\mathbf{u})}{u_{(j+1)-1}} \right), \end{aligned}$$

that is to say

$$\sum_i A_{i,j} = \alpha h_j > 0.$$

□

Theorem 1. *Assume that $f > 0$ and $g > 0$. Let A and $\mathbf{b}$ be defined by (27) and (28) through (31), or (32) through (35), or (36) through (39), depending on the boundary conditions, with $s_{i+\frac{1}{2}} = 0, \forall i$. Then $A^{-1}\mathbf{b} = \mathbf{u} \geq \mathbf{0}$.*

Proof. As A^t is a strict M-matrix A is invertible and its inverse has only positive entries (see for example [22], Corollary 3.20). In view of Remark 3, the right hand side is nonnegative, hence $\mathbf{u} = A^{-1}\mathbf{b} \geq \mathbf{0}$.

□

2.2.2 Symmetric version: LMP structure

Proposition 4. *Assume that $\mathbf{u} > \mathbf{0}$, the matrix A defined by (27) and (28) through (31), or (32) through (35), or (36) through (39), depending on the boundary conditions, is symmetric.*

Proof. Let $x_{i+\frac{1}{2}}$, be an interior vertex of the mesh. If condition (16) is satisfied for this vertex, we use the definition of the flux (15), then symmetrization condition leads to $A_{i,i+1} = A_{i+1,i}$. Else, the flux is defined by (17), and once again $A_{i,i+1} = A_{i+1,i}$.

□

Proposition 5. *Assume that* $\mathbf{u} > \mathbf{0}$*, let* A *be defined by (27) and (28) through (31), or (32) through (35), or (36) through (39), depending on the boundary conditions, then the matrix* A *is a strict M-matrix.*

Proof. As for Proposition 3, it can be proved that the matrix A is the transpose of a strict M-matrix. Besides, A is symmetric, so A is itself a strict M-matrix. □

Definition 2. *This definition is taken from [9]. We say that a scheme for (3) has the local maximum principle structure (LMP structure for short) if it can be written in the form*

$$\forall i \in [\![1,n]\!] : \sum_{j=1}^{n} \lambda_{i,j}(\mathbf{u})(u_i - u_j) + \lambda_{i,0}(\mathbf{u})(u_i - u_{\frac{1}{2}}) + \lambda_{i,n+1}(\mathbf{u})(u_i - u_{n+\frac{1}{2}}) = f_i h_i, \tag{41}$$

for some functions $\lambda_{i,j} : \mathbb{R}^n \to \mathbb{R}^+$ *satisfying,*

$$\lambda_{1,0} > 0, \lambda_{n,n+1} > 0, \text{ and } \forall i \in [\![1,n-1]\!] : \lambda_{i,i+1} > 0. \tag{42}$$

Theorem 2. *Assume that* $f > 0$*, and* $g > 0$*. Let* A *and* $\mathbf{b}$ *be defined by (27) and (28) through (31), or (32) through (35), or (36) through (39), depending on the boundary conditions. Assume that we have applied the symmetrization procedure defined in Section 1.4. Then* $A^{-1}\mathbf{b} = \mathbf{u} \geq 0$*. If moreover* $\alpha = 0$*, the scheme has the LMP structure.*

2.3 Consistency of the fluxes

Proposition 6. *Let* $k \in \mathbb{N}^*$ *and* $\{x_i\}_{1 \leq i \leq n}$ *be a mesh satisfying* $x_i < x_{i+1}$*,* $\forall i \in [\![1,n-1]\!]$*. Let* $\bar{u} \in \mathcal{C}^k(\Omega)$ *be the exact solution of (1). The fluxes defined by (11) are consistent to order* k*, that is to say*

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + \mathcal{O}(h^k).$$

Proof. Let $\bar{u} \in \mathcal{C}^k(\Omega)$ the exact solution, a Taylor expansion gives

$$\bar{u}(x) = \sum_{\ell=0}^{k} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) \frac{(x - x_{i+\frac{1}{2}})^\ell}{\ell!} + \mathcal{O}\left((x - x_{i+\frac{1}{2}})^{k+1}\right) = P(x) + \mathcal{O}\left((x - x_{i+\frac{1}{2}})^{k+1}\right),$$

where P is the k-th order polynomial

$$P(x) = \sum_{\ell=0}^{k} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) \frac{(x - x_{i+\frac{1}{2}})^\ell}{\ell!},$$

such that

$$\frac{d^\ell P}{dx^\ell}(x_{i+\frac{1}{2}}) = \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}), \quad \forall \ell \in [\![1,k]\!]. \tag{43}$$

Applying our expression of the flux to $\bar{\mathbf{u}}$ gives

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} \left[\bar{u}(x_{i+1}) - \bar{u}(x_i) \right] + r_{i+\frac{1}{2}}(\bar{\mathbf{u}}) \right),$$

with the following expression of the remainder

$$r_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = -\frac{1}{h_{i+\frac{1}{2}}} \sum_{\ell=2}^{k} \frac{h_{i+1}^\ell + (-1)^{\ell+1} h_i^\ell}{(\ell+1)!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}).$$

Thus Equation (43) leads to

$$r_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = -\frac{1}{h_{i+\frac{1}{2}}} \sum_{\ell=2}^{k} \frac{h_{i+1}^\ell + (-1)^{\ell+1} h_i^\ell}{(\ell+1)!} \frac{d^\ell P}{dx^\ell}(x_{i+\frac{1}{2}}),$$

that is to say $r_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = r_{i+\frac{1}{2}}(\mathbf{p})$ with $\mathbf{p} = (p_i)$ defined by

$$p_i = \frac{1}{x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} P(x)dx.$$

So the flux becomes

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} \left[P(x_{i+1}) + \mathcal{O}\left((x_{i+1} - x_{i+\frac{1}{2}})^{k+1} \right) - P(x_i) - \mathcal{O}\left((x_i - x_{i+\frac{1}{2}})^{k+1} \right) \right] + r_{i+\frac{1}{2}}(\mathbf{p}) \right).$$

We prove in Appendix A that the flux is exact for polynomials of degree k, that is to say

$$\frac{1}{h_{i+\frac{1}{2}}}(P(x_{i+1}) - P(x_i)) + r_{i+\frac{1}{2}}(\mathbf{p}) = \frac{dP}{dx}(x_{i+\frac{1}{2}}),$$

which leads to

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \kappa_{i+\frac{1}{2}} \frac{dP}{dx}(x_{i+\frac{1}{2}}) + \mathcal{O}(h^k),$$

and finally

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + \mathcal{O}(h^k).$$

The fluxes are consistent to order k. □

Remark 5. *This proposition can be extended to the boundary fluxes. Indeed, for a Neumann boundary condition, the consistency is obvious and for Dirichlet or mixed boundary conditions, the proof is similar.*

2.4 Convergence of the scheme at order k

Consider again problem (3) with $\beta = 0$, $\gamma = 1$,

$$\begin{cases} -\dfrac{d}{dx}\left(\kappa \dfrac{d\bar{u}}{dx}\right) + \alpha \bar{u} = f & \text{in } \Omega, \\ \kappa \dfrac{d\bar{u}}{dn} = 0 & \text{on } \partial\Omega. \end{cases} \tag{44}$$

We will start by proving that the scheme is convergent at order $k-1$ in L^1 norm. Next, this will allow us to prove the convergence of the fluxes at order $k-1$ in L^2 norm. Then, we will use these two results to show that the remainder of the scheme is $\mathcal{O}(h^{\frac{1}{2}})$. Finally, owing to these results, we will be able to prove that the scheme is convergent at order k in the H^1 norm defined in (8).

2.4.1 Convergence at the order $k-1$

The scheme reads as

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u}) + \alpha h_i u_i = h_i f_i, \quad \forall i \in [\![1, n]\!], \tag{45}$$

with $\forall i \in [\![1, n-1]\!]$,

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) \right) = \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) u_{i+1} - \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i, \tag{46}$$

and

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = 0. \tag{47}$$

Proposition 7 (Convergence at order $k-1$ in L^1 norm)**.** *Let $k \in \mathbb{N}^*$, $\bar{u} \in \mathcal{C}^k(\Omega)$ be the exact solution of (44) and assume that $\bar{\mathbf{u}} \geq \mathbf{0}$. Let $\mathbf{e} = (\bar{u}_i - u_i)_{1 \leq i \leq n}$, where $\mathbf{u}$ is the solution of the scheme (45)-(46)-(47) and assume that $\mathbf{u} > \mathbf{0}$. Then,*

$$\|\mathbf{e}\|_{L^1} \leq Ch^{k-1},$$

with $\|\cdot\|_{L^1}$ defined by (6).

Proof. On the one hand the numerical flux defined by (46) satisfies (45) and on the other hand, the exact flux $\bar{\mathcal{F}}_{i+\frac{1}{2}} = \kappa_{i+\frac{1}{2}} \dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}})$ satisfies

$$-\bar{\mathcal{F}}_{i+\frac{1}{2}} + \bar{\mathcal{F}}_{i-\frac{1}{2}} + \alpha h_i \bar{u}_i = h_i f_i, \qquad \forall i \in [\![1, n]\!].$$

Subtracting (45) from this equation gives

$$-(\bar{\mathcal{F}}_{i+\frac{1}{2}} - \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})) + (\bar{\mathcal{F}}_{i-\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) + \alpha h_i(\bar{u}_i - u_i) = 0, \qquad \forall i \in [\![1, n]\!].$$

Besides, the consistency of the fluxes gives that there exists a constant $C > 0$ such as

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \bar{\mathcal{F}}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}}, \qquad \forall i \in [\![1, n,]\!] \quad \text{with } |R_{i+\frac{1}{2}}| \leq Ch^k, \text{ where } k \text{ is the order.} \tag{48}$$

These last two equations imply

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{e}) + \alpha h_i e_i = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}, \qquad \forall i \in [\![1, n]\!].$$

By choosing $\Delta \in \mathbb{R}^+$ such that

$$-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta \geq 0, \qquad \forall i \in [\![1, n]\!],$$

and adding it to e_i leads to

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{e}) + \alpha h_i(e_i + \Delta) = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta \geq 0, \qquad \forall i \in [\![1, n]\!].$$

The flux is not modified since the remainder only involves derivatives (Δ being a constant, it no longer appears in the derivatives)

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e} + \boldsymbol{\Delta}) = \kappa_{i+\frac{1}{2}} \left(\frac{e_{i+1} + \Delta - e_i - \Delta}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{e}) \right) = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}), \qquad \forall i \in [\![1, n]\!].$$

The corresponding matrix system writes

$$A(\mathbf{e} + \boldsymbol{\Delta})(\mathbf{e} + \boldsymbol{\Delta}) = \mathbf{R} + \alpha \mathbf{h} \boldsymbol{\Delta},$$

with

$$(\mathbf{e} + \boldsymbol{\Delta})_i = e_i + \Delta, \qquad (\mathbf{R} + \alpha \mathbf{h} \boldsymbol{\Delta})_i = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta \geq 0, \qquad \forall i \in [\![1, n]\!].$$

The right hand side being nonnegative, Theorem 1 applied to $\mathbf{e} + \boldsymbol{\Delta}$ assures us that $\mathbf{e} + \boldsymbol{\Delta} > \mathbf{0}$ and Proposition 2 applied to $\mathbf{e} + \boldsymbol{\Delta}$ leads to

$$\sum_{i=1}^{n} \alpha h_i (e_i + \Delta) = \sum_{i=1}^{n} \left(-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta \right). \tag{49}$$

So the L^1-norm of the error can be rewritten as

$$\sum_{i=1}^{n} h_i |e_i| = \sum_{i=1}^{n} h_i |e_i + \Delta - \Delta| \leq \sum_{i=1}^{n} h_i |e_i + \Delta| + \Delta \sum_{i=1}^{n} h_i.$$

Since $\displaystyle\sum_{i=1}^{n} h_i = 1$ and $e_i + \Delta \geq 0$, (49) leads to

$$\sum_{i=1}^{n} h_i|e_i| \leq \frac{1}{\alpha}\sum_{i=1}^{n}\left(-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}\right) + 2\Delta.$$

Hence, inequality (48) gives

$$\sum_{i=1}^{n} h_i|e_i| \leq C\sum_{i=1}^{n} h^k + 2\Delta.$$

Choosing for Δ the smallest value such that $-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta \geq 0$, that is, $\Delta = \frac{1}{\alpha}\max\limits_{1\leq i\leq n}\left(\frac{R_{i+\frac{1}{2}} - R_{i-\frac{1}{2}}}{h_i}\right)$, and using (48) lead to

$$\|\mathbf{e}\|_{L^1} = \sum_{i=1}^{n} h_i|e_i| \leq nCh^k + 2Ch^{k-1} = Ch^{k-1}.$$

So, the scheme converges at order $k-1$. □

2.4.2 Convergence of the fluxes

Let us denote by $H_M = \{(u_i)_{1\leq i\leq n}\}$ the set of cell values, $H_E = \{(f_{i+\frac{1}{2}})_{1\leq i\leq n-1}\}$ the set of node values and consider homogeneous Neumann boundary conditions, that is, for all $\mathbf{f} \in H_E$

$$f_{\frac{1}{2}} = f_{n+\frac{1}{2}} = 0. \tag{50}$$

Let us define the scalar products

$$\begin{cases} (\mathbf{u}|\mathbf{v})_{H_M} = \displaystyle\sum_{i=1}^{n} h_i u_i v_i, \\ (\mathbf{f}|\mathbf{g})_{H_E} = \displaystyle\sum_{i=1}^{n-1} h_{i+\frac{1}{2}} f_{i+\frac{1}{2}} g_{i+\frac{1}{2}}, \end{cases} \tag{51}$$

and the operators

$$\begin{cases} D : H_M \longrightarrow H_E \text{ defined by} & (D\mathbf{u})_{i+\frac{1}{2}} = \dfrac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}, \qquad 1 \leq i \leq n-1, \\ D^* : H_E \longrightarrow H_M \text{ defined by} & (D^*\mathbf{f})_i = -\dfrac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{h_i}, \quad 1 \leq i \leq n. \end{cases}$$

Proposition 8. *If condition (50) is satisfied the operators D and D^* are adjoints of each other, that is to say that $(D\mathbf{u}|\mathbf{f})_{H_E} = (\mathbf{u}|D^*\mathbf{f})_{H_M}$, $\forall \mathbf{u} \in H_M$, $\forall \mathbf{f} \in H_E$.*

Proof. The definition of the scalar product gives

$$(D\mathbf{u}|\mathbf{f})_{H_E} = \sum_{i=1}^{n-1} h_{i+\frac{1}{2}} (D\mathbf{u})_{i+\frac{1}{2}} f_{i+\frac{1}{2}},$$

which means

$$(D\mathbf{u}|\mathbf{f})_{H_E} = \sum_{i=1}^{n-1} (u_{i+1} - u_i) f_{i+\frac{1}{2}}.$$

The two sums can be separated

$$(D\mathbf{u}|\mathbf{f})_{H_E} = \sum_{i=1}^{n-1} u_{i+1} f_{i+\frac{1}{2}} - \sum_{i=1}^{n-1} u_i f_{i+\frac{1}{2}}.$$

We shift the index of the first sum, which gives

$$(D\mathbf{u}|\mathbf{f})_{H_E} = \sum_{i=2}^{n} u_i f_{i-\frac{1}{2}} - \sum_{i=1}^{n-1} u_i f_{i+\frac{1}{2}}.$$

Then, the sums can be recombined as follows

$$(D\mathbf{u}|\mathbf{f})_{H_E} = u_n f_{n-\frac{1}{2}} - u_1 f_{\frac{3}{2}} - \sum_{i=2}^{n-1} u_i (f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}).$$

Condition (50) allows us to insert the boundary terms which are zero

$$(D\mathbf{u}|\mathbf{f})_{H_E} = u_n (f_{n-\frac{1}{2}} - f_{n+\frac{1}{2}}) - u_1 (f_{\frac{3}{2}} - f_{\frac{1}{2}}) - \sum_{i=2}^{n-1} u_i (f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}) = -\sum_{i=1}^{n} u_i (f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}) = (\mathbf{u}, D^*\mathbf{f})_{H_M}.$$

Thus, the operators D^* and D are adjoints of each other.

□

Proposition 9 (Convergence of the fluxes at order $k-1$)**.** *Let $k \in \mathbb{N}^*$, $\bar{u} \in \mathcal{C}^k(\Omega)$ be the exact solution of (44) and assume that $\bar{u} \geq 0$. Let us denote $\mathbf{r}(\mathbf{e}) \in H_E$ the vector whose components are $r_{i+\frac{1}{2}}(\mathbf{e}), \forall i \in [\![0, n]\!]$ the remainders defined by (12) and the vector $\mathbf{e} \in H_M$ defined by $e_i = \bar{u}_i - u_i, \forall i \in [\![1, n]\!]$. Assume that $u_i > 0, \forall i \in [\![1, n]\!]$. Then we have*

$$\|\boldsymbol{\mathcal{F}}(\mathbf{u}) - \bar{\boldsymbol{\mathcal{F}}}\|_{H_E} \leq C h^{k-1},$$

where $\boldsymbol{\mathcal{F}}(\mathbf{u}) \in H_E$ is defined by $(\boldsymbol{\mathcal{F}}(\mathbf{u}))_{i+\frac{1}{2}} = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}), \forall i \in [\![0, n]\!]$, with $\mathcal{F}_{i+\frac{1}{2}}$ given by (46) and (47), and $\bar{\boldsymbol{\mathcal{F}}}$ is defined by $(\bar{\boldsymbol{\mathcal{F}}})_{i+\frac{1}{2}} = \bar{\mathcal{F}}_{i+\frac{1}{2}}$, with $\bar{\mathcal{F}}_{i+\frac{1}{2}} = \kappa_{i+\frac{1}{2}} \dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}), \forall i \in [\![0, n]\!]$.

Proof. The scheme

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u}) + \alpha h_i u_i = h_i f_i, \qquad \forall i \in [\![1, n]\!],$$

can be written as

$$D^* \boldsymbol{\kappa} (D\mathbf{u} + \mathbf{r}(\mathbf{u})) + \alpha \mathbf{u} = \mathbf{f}.$$

Besides, the exact flux $\bar{\mathcal{F}}_{i+\frac{1}{2}} = \kappa_{i+\frac{1}{2}} \dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}), \forall i \in [\![1, n]\!]$ also satisfies

$$-\bar{\mathcal{F}}_{i+\frac{1}{2}} + \bar{\mathcal{F}}_{i-\frac{1}{2}} + \alpha h_i \bar{u}_i = h_i f_i, \qquad \forall i \in [\![1, n]\!].$$

Since the fluxes are consistent there exists C such that

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \bar{\mathcal{F}}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}}, \qquad \text{with } |R_{i+\frac{1}{2}}| \leq C h^k, \qquad \forall i \in [\![1, n]\!]. \tag{52}$$

Thus, we have

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{e}) + \alpha h_i u_i = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}, \qquad \forall i \in [\![1, n]\!],$$

that can be written

$$D^* \boldsymbol{\kappa} (D\mathbf{e} + \mathbf{r}(\mathbf{e})) + \alpha \mathbf{e} = D^* \mathbf{R}.$$

Given $\mathbf{v} \in H_M$, we take the scalar product of this equation with $\mathbf{v}$

$$(D^* \boldsymbol{\kappa} (D\mathbf{e} + \mathbf{r}(\mathbf{e}))|\mathbf{v})_{H_M} + (\alpha \mathbf{e}|\mathbf{v})_{H_M} = (D^* \mathbf{R}|\mathbf{v})_{H_M},$$

that is to say

$$(D^* (\boldsymbol{\kappa} (D\mathbf{e} + \mathbf{r}(\mathbf{e})) - \mathbf{R})|\mathbf{v})_{H_M} + (\alpha \mathbf{e}|\mathbf{v})_{H_M} = 0.$$

Besides $\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))-\mathbf{R}$ can be rewritten as

$$\kappa_{i+\frac{1}{2}}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))_{i+\frac{1}{2}}-R_{i+\frac{1}{2}}=-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})+\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}})-R_{i+\frac{1}{2}}=-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})+\bar{\mathcal{F}}_{i+\frac{1}{2}},\qquad\forall i\in[\![1,n]\!],$$

and $\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})$ and $\bar{\mathcal{F}}_{i+\frac{1}{2}}$ satisfy (50), so $\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))-\mathbf{R}$ satisfies (50) too.

Using Proposition 8 provides

$$(\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))|D\mathbf{v})_{H_E}+(\alpha\mathbf{e}|\mathbf{v})_{H_M}=(\mathbf{R}|D\mathbf{v})_{H_E}.$$

We define $\mathbf{v}\in H_M$ by induction as follow

$$\begin{cases} v_1=0, \\ \\ v_{i+1}=h_{i+\frac{1}{2}}\kappa_{i+\frac{1}{2}}\left(\dfrac{e_{i+1}-e_i}{h_{i+\frac{1}{2}}}+r_{i+\frac{1}{2}}\right)+v_i \qquad\qquad \forall i\in[\![1,n-1]\!], \end{cases}$$

whence $D\mathbf{v}=\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))$. We thus have

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}^2+(\alpha\mathbf{e}|\mathbf{v})_{H_M}=(\mathbf{R}|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e})))_{H_E}.$$

The Cauchy-Schwarz inequality leads to

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}^2+(\alpha\mathbf{e}|\mathbf{v})_{H_M}\leq\|\mathbf{R}\|_{H_E}\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}. \tag{53}$$

Besides, we have

$$(\alpha\mathbf{e}|\mathbf{v})_{H_M}=\alpha\sum_{i=1}^{n}h_ie_iv_i.$$

Replacing v_i by its expression leads to

$$(\alpha\mathbf{e}|\mathbf{v})_{H_M}=\alpha\sum_{i=1}^{n}h_ie_i\sum_{j=1}^{i-1}h_{j+\frac{1}{2}}\kappa_{j+\frac{1}{2}}\left(\frac{e_{j+1}-e_j}{h_{j+\frac{1}{2}}}+r_{j+\frac{1}{2}}\right).$$

The Cauchy-Schwarz inequality gives

$$|(\alpha\mathbf{e}|\mathbf{v})_{H_M}|\leq\alpha\sum_{i=1}^{n}h_i|e_i|\left(\sum_{j=1}^{i-1}h_{j+\frac{1}{2}}\left(\kappa_{j+\frac{1}{2}}\left(\frac{e_{j+1}-e_j}{h_{j+\frac{1}{2}}}+r_{j+\frac{1}{2}}\right)\right)^2\right)^{1/2}\left(\sum_{j=1}^{i-1}h_{j+\frac{1}{2}}\right)^{1/2},$$

hence

$$|(\alpha\mathbf{e}|\mathbf{v})_{H_M}|\leq\alpha\left(\sum_{i=1}^{n}h_i|e_i|\right)\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}.$$

Inserting this estimate into (53), we have

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}^2\leq\alpha\left(\sum_{i=1}^{n}h_i|e_i|\right)\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}+\|\mathbf{R}\|_{H_E}\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E},$$

hence

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}\leq\|\mathbf{R}\|_{H_E}+\alpha\sum_{i=1}^{n}h_i|e_i|.$$

Equation (52) gives

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E} \leq Ch^k + \alpha \sum_{i=1}^{n} h_i |e_i|.$$

Proposition 7 gives

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E} \leq Ch^k + \alpha C h^{k-1}. \tag{54}$$

Recalling that

$$(\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e})))_{i+\frac{1}{2}} = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) = \mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) - \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}),$$

we infer

$$\|\boldsymbol{\mathcal{F}}(\mathbf{u}) - \bar{\boldsymbol{\mathcal{F}}}\|_{H_E} = \|\boldsymbol{\mathcal{F}}(\mathbf{u}) - \boldsymbol{\mathcal{F}}(\bar{\mathbf{u}}) + \mathbf{R}\|_{H_E} \leq \|\boldsymbol{\mathcal{F}}(\mathbf{u}) - \boldsymbol{\mathcal{F}}(\bar{\mathbf{u}})\|_{H_E} + \|\mathbf{R}\|_{H_E} \leq Ch^{k-1}.$$

So the fluxes are convergent at order $k-1$. □

2.4.3 Estimation of the remainder

Lemma 1. *Let $k \in \mathbb{N}^*, k > 2$, $\bar{u} \in \mathcal{C}^k(\Omega)$ be the exact solution of (44) and assume that $\bar{u} \geq 0$. Let $\mathbf{u} \in \mathbb{R}^n$ be the solution of (45), (46) and (47) and assume that $u_i > 0, \forall i \in [\![1, n]\!]$. Let the remainder $\mathbf{r}(\mathbf{u}) \in \mathbb{R}^{n+1}$ be defined by $(\mathbf{r}(\mathbf{u}))_{i+\frac{1}{2}} = r_{i+\frac{1}{2}}(\mathbf{u}), \forall i \in [\![0, n]\!]$, $r_{i+\frac{1}{2}}$ being the remainder of the flux given by (12). Then we have*

$$\|\mathbf{r}(\mathbf{u})\|_{L^2} \leq Ch^{\frac{1}{2}}.$$

Proof. Considering the scheme (45), Proposition 9 gives

$$\|\mathcal{F}(\mathbf{u}) - \bar{\mathcal{F}}\|_{L^2} \leq Ch^{k-1},$$

and Equation (5) yields

$$\|\mathcal{F}(\mathbf{u}) - \bar{\mathcal{F}}\|_{\infty} \leq \sqrt{\max_{0\leq i\leq n} (\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \bar{\mathcal{F}}_{i+\frac{1}{2}})^2} = \frac{C}{\sqrt{h}} \sqrt{\sum_{i=0}^{n} h_{i+\frac{1}{2}} (\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \bar{\mathcal{F}}_{i+\frac{1}{2}})^2} = \frac{C}{\sqrt{h}} \|\mathcal{F}(\mathbf{u}) - \bar{\mathcal{F}}\|_{L^2} \leq Ch^{k-\frac{3}{2}},$$

that is to say

$$\bar{\mathcal{F}}_{i+\frac{1}{2}} = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{O}(h^{k-\frac{3}{2}}).$$

Proposition 6 implies

$$\bar{\mathcal{F}}_{i+\frac{1}{2}} = \mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) + \mathcal{O}(h^k). \tag{55}$$

The difference between these last two equalities yields

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{O}(h^{k-\frac{3}{2}}). \tag{56}$$

We use that the two point flux is consistent of order 1 and (55). This gives

$$\kappa_{i+\frac{1}{2}} \left(\frac{\bar{u}_{i+1} - \bar{u}_i}{h_{i+\frac{1}{2}}} \right) = \bar{\mathcal{F}}_{i+\frac{1}{2}} + \mathcal{O}(h). \tag{57}$$

The expression of $\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}})$ is given by

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \kappa_{i+\frac{1}{2}} \left(\frac{\bar{u}_{i+1} - \bar{u}_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\bar{\mathbf{u}}) \right),$$

from which we have

$$r_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \frac{1}{\kappa_{i+\frac{1}{2}}} \mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) - \frac{\bar{u}_{i+1} - \bar{u}_i}{h_{i+\frac{1}{2}}}. \tag{58}$$

Using (57), this yields

$$r_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \frac{1}{\kappa_{i+\frac{1}{2}}}\left(\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) - \bar{\mathcal{F}}_{i+\frac{1}{2}}\right) + \mathcal{O}(h),$$

and (55) gives

$$r_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \mathcal{O}(h^k) + \mathcal{O}(h) = \mathcal{O}(h). \tag{59}$$

The expression of the remainder is given by

$$r_{i+\frac{1}{2}}(\mathbf{u}) = \frac{1}{\kappa_{i+\frac{1}{2}}}\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}},$$

which means

$$\sum_{i=0}^{n} h_{i+\frac{1}{2}}|r_{i+\frac{1}{2}}(\mathbf{u})| = \sum_{i=0}^{n} h_{i+\frac{1}{2}}\left|\frac{1}{\kappa_{i+\frac{1}{2}}}\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}}\right|.$$

that is to say

$$\sum_{i=0}^{n} h_{i+\frac{1}{2}}|r_{i+\frac{1}{2}}(\mathbf{u})| = \sum_{i=0}^{n} h_{i+\frac{1}{2}}\left|\frac{1}{\kappa_{i+\frac{1}{2}}}\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) - \frac{\bar{u}_{i+1}-\bar{u}_i}{h_{i+\frac{1}{2}}} + \frac{1}{\kappa_{i+\frac{1}{2}}}\left(\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}})\right) + \left(\frac{\bar{u}_{i+1}-\bar{u}_i}{h_{i+\frac{1}{2}}} - \frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}}\right)\right|.$$

Equation (58) leads to

$$\sum_{i=0}^{n} h_{i+\frac{1}{2}}|r_{i+\frac{1}{2}}(\mathbf{u})| \leq \sum_{i=0}^{n} h_{i+\frac{1}{2}}\left|r_{i+\frac{1}{2}}(\bar{\mathbf{u}})\right| + \sum_{i=0}^{n} h_{i+\frac{1}{2}}\left|\frac{1}{\kappa_{i+\frac{1}{2}}}\left(\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}})\right)\right| + \sum_{i=0}^{n}|\bar{u}_{i+1}-u_{i+1}| + \sum_{i=0}^{n}|\bar{u}_i - u_i|.$$

Equations (2), (56) and (59) give

$$\sum_{i=0}^{n} h_{i+\frac{1}{2}}|r_{i+\frac{1}{2}}(\mathbf{u})| \leq Ch\sum_{i=0}^{n} h_{i+\frac{1}{2}} + \frac{C}{\kappa_0}h^{k-\frac{3}{2}}\sum_{i=0}^{n} h_{i+\frac{1}{2}} + \sum_{i=0}^{n} h_{i+1}\frac{|\bar{u}_{i+1}-u_{i+1}|}{h_{i+1}} + \sum_{i=0}^{n} h_i\frac{|\bar{u}_i - u_i|}{h_i}.$$

Equation (5) yields

$$\sum_{i=0}^{n} h_{i+\frac{1}{2}}|r_{i+\frac{1}{2}}(\mathbf{u})| \leq Ch + \frac{C}{\kappa_0}h^{k-\frac{3}{2}} + \frac{C}{h}\sum_{i=0}^{n} h_{i+1}\left|\bar{u}_{i+1}-u_{i+1}\right| + \frac{C}{h}\sum_{i=0}^{n} h_i\left|\bar{u}_i - u_i\right|.$$

Proposition 7 gives

$$\sum_{i=1}^{n} h_i|\bar{u}_i - u_i| = \mathcal{O}(h^{k-1}),$$

which leads to

$$\sum_{i=0}^{n} h_{i+\frac{1}{2}}|r_{i+\frac{1}{2}}(\mathbf{u})| \leq Ch + \frac{C}{\kappa_0}h^{k-\frac{3}{2}} + Ch^{k-2},$$

and, for $k > 2$,

$$\|\mathbf{r}(\mathbf{u})\|_{L^1} = \sum_{i=0}^{n} h_{i+\frac{1}{2}}|r_{i+\frac{1}{2}}(\mathbf{u})| \leq Ch. \tag{60}$$

Besides, Equations (5) yields

$$\|\mathbf{r}(\mathbf{u})\|_\infty = \max_{0\leq i\leq n}|r_{i+\frac{1}{2}}(\mathbf{u})| \leq \frac{C}{h}\sum_{i=0}^{n} h_{i+\frac{1}{2}}|r_{i+\frac{1}{2}}(\mathbf{u})| \leq \frac{C}{h}\|\mathbf{r}(\mathbf{u})\|_{L^1},$$

and Equation (60) gives

$$\|\mathbf{r}(\mathbf{u})\|_\infty \leq C. \tag{61}$$

Moreover, Equations (60) and (61) leads to

$$\|\mathbf{r}(\mathbf{u})\|_{L^2}^2 = \sum_{i=0}^{n} h_{i+\frac{1}{2}} r_{i+\frac{1}{2}}(\mathbf{u})^2 \leq \|\mathbf{r}(\mathbf{u})\|_\infty \sum_{i=0}^{n} h_{i+\frac{1}{2}} |r_{i+\frac{1}{2}}(\mathbf{u})| \leq \|\mathbf{r}(\mathbf{u})\|_\infty \|\mathbf{r}(\mathbf{u})\|_{L^1} \leq Ch,$$

which means

$$\|\mathbf{r}(\mathbf{u})\|_{L^2} \leq Ch^{\frac{1}{2}}.$$

□

2.4.4 Convergence at order k

Theorem 3 (Convergence at order k in H^1 norm)**.** *Let $\{x_i\}_{1\leq i\leq n}$ be a mesh satisfying (4) and (5). Let $k \in \mathbb{N}^*, k > 2$, $\bar{u} \in \mathcal{C}^k(\Omega)$ be the exact solution of (44) and assume that $\bar{\mathbf{u}} \geq \mathbf{0}$. Let $\mathbf{u} \in \mathbb{R}^n$ be the solution of (45), (46) and (47) and assume that $\mathbf{u} > \mathbf{0}$. Let us denote the vector $\mathbf{e} = (\bar{u}_i - u_i)_{1\leq i\leq n}$. Then we have*

$$\|\mathbf{e}\|_{H^1} \leq Ch^k,$$

with $\|\cdot\|_{H^1}$ defined by (8).

The following proof is inspired by a proof done in [11], Chapter 2, Section 6.1.

Proof. Let us recall the form of the scheme (45)

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u}) + \alpha h_i u_i = h_i f_i, \qquad \forall i \in [\![1, n]\!].$$

Besides, the integration of (44) on $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ gives

$$-\underbrace{\kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}})}_{\bar{\mathcal{F}}_{i+\frac{1}{2}}} + \underbrace{\kappa_{i-\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i-\frac{1}{2}})}_{\bar{\mathcal{F}}_{i-\frac{1}{2}}} + \alpha h_i \bar{u}_i = h_i f_i, \qquad \forall i \in [\![1, n]\!].$$

Subtracting these last two equalities yields

$$-(\bar{\mathcal{F}}_{i+\frac{1}{2}} - \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})) + (\bar{\mathcal{F}}_{i-\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) + \alpha h_i e_i = 0, \qquad \forall i \in [\![1, n]\!], \tag{62}$$

with $e_i = \bar{u}_i - u_i$. Multiplying (62) by e_i and summing on all the cells leads to

$$-\sum_{i=1}^{n} (\bar{\mathcal{F}}_{i+\frac{1}{2}} - \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}))e_i + \sum_{i=1}^{n} (\bar{\mathcal{F}}_{i-\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u}))e_i + \alpha \sum_{i=1}^{n} h_i e_i^2 = 0.$$

A discrete integration by parts with homogeneous Neumann boundary conditions gives

$$\sum_{i=1}^{n-1} (\bar{\mathcal{F}}_{i+\frac{1}{2}} - \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}))(e_{i+1} - e_i) + \sum_{i=1}^{n} \alpha h_i e_i^2 = 0. \tag{63}$$

The consistency of the fluxes given by (55) implies that

$$\bar{\mathcal{F}}_{i+\frac{1}{2}} = \mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) - R_{i+\frac{1}{2}}, \tag{64}$$

with

$$|R_{i+\frac{1}{2}}| \leq Ch^k. \tag{65}$$

Using (64) and replacing the flux by its definition (46), (63) becomes

$$\sum_{i=1}^{n-1}\kappa_{i+\frac{1}{2}}\frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}}+\sum_{i=1}^{n-1}\kappa_{i+\frac{1}{2}}(r_{i+\frac{1}{2}}(\bar{\mathbf{u}})-r_{i+\frac{1}{2}}(\mathbf{u}))(e_{i+1}-e_i)+\sum_{i=1}^{n}\alpha h_i e_i^2=\sum_{i=1}^{n-1}R_{i+\frac{1}{2}}(e_{i+1}-e_i). \tag{66}$$

Let us consider the function

$$\begin{aligned}(\Omega,\mathbb{R}^k)&\longrightarrow \mathbb{R},\\ (x,\boldsymbol{\phi})&\longmapsto \sum_{i=1}^{n-1}r_{i+\frac{1}{2}}(\boldsymbol{\phi})\mathbb{1}_{[x_i,x_{i+1}]}(x).\end{aligned}$$

This function is continuous and linear with respect to $\boldsymbol{\phi}$ and piecewise constant with respect to x. The continuity of r with respect to $\boldsymbol{\phi}$ implies that $\exists\lambda, \|r(x,\boldsymbol{\phi})\|_{L^2(\Omega)}\leq\lambda(x)\|\boldsymbol{\phi}\|_{L^2},\forall\boldsymbol{\phi}$. Lemma 1 shows that $\lambda(x)\leq Ch^{\frac{1}{2}}$. Then

$$\begin{aligned}\|r(x,\boldsymbol{\phi})\|^2_{L^2(\Omega)}=\int_0^1 r(x,\boldsymbol{\phi})^2dx&=\sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}}r(x,\boldsymbol{\phi})^2dx\\ &=\sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}}\left(r_{i+\frac{1}{2}}(\boldsymbol{\phi})\right)^2dx=\sum_{i=1}^{n-1}h_{i+\frac{1}{2}}r^2_{i+\frac{1}{2}}(\boldsymbol{\phi})=\|\mathbf{r}(\boldsymbol{\phi})\|^2_{L^2},\end{aligned}$$

that leads to

$$\|\mathbf{r}(\boldsymbol{\phi})\|_{L^2}\leq Ch^{\frac{1}{2}}\|\boldsymbol{\phi}\|_{L^2}. \tag{67}$$

We have

$$\left|\sum_{i=1}^{n-1}\kappa_{i+\frac{1}{2}}(r_{i+\frac{1}{2}}(\bar{\mathbf{u}})-r_{i+\frac{1}{2}}(\mathbf{u}))(e_{i+1}-e_i)\right|\leq\max_i(\kappa_{i+\frac{1}{2}})\sum_{i=1}^{n-1}\sqrt{h_{i+\frac{1}{2}}}|r_{i+\frac{1}{2}}(\mathbf{e})|\frac{|e_{i+1}-e_i|}{\sqrt{h_{i+\frac{1}{2}}}}.$$

The Cauchy Schwarz inequality gives

$$\left|\sum_{i=1}^{n-1}\kappa_{i+\frac{1}{2}}(r_{i+\frac{1}{2}}(\bar{\mathbf{u}})-r_{i+\frac{1}{2}}(\mathbf{u}))(e_{i+1}-e_i)\right|\leq\max_i(\kappa_{i+\frac{1}{2}})\|r(\mathbf{e})\|_{L^2}\left(\sum_{i=1}^{n-1}\frac{|e_{i+1}-e_i|^2}{h_{i+\frac{1}{2}}}\right)^{1/2},$$

and (67) yields

$$\left|\sum_{i=1}^{n-1}\kappa_{i+\frac{1}{2}}(r_{i+\frac{1}{2}}(\bar{\mathbf{u}})-r_{i+\frac{1}{2}}(\mathbf{u}))(e_{i+1}-e_i)\right|\leq C\max_i(\kappa_{i+\frac{1}{2}})\sqrt{h}\|\mathbf{e}\|_{L^2}\left(\sum_{i=1}^{n-1}\frac{|e_{i+1}-e_i|^2}{h_{i+\frac{1}{2}}}\right)^{1/2}.$$

The inequality $AB\leq\frac{1}{2}(A^2+B^2)$ leads to

$$\left|\sum_{i=1}^{n-1}\kappa_{i+\frac{1}{2}}(r_{i+\frac{1}{2}}(\bar{\mathbf{u}})-r_{i+\frac{1}{2}}(\mathbf{u}))(e_{i+1}-e_i)\right|\leq C\max_i(\kappa_{i+\frac{1}{2}})\sqrt{h}\left(\sum_{i=1}^{n}h_i|e_i|^2+\sum_{i=1}^{n-1}\frac{|e_{i+1}-e_i|^2}{h_{i+\frac{1}{2}}}\right). \tag{68}$$

From (2), (65) and (68) we deduce that (66) becomes

$$\kappa_0\sum_{i=1}^{n-1}\frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}}+\sum_{i=1}^{n}\alpha h_i|e_i|^2\leq Ch^k\sum_{i=1}^{n-1}|e_{i+1}-e_i|+C\sqrt{h}\sum_{i=1}^{n-1}\frac{|e_{i+1}-e_i|^2}{h_{i+\frac{1}{2}}}+C\sqrt{h}\sum_{i=1}^{n}h_i|e_i|^2, \tag{69}$$

that is to say

$$\left(\kappa_0-C\sqrt{h}\right)\sum_{i=1}^{n-1}\frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}}+\left(\alpha-C\sqrt{h}\right)\sum_{i=1}^{n}h_i|e_i|^2\leq Ch^k\sum_{i=1}^{n-1}|e_{i+1}-e_i|.$$

Choosing h such as $\kappa_0 - C\sqrt{h} \geq \min\left(\frac{\kappa_0}{2}, \frac{\alpha}{2}\right)$ and $\alpha - C\sqrt{h} \geq \min\left(\frac{\kappa_0}{2}, \frac{\alpha}{2}\right)$, inserting $\sqrt{h_{i+\frac{1}{2}}}$ in the right-hand side and applying the Cauchy-Schwarz inequality gives

$$\sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{n} h_i |e_i|^2 \leq \frac{1}{\min\left(\frac{\kappa_0}{2}, \frac{\alpha}{2}\right)} C h^k \left(\sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}}\right)^{1/2} \left(\sum_{i=1}^{n-1} h_{i+\frac{1}{2}}\right)^{1/2}.$$

Since $\sum_{i=1}^{n-1} h_{i+\frac{1}{2}} \leq 1$,

$$\sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{n} h_i |e_i|^2 \leq C h^k \left(\sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}}\right)^{1/2}.$$

Besides

$$\sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}} \leq \sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{n} h_i |e_i|^2,$$

that yields

$$\sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{n} h_i |e_i|^2 \leq C h^k \left(\sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{n} h_i |e_i|^2\right)^{1/2},$$

that is to say

$$\sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{n} h_i |e_i|^2 \leq C h^{2k}, \tag{70}$$

which means that $\|\mathbf{e}\|_{H^1} \leq C h^k$. So, the scheme is convergent at order k. □

Proposition 10. *Let $k \in \mathbb{N}^*$ and $\mathbf{e} \in \mathbb{R}^n$. If*

$$\|\mathbf{e}\|_{H^1}^2 = \sum_{i=1}^{n-1} \frac{(e_{i+1}-e_i)^2}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{n} h_i |e_i|^2 \leq C h^{2k},$$

then, we have

$$|e_i| \leq C h^k, \quad \forall i \in [\![1, n]\!].$$

Proof. First, there exists $i_0 \in [\![1, n]\!]$ such that

$$|e_{i_0}| \leq C h^k. \tag{71}$$

Indeed, if it is not the case,

$$|e_i|^2 > C h^{2k}, \quad \forall i \in [\![1, n]\!],$$

that is to say

$$\sum_{i=1}^{n} h_i |e_i|^2 > C h^{2k} \underbrace{\sum_{i=1}^{n} h_i}_{=1} = C h^{2k},$$

which is in contradiction with (70). Then, let $j > i_0$, we have

$$e_j = \sum_{\ell=i_0+1}^{j} (e_\ell - e_{\ell-1}) + e_{i_0},$$

which gives

$$|e_j| \leq \sum_{\ell=i_0+1}^{j} \sqrt{h_{\ell-\frac{1}{2}}} \frac{|e_\ell - e_{\ell-1}|}{\sqrt{h_{\ell-\frac{1}{2}}}} + |e_{i_0}|.$$

The Cauchy-Schwarz inequality and inequality (71) lead to

$$|e_j| \leq \left(\sum_{\ell=i_0+1}^{j} h_{\ell-\frac{1}{2}} \right)^{1/2} \left(\sum_{\ell=i_0+1}^{j} \frac{|e_\ell - e_{\ell-1}|^2}{h_{\ell-\frac{1}{2}}} \right)^{1/2} + Ch^k.$$

Since $\sum_{\ell=i_0+1}^{j} h_{\ell-\frac{1}{2}} \leq 1$ inequality (70) yields $|e_j| \leq Ch^k$.

The same proof can be done in the case of $j < i_0$ with $e_j = -\sum_{\ell=j+1}^{i_0} (e_i - e_{l-1}) + e_{i_0}$. □

Thus, the scheme is also convergent at order k with the L^∞-norm

$$|e_i| \leq Ch^k, \quad \forall i \in [\![1, n]\!].$$

2.4.5 Asymptotic behavior of the symmetry condition

Lemma 2. *Let $\{x_i\}_{1\leq i\leq n}$ be a mesh satisfying (4) and (5). Let $k \in \mathbb{N}^*, k > 2$, $\bar{u} \in \mathcal{C}^k(\Omega)$ be the exact solution of (44) and assume that $\bar{u} \geq 0$. Let $\mathbf{u} \in \mathbb{R}^n$ be the solution of (45), (46) and (47) and assume that $u_i > 0, \forall i \in [\![1, n]\!]$. Assume moreover that $\frac{d\bar{u}}{dx} \neq 0$ on Ω, then the condition (16) is asymptotically fulfilled as $h \to 0$.*

Proof. Theorem 3 shows that

$$\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) = \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + O(h^{k-1}),$$

and Proposition 10 that

$$u_{i+1} - u_i = \bar{u}_{i+1} - \bar{u}_i + O(h^k) = h_{i+\frac{1}{2}} \left(\frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + O(h) \right).$$

Then since $\frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) \neq 0$, for h small enough these two quantities have the same sign. □

Remark 6. *Because Lemma 1 requires that $k > 2$ to hold, we have only proven the arbitrary order of convergence of the method for $k \geq 3$. However, it is proven in [13] that the scheme defined by the fluxes (17) is second-order accurate. This allows us to claim that we have designed a provably arbitrarily high order method.*

2.5 The case of discontinuous diffusion coefficient κ

In the case where κ is discontinuous at the node $x_{i+\frac{1}{2}}$, we compute two fluxes $\mathcal{F}^L_{i+\frac{1}{2}}(\mathbf{u})$ and $\mathcal{F}^R_{i+\frac{1}{2}}(\mathbf{u})$. The first one is computed using a Taylor expansion in $[x_i, x_{i+\frac{1}{2}}]$ while the second one is computed via a Taylor expansion on $[x_{i+\frac{1}{2}}, x_{i+1}]$. Thus, we use two polynomial reconstructions, one on the left and the other on the right of $x_{i+\frac{1}{2}}$. For each node, we shift the stencil so that it does not cross the node where the discontinuity is located. Let us denote

$$\mathcal{F}^R_{i+\frac{1}{2}}(\mathbf{u}) = \kappa^R_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_{i+\frac{1}{2}}}{\frac{h_{i+1}}{2}} + r^R_{i+\frac{1}{2}}(\mathbf{u}) \right) \quad \text{and} \quad \mathcal{F}^L_{i+\frac{1}{2}}(\mathbf{u}) = \kappa^L_{i+\frac{1}{2}} \left(\frac{u_{i+\frac{1}{2}} - u_i}{\frac{h_i}{2}} + r^L_{i+\frac{1}{2}}(\mathbf{u}) \right),$$

with

$$\kappa^R_{i+\frac{1}{2}} = \kappa(x_{i+\frac{1}{2}} + \epsilon) \quad \text{and} \quad \kappa^L_{i+\frac{1}{2}} = \kappa(x_{i+\frac{1}{2}} - \epsilon),$$

where $r^R_{i+\frac{1}{2}}(\mathbf{u})$ (resp. $r^L_{i+\frac{1}{2}}(\mathbf{u})$) denotes the remainder associated with the polynomial reconstruction of the solution using the cells located at the right (resp. left) of the node $x_{i+\frac{1}{2}}$.

Thus, the continuous problem imposing the equality of the fluxes, we also impose it at the discrete level, that is to say $\mathcal{F}^R_{i+\frac{1}{2}}(\mathbf{u}) = \mathcal{F}^L_{i+\frac{1}{2}}(\mathbf{u})$ which leads to

$$\kappa^R_{i+\frac{1}{2}}\left(\frac{u_{i+1}-u_{i+\frac{1}{2}}}{\frac{h_{i+1}}{2}} + r^R_{i+\frac{1}{2}}(\mathbf{u})\right) = \kappa^L_{i+\frac{1}{2}}\left(\frac{u_{i+\frac{1}{2}}-u_i}{\frac{h_i}{2}} + r^L_{i+\frac{1}{2}}(\mathbf{u})\right),$$

which yields

$$u_{i+\frac{1}{2}} = \frac{h_i h_{i+1}}{2(h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i\kappa^R_{i+\frac{1}{2}})}\left[2\left(\frac{\kappa^R_{i+\frac{1}{2}}u_{i+1}}{h_{i+1}} + \frac{\kappa^L_{i+\frac{1}{2}}u_i}{h_i}\right) + \kappa^R_{i+\frac{1}{2}}r^R_{i+\frac{1}{2}}(\mathbf{u}) - \kappa^L_{i+\frac{1}{2}}r^L_{i+\frac{1}{2}}(\mathbf{u})\right].$$

Replacing $u_{i+\frac{1}{2}}$ by its expression in $\mathcal{F}^L_{i+\frac{1}{2}}$ or $\mathcal{F}^R_{i+\frac{1}{2}}$ gives

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \mathcal{F}^L_{i+\frac{1}{2}}(\mathbf{u}) = \mathcal{F}^R_{i+\frac{1}{2}}(\mathbf{u}) = \frac{2\kappa^L_{i+\frac{1}{2}}\kappa^R_{i+\frac{1}{2}}}{h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i\kappa^R_{i+\frac{1}{2}}}\left[(u_{i+1}-u_i) + \frac{1}{2}\left(h_{i+1}r^R_{i+\frac{1}{2}}(\mathbf{u}) + h_i r^L_{i+\frac{1}{2}}(\mathbf{u})\right)\right],$$

that is

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \tilde{\alpha}_{i+\frac{1}{2}}(u_{i+1}-u_i) + \tilde{r}_{i+\frac{1}{2}}(\mathbf{u}) \tag{72}$$

with

$$\tilde{\alpha}_{i+\frac{1}{2}} = \frac{2\kappa^L_{i+\frac{1}{2}}\kappa^R_{i+\frac{1}{2}}}{h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i\kappa^R_{i+\frac{1}{2}}}, \quad \tilde{r}_{i+\frac{1}{2}}(\mathbf{u}) = \frac{h_{i+1}\kappa^L_{i+\frac{1}{2}}\kappa^R_{i+\frac{1}{2}}}{h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i\kappa^R_{i+\frac{1}{2}}}r^R_{i+\frac{1}{2}}(\mathbf{u}) + \frac{h_i\kappa^L_{i+\frac{1}{2}}\kappa^R_{i+\frac{1}{2}}}{h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i\kappa^R_{i+\frac{1}{2}}}r^L_{i+\frac{1}{2}}(\mathbf{u}).$$

The coefficient $\tilde{\alpha}_{i+\frac{1}{2}}$ being positive the trick to obtain monotonicity (Section 1.3) and the step of symmetrization can be done again for this scheme. Besides, the previous analysis applies to this case. In the case where the condition of symmetrization is not satisfied, the flux (72) is replaced by the first-order approximation

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \tilde{\alpha}_{i+\frac{1}{2}}(u_{i+1}-u_i).$$

Remark 7. *In the case of a discontinuous right hand side f, we use the same type of strategy. The reconstruction is made on each side of the discontinuity to deal with the possible discontinuity of the second derivative.*

3 Numerical implementation

3.1 Division by zero

In the previous sections, we have assumed $u_i > 0, \forall i \in [\![1,n]\!]$, but in practice, u_i may vanish. In order to circumvent this difficulty, it is possible to add a term proportional to h^k to the denominator in the flux (as well as in the expression of $s_{i+\frac{1}{2}}$). Let $\epsilon > 0$, the flux is given by[1]

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}}\left[\left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}+\epsilon h^k}\right)u_{i+1} - \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i+\epsilon h^k}\right)u_i\right].$$

[1] In the benchmarks we have chosen $\epsilon = 10^{-11}$.

3.2 Fixed point for nonlinearity

The system obtained is of the form $A\mathbf{u} = \mathbf{b}$, A being a matrix dependent on the solution. So, a fixed point algorithm is necessary to solve this system. We start with an initial guess $\mathbf{u}^0$, compute the matrix $A(\mathbf{u}^0)$ and solve $A(\mathbf{u}^0)\mathbf{u}^1 = \mathbf{b}$. Repeating this process, we build a sequence $\mathbf{u}^n$ that, if it converges, tends to the solution of the scheme. We perform this algorithm until the difference between the solution obtained between two iterations is small enough. Then, the following loop is performed

$$\begin{aligned}
&\nu = 0\\
&A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b}\\
&\text{While } |\mathbf{u}^{\nu+1} - \mathbf{u}^\nu| > \epsilon\\
&\quad A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b}\\
&\quad \nu = \nu + 1.
\end{aligned}$$

Unfortunately, we have no proof of convergence of this algorithm. However, thanks to Theorem 1 or 2, we have the following safety proposition:

Proposition 11. *Assume that $f \geq 0$, $g \geq 0$, and either $||f||_{L^2(\Omega)} > 0$, $g(0) > 0$ or $g(1) > 0$. Assume moreover that $\mathbf{u}^0 > \mathbf{0}$. Then $\forall \nu, \mathbf{u}^\nu > \mathbf{0}$.*

To prove this property, we need to introduce the concept of irreducible matrix. We quote here [22, Definition 1.15].

Definition 3. *A $n \times n$ matrix A is **reducible** if there exits a $n \times n$ permutation matrix P such that*

$$PAP^T = \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix},$$

*where $A_{1,1}$ is a $r \times r$ submatrix and $A_{2,2}$ is a $(n-r) \times (n-r)$ submatrix, where $1 \leq r < n$. If no such permutation matrix exists, then A is **irreducible**.*

The matrix of the scheme can be proven to be irreducible in view of the following Lemma (see [22, Theorem 1.17]).

Lemma 3. *To any matrix A we associate the graph of nodes $1, 2, ..., n$ and of directed edges connecting i to j if $A_{ij} \neq 0$. Then A is irreducible if and only if for any pair $i \neq j$ there exists a chain of edges that allows to go from i to j,*

$$A_{i,k1} \neq 0 \to A_{k1,k2} \neq 0 \to \cdots \to A_{km,j} \neq 0.$$

With these definitions we can make use of the following theorem (see [22], Corollary 3.20).

Theorem 4. *If A is an irreducible strict M-matrix, then it is invertible and $\forall i, j : (A^{-1})_{ij} > 0$.*

We are now in position to prove Proposition 11.

Proof of Proposition 11. We argue by induction. We assume that $\mathbf{u}^\nu > 0$. Thus $A^T(\mathbf{u}^\nu)$ is a strict M-matrix (see Theorem 1 or 2). It is easy to check, that $A^T(\mathbf{u}^\nu)$ is also irreducible. Thus all the entries of $A^{-T}(\mathbf{u}^\nu)$ are positive, using Theorem 4, and consequently all the entries of $A^{-1}(\mathbf{u}^\nu)$ are positive. Using Remark 3, we know that all components of $\mathbf{b}$ are non-negative. Moreover, because of the assumption that either $||f||_{L^2(\Omega)} > 0$, $g(0) > 0$ or $g(1) > 0$, at least one component of $\mathbf{b}$ is non zero. It yields

$$\forall i \in [\![1, n]\!] : u_i^{\nu+1} = \sum_{j=1}^{n} A_{ij}^{-1} b_j > 0,$$

as a sum of non-negative numbers which are not all zeros. □

Proposition 11 shows that the condition $\mathbf{u}^\nu > \mathbf{0}$ remains satisfied during the fixed point procedure, which allows to always define $A(\mathbf{u}^\nu)$.

4 Numerical experiments

Given $\Omega =]0,1[$, κ a diffusion coefficient and g a function defined on $\partial\Omega$ consider Problem (3) with $\alpha = 0$, $\beta = 1$, $\gamma = 0$

$$\begin{cases} -\dfrac{d}{dx}\left(\kappa \dfrac{d\bar{u}}{dx}\right) = f & \text{in } \Omega, \\ \bar{u} = g & \text{on } \partial\Omega. \end{cases} \tag{73}$$

We will use three types of meshes:

1. regular meshes such that the possible discontinuities of the diffusion coefficient κ are placed at their vertices,
2. deformed meshes, whose deformation is given by: $x \to x + 0.65x(1-x)(0.5-x)\sin(0.8\pi)$,
3. random meshes, an example of which with 8 cells being given in Figure 1.

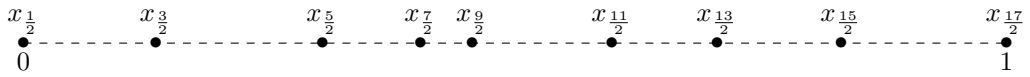


Figure 1: A random mesh with 8 cells.

4.1 L^2 convergence for polynomial solutions

Given $\kappa = 1$, $f(x) = -6x$ (resp. $f(x) = -72x^7$), $g(0) = 1$ and $g(1) = 2$, the function $\bar{u}(x) = x^3 + 1$ (resp. $\bar{u}(x) = x^9 + 1$) is solution to (73). We perform a convergence study for these problems on a deformed mesh with 64 cells. The L^2-error between the exact $\bar{u}$ and approximated u solutions are reported in the Table 1.

Order	$\bar{u}(x) = x^3 + 1$	$\bar{u}(x) = x^9 + 1$
1	1.64e-04	1.56e-03
2	3.46e-06	7.00e-04
3	4.53e-15	2.70e-04
4	3.79e-15	1.39e-06
5	8.15e-15	7.43e-07
6	2.57e-14	5.24e-10
7	4.21e-15	7.07e-09
8	5.02e-15	6.58e-13
9	7.86e-15	8.17e-15

Table 1: The L^2-error between the exact $\bar{u}$ and approximated u solutions.

The proof of exactness for polynomial of degree k (see appendix A) shows that the numerical solution must be exact for an order greater than 3 (resp. 9). The table of convergence (1) agrees with the theory since the error is zero, to machine precision, for the order greater than 3 (resp. 9).

4.2 L^2 convergence for a smooth diffusion coefficient

Given $\kappa = \exp(x)$, $f(x) = 4\exp(x) + 4x\exp(x) - \pi\cos(\pi x)\exp(x) + \pi^2\exp(x)\sin(\pi x)$ (note that f is positive), $g(0) = 4$ and $g(1) = 2$, the function $\bar{u}(x) = \sin(\pi x) - 2x^2 + 4$ is solution to (73). We perform a convergence study for this problem with the non-symmetric and symmetric schemes on the deformed mesh. The L^2-error and the L^2-error of the fluxes between the exact $\bar{u}$ and approximated u solutions are reported in Figures 2 to 5.

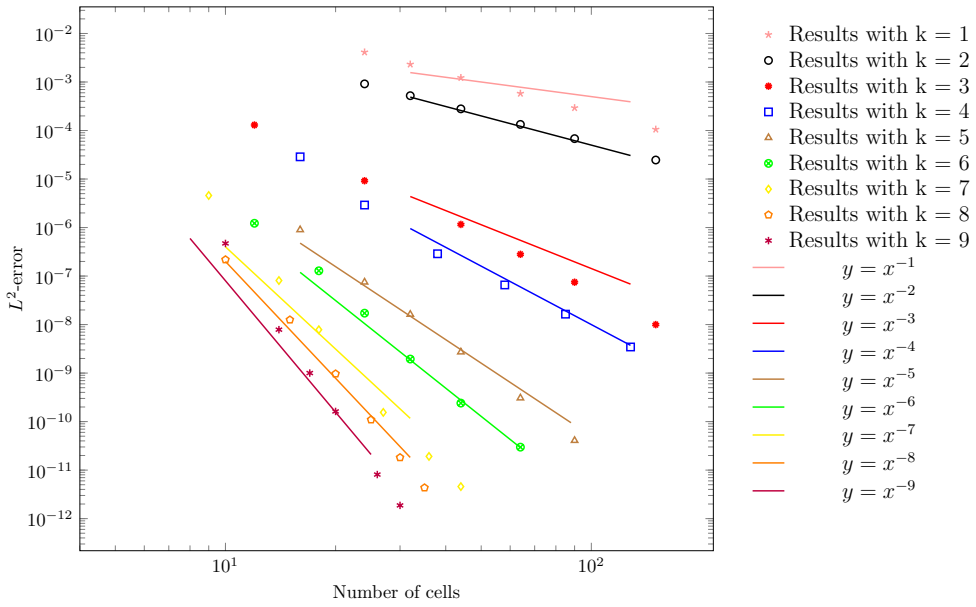


Figure 2: L^2-error with the non-symmetric scheme for problem of Sec. 4.2.

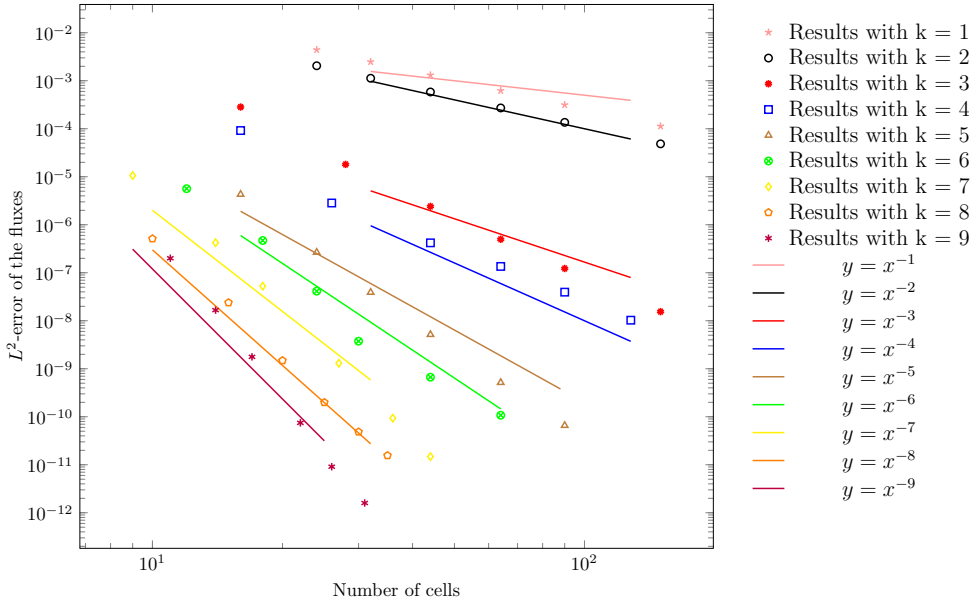


Figure 3: L^2-error of the fluxes with the non-symmetric scheme for problem of Sec. 4.2.

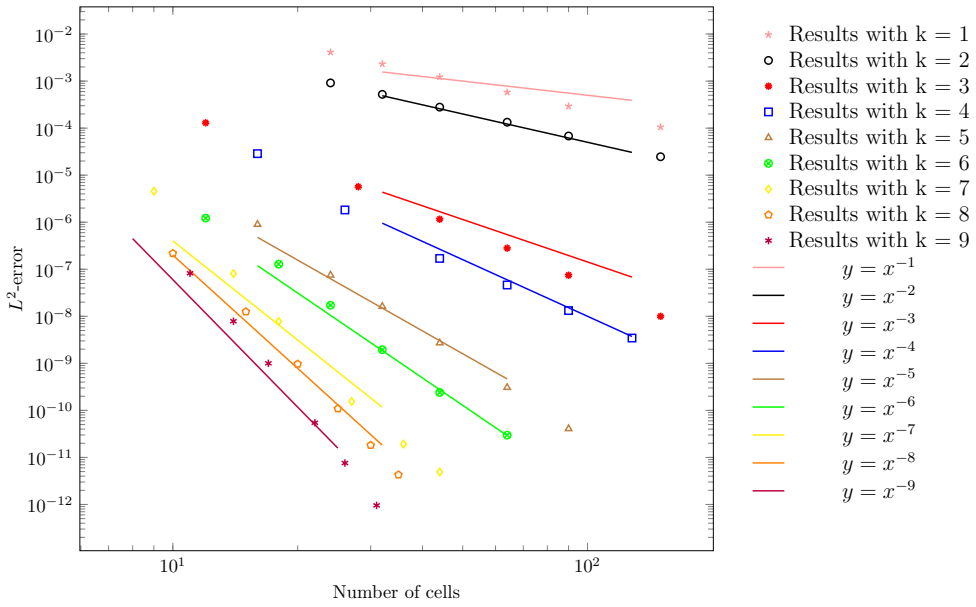


Figure 4: L^2-error with the symmetric scheme for problem of Sec. 4.2.

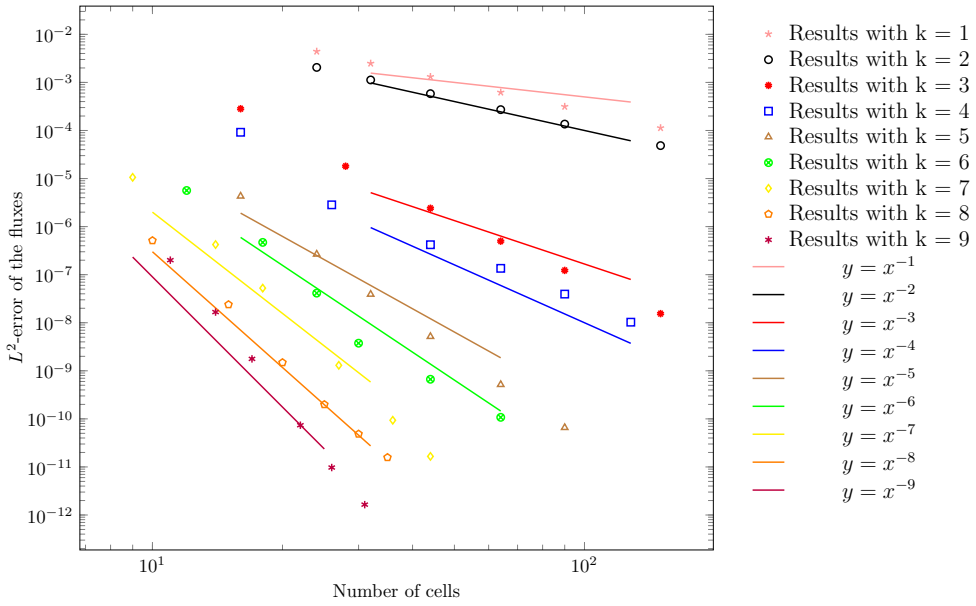


Figure 5: L^2-error of the fluxes with the symmetric scheme for problem of Sec. 4.2.

The results show that the numerical convergence order is at worst equal to the theorical order k (for the theorical order 4 one obtains convergence at order 4) or better (for the theorical order 3 one obtains the order 4). Besides, the results are qualitatively the same for the symmetric case and for the non-symmetric case. We observe similar convergence orders in L^2 norm and in L^2 norm of the fluxes. We also perform a convergence study for the same problem on the random mesh: see Figures 6 to 9.

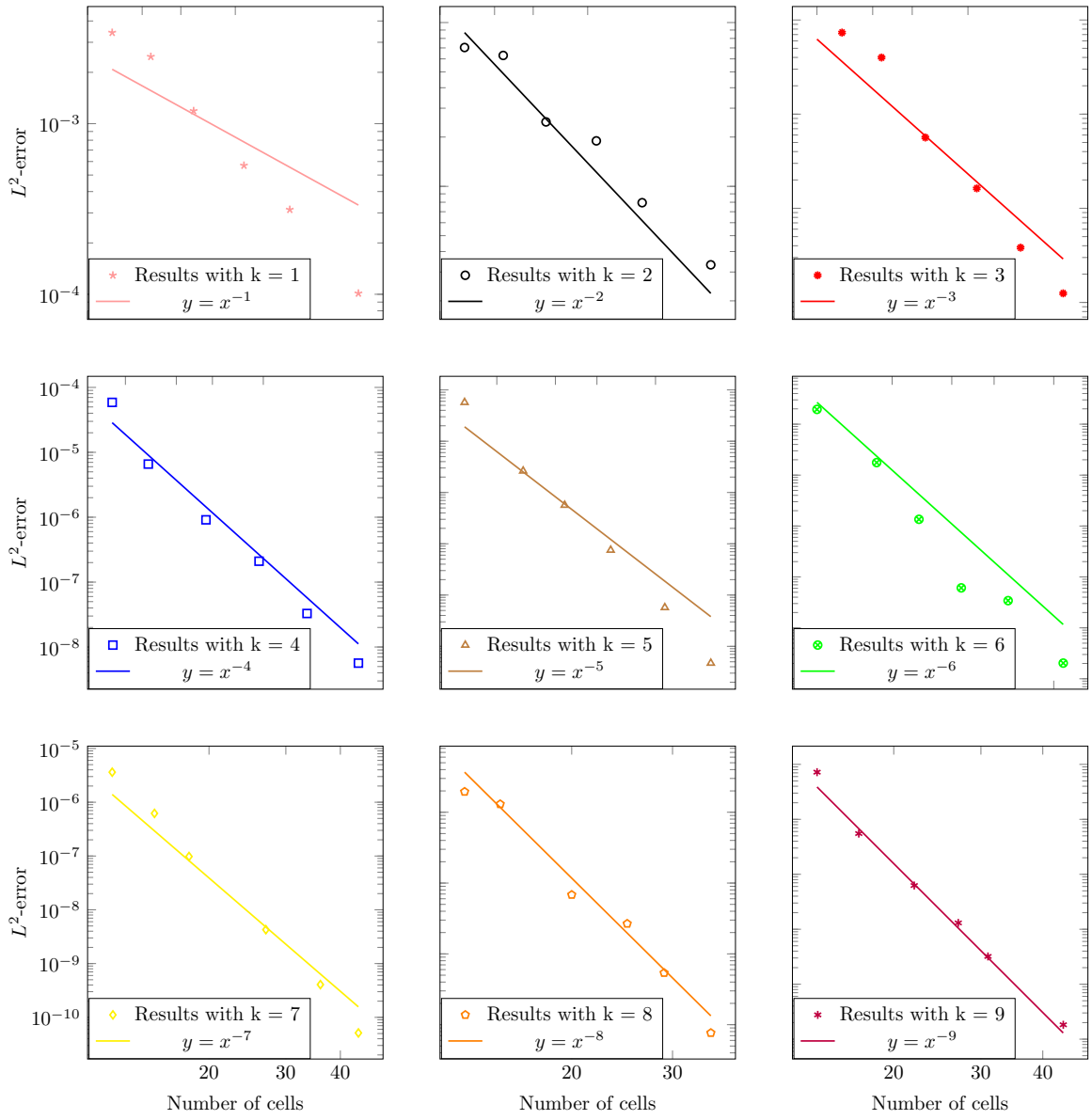


Figure 6: L^2-error with non-symmetric scheme and random mesh for problem of Sec. 4.2.

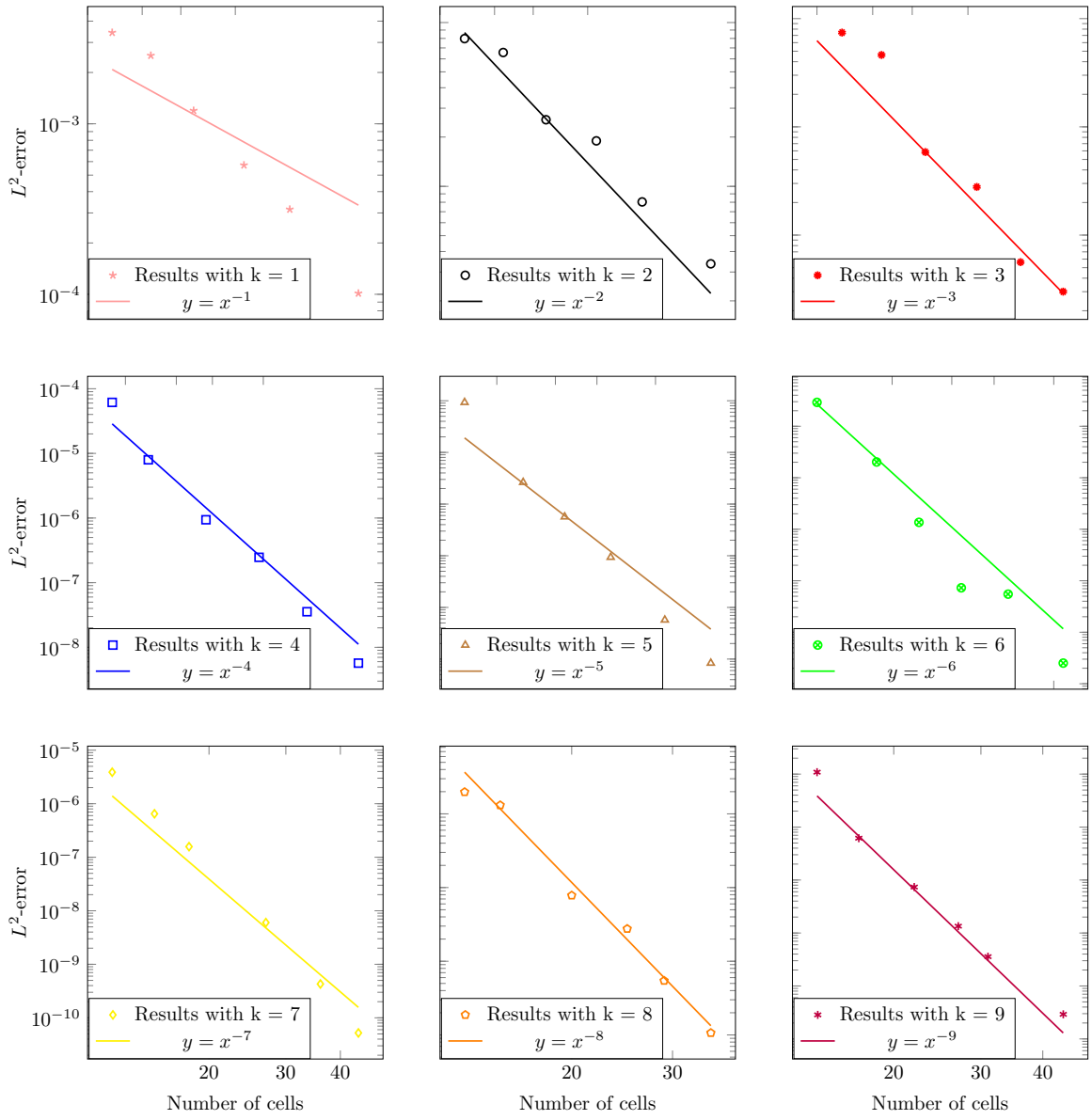


Figure 7: L^2-error on the fluxes with non-symmetric scheme and random mesh for problem of Sec. 4.2.

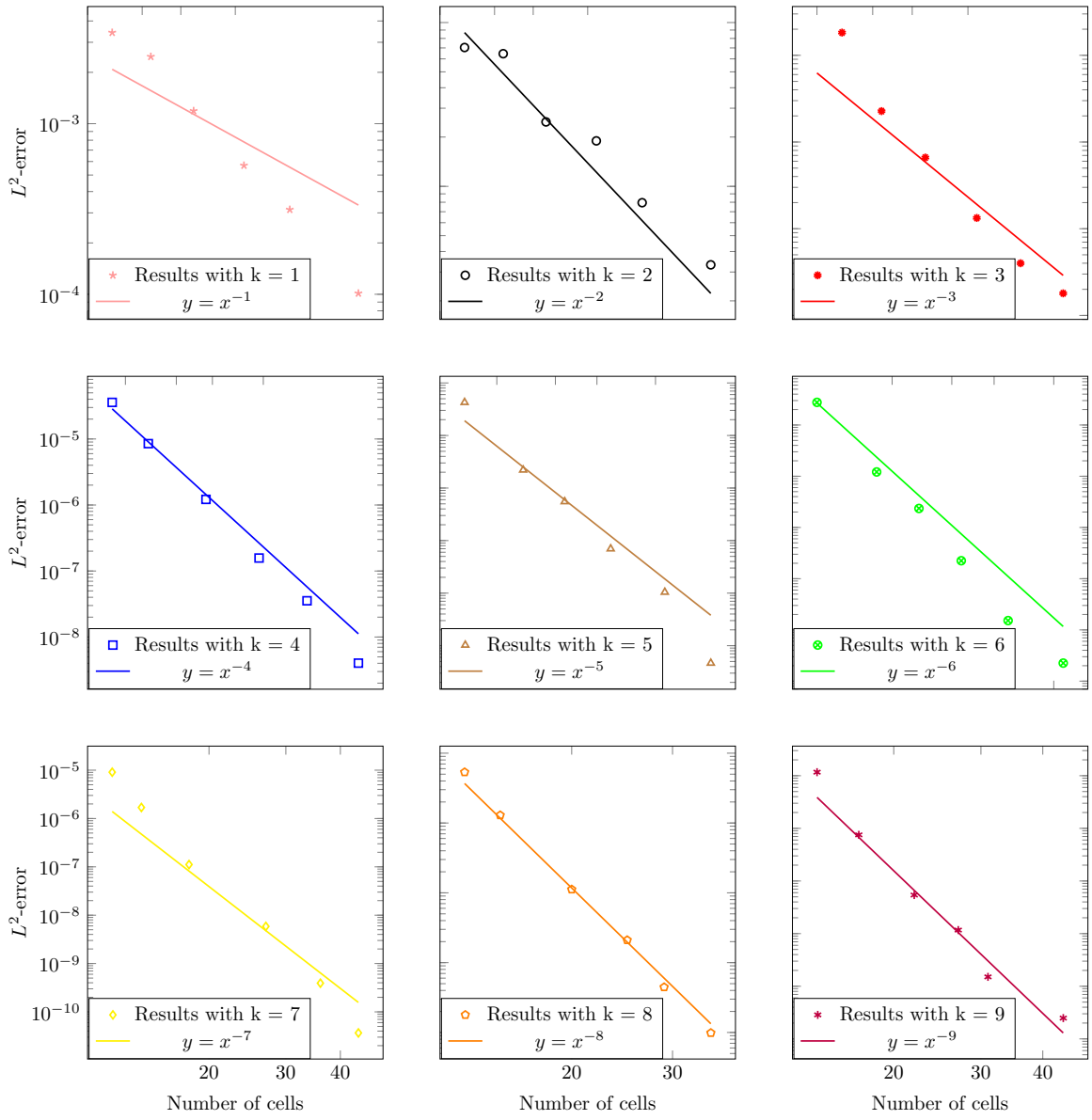


Figure 8: L^2-error with symmetric scheme and random mesh for problem of Sec. 4.2.

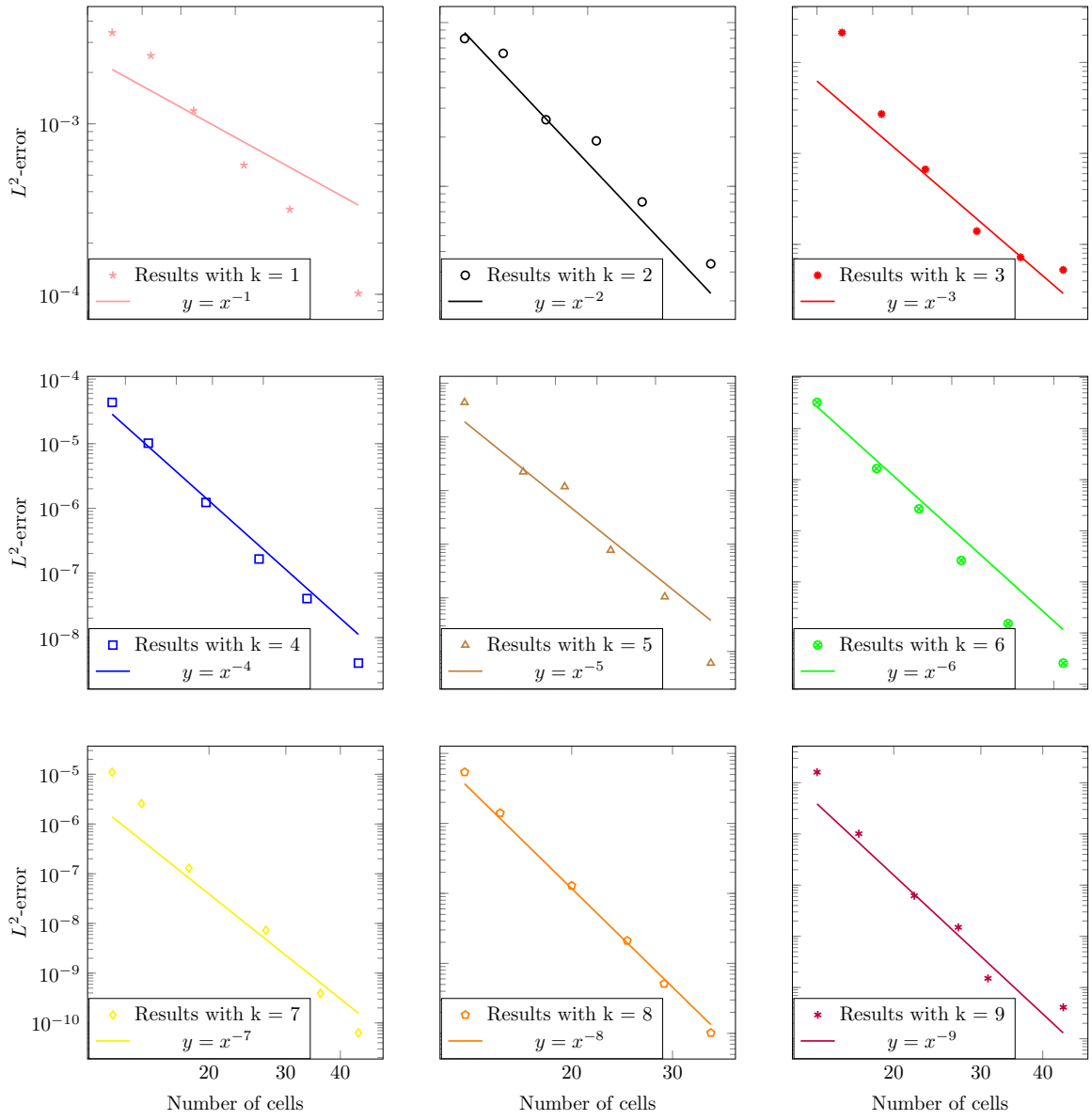


Figure 9: L^2-error on the fluxes with the symmetric scheme and random mesh for problem of Sec. 4.2.

4.3 Comparison with a non-monotone scheme

Given $\kappa = 1$, $f = \pi^2 \sin(\pi x)$ (note that f is positive), $g(0) = g(1) = 0$, the function $\bar{u}(x) = \sin(\pi x)$ is solution to (73). We perform a study for this problem on a deformed mesh for the symmetrical scheme at order 3. Results are summarized in Table 2. We can see that the solution obtained with the monotone scheme is always positive while the one obtained with the non-monotone scheme has a negative component.

Number of cells	High order monotone scheme	High order non monotone scheme
8	0	74
16	0	8
32	0	4
64	0	37
128	0	4

Table 2: Comparison of the number of negative components between the monotone and non-monotone schemes.

4.4 Discontinuous right hand side

Given $\kappa = 1$ and

$$f(x) = \mathbb{1}_{\{x > \frac{1}{2}\}}(x), \quad g(0) = \frac{1}{8}, \quad g(1) = \frac{1}{2},$$

the function

$$\bar{u}(x) = \left(\frac{1}{2}x + \frac{1}{8}\right)\mathbb{1}_{\{x \leq \frac{1}{2}\}}(x) + \left(-\frac{1}{2}x^2 + x\right)\mathbb{1}_{\{x > \frac{1}{2}\}}(x)$$

is solution to (73). We perform a convergence study for this problem, using the method described in Section 2.5, on a cartesian mesh for order 1 to 9. Results are summarized in Figure 10. These graphs show that, in the case of a discontinuous right hand side, if we apply the method considered for a discontinuous κ, the results are similar to those of the continuous case.

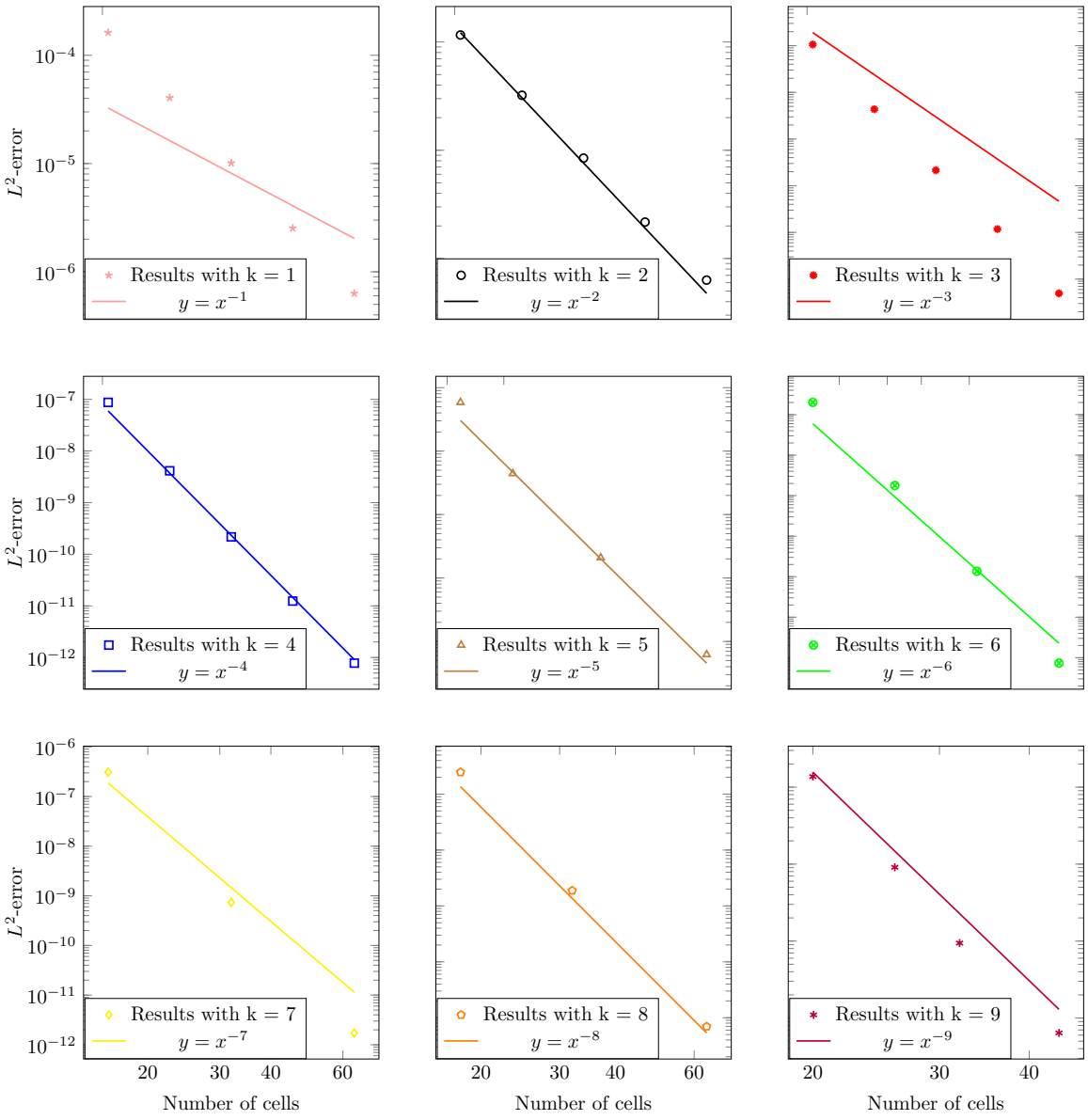


Figure 10: L^2-error with symmetric scheme and discontinuous right hand size for problem of Sec. 4.4.

4.5 Discontinuous diffusion coefficient κ

Given κ such that

$$\kappa(x) = \begin{cases} 1 & \text{if} \quad x \leq \dfrac{1}{2}, \\ 2 & \text{if} \quad x > \dfrac{1}{2}, \end{cases}$$

and $f(x) = \pi^2 \sin(\pi x)$, the function

$$\bar{u}(x) = (\sin(\pi x) + 2x)\, \mathbb{1}_{\{x \leq \frac{1}{2}\}}(x) + \left(\frac{1}{2}\sin(\pi x) + x + 1\right) \mathbb{1}_{\{x > \frac{1}{2}\}}(x),$$

is solution to (73). We perform a convergence study for this problem, using the method described in Section 2.5, on a cartesian mesh for order 1 to 9. Results are summarized in Figure 11. These graphs show that, in the case of a discontinuous κ, the results are similar to those of the continuous case.

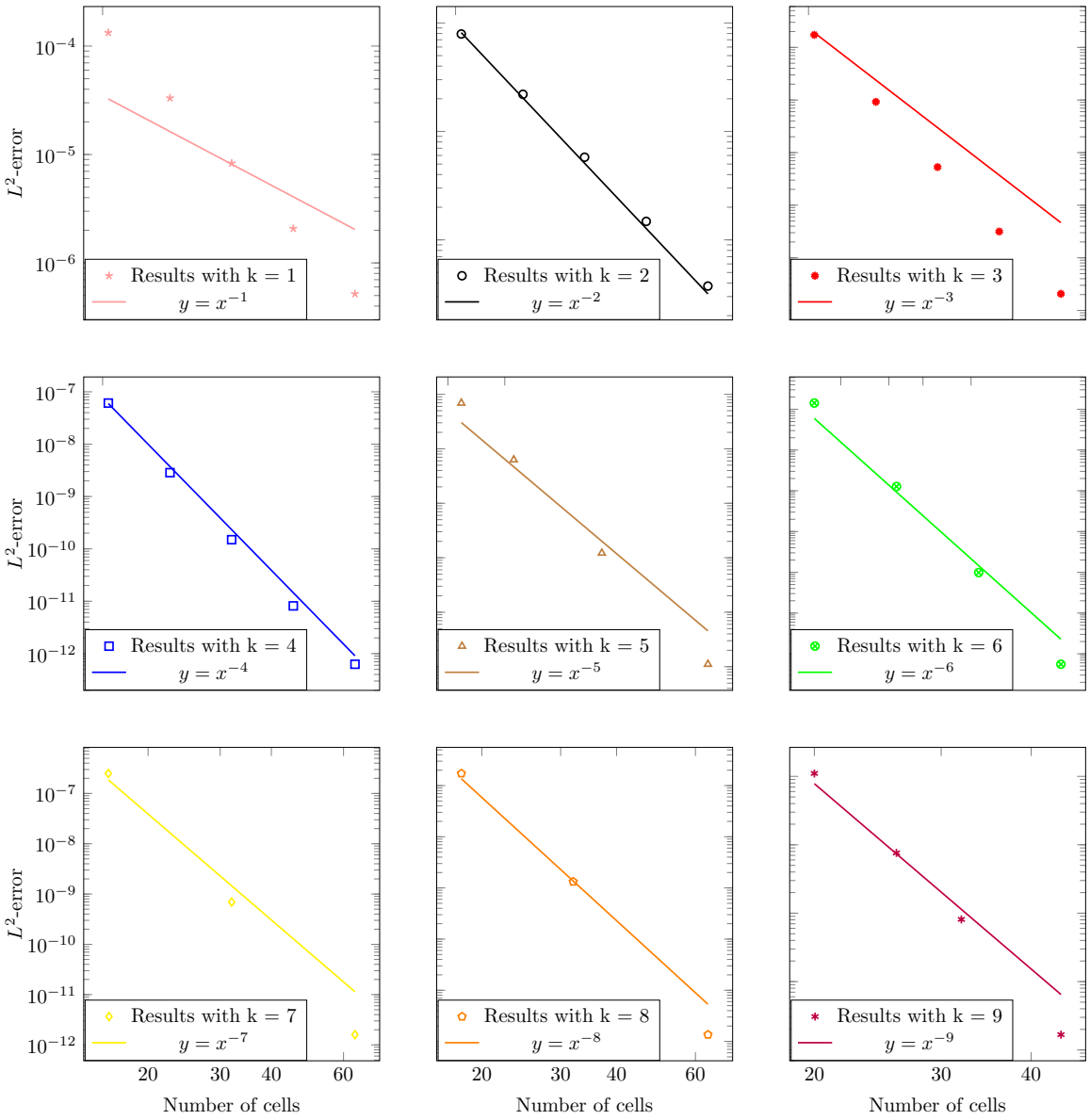


Figure 11: L^2-error with symmetric scheme and discontinuous κ for problem of Sec. 4.5.

5 Concluding remarks

In this paper we have proposed an arbitrary-order monotone scheme for the elliptic problem (3), on arbitrary 1D meshes. The properties of convergence at a given order, and the preservation of the positivity of the discrete solution have been proven. We also proposed a symmetrized version of the method. We have shown how to extend these schemes to the case of a discontinuous diffusion coefficient. These properties have been illustrated numerically up to the order 9. In future works, we aim to extend these schemes to higher spatial dimensions and to parabolic problems.

Acknowledgement

The authors thank Christophe Buet, Clément Cancès, Stphane Del Pino and Christophe Le Potier for fruitful discussions about this work and are indebted to Stéphane Del Pino for his help in the implementation of the method.

A Exactness for polynomials of degree k

To simplify the calculation let us take a polynomial of degree k centered on $x_{i+\frac{1}{2}}$ as an exact solution in order to demonstrate that the approximation of $\dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}})$ is exact for polynomials of degree k. For

$$\bar{u}(x) = \sum_{p=0}^{k} a_{i+\frac{1}{2},p}(x - x_{i+\frac{1}{2}})^p,$$

we obtain

$$\frac{d^\ell \bar{u}}{dx^\ell}(x) = \sum_{p=\ell}^{k} \frac{p!}{(p-\ell)!} a_{i+\frac{1}{2},p}(x - x_{i+\frac{1}{2}})^{p-\ell},$$

that is

$$\frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) = \ell! a_{i+\frac{1}{2},\ell}.$$

Besides, mean values were used to estimate the values of u at the centers of the cells, so

$$\bar{u}_{i+1} = \frac{1}{h_{i+1}} \int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} \sum_{p=0}^{k} a_{i+\frac{1}{2},p}(x - x_{i+\frac{1}{2}})^p = \sum_{p=0}^{k} a_{i+\frac{1}{2},p} \frac{h_{i+1}^p}{p+1},$$

and

$$\bar{u}_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \sum_{p=0}^{k} a_{i+\frac{1}{2},n}(x - x_{i+\frac{1}{2}})^p = \sum_{p=0}^{k} a_{i+\frac{1}{2},p} \frac{(-1)^p h_i^p}{p+1}.$$

The flux

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \frac{\kappa_{i+\frac{1}{2}}}{h_{i+\frac{1}{2}}} \left[\bar{u}_{i+1} - \bar{u}_i - \sum_{p=2}^{k} \frac{h_{i+1}^p + (-1)^{p+1} h_i^p}{(p+1)!} \frac{d^p \bar{u}}{dx^p}(x_{i+\frac{1}{2}}) \right],$$

becomes

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \frac{\kappa_{i+\frac{1}{2}}}{h_{i+\frac{1}{2}}} \left(\left[\sum_{p=0}^{k} a_{i+\frac{1}{2},p} \frac{h_{i+1}^p}{p+1} - \sum_{p=0}^{k} a_{i+\frac{1}{2},p} \frac{(-1)^p h_i^p}{p+1} \right] - \sum_{p=2}^{k} \frac{h_{i+1}^p + (-1)^{p+1} h_i^p}{(p+1)!} p! a_{i+\frac{1}{2},p} \right),$$

that is to say

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \kappa_{i+\frac{1}{2}} \left(a_{i+\frac{1}{2},1} + \sum_{p=2}^{k} a_{i+\frac{1}{2},p} \frac{h_{i+1}^p + (-1)^{p+1} h_i^p}{h_{i+\frac{1}{2}}(p+1)} - \sum_{p=2}^{k} \frac{h_{i+1}^p + (-1)^{p+1} h_i^p}{h_{i+\frac{1}{2}}(p+1)} a_{i+\frac{1}{2},p} \right) = \kappa_{i+\frac{1}{2}} a_{i+\frac{1}{2},1}.$$

The flux is exact for polynomials of degree k.

References

[1] L. Beirão da Veiga, F. Brezzi, L. D. Marini, and A. Russo. Virtual element method for general second-order elliptic problems on polygonal meshes. *Mathematical Models and Methods in Applied Sciences*, 26(04):729–750, 2016.

[2] 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 (electronic), 2005.

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

[4] J.-S. Camier and F. Hermeline. A monotone nonlinear finite volume method for approximating diffusion operators on general meshes. *Int. J. Numer. Meth. Engng*, 107:496–519, 2016.

[5] P. Ciarlet. *The Finite Element Method for elliptic problems*, volume 40. SIAM, Philadelphia, 2002.

[6] Bruno Després. Non linear schemes for the heat equation in 1d. *ESAIM: M2AN*, 48(1):107–134, 2014.

[7] Daniele Antonio Di Pietro and Jérôme Droniou. *The Hybrid High-Order method for polytopal meshes*, volume 19. Springer, 2019.

[8] Daniele Antonio Di Pietro, Jérôme Droniou, and Alexandre Ern. *Mathematical aspects of discontinuous Galerkin methods*. Springer, 2021.

[9] Jérôme Droniou and Christophe Le Potier. Construction and convergence study of schemes preserving the elliptic local maximum principle. *SIAM J. Numer. Anal.*, 49(2):459–490, 2011.

[10] L.C. Evans. Application of nonlinear semigroup theory to certain partial differential equations. *Nonlinear Evolution Equations*, pages 163–188, 1978.

[11] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In Ph. G. Ciarlet and J.-L. Lions, editors, *Handbook of numerical analysis*, volume VII. North-Holland, Amsterdam, 2000.

[12] Robert Eymard, Thierry Gallouët, Cindy Guichard, Raphaèle Herbin, and Roland Masson. TP or not TP, that is the question. *Computational Geosciences*, 18(3-4):285–296, 2014.

[13] P. Forsyth and P. Sammon. Quadratic convergence for cell-centered grids. *Applied Numerical Mathematics*, 4(5):377–394, 1988.

[14] Yanni Gao, Guangwei Yuan, Shuai Wang, and Xudeng Hang. A finite volume element scheme with a monotonicity correction for anisotropic diffusion problems on general quadrilateral meshes. *Journal of Computational Physics*, 407:109143, 2020.

[15] Christophe Le Potier. Schéma volumes finis monotone pour des opérateurs de diffusion fortement anisotropes sur des maillages de triangles non structurés. *Comptes Rendus Mathematique*, 341(12):787–792, 2005.

[16] Christophe Le Potier. Correction non linéaire et principe du maximum pour la discrétisation d'opérateurs de diffusion avec des schémas volumes finis centrés sur les mailles. *Comptes Rendus Mathematique*, 348(11-12):691–695, 2010.

[17] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Yu. Vassilevski. Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *Journal of Computational Physics*, 227(1):492–512, 2007.

[18] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes. *Journal of Computational Physics*, 228(3):703–716, 2009.

[19] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. Minimal stencil finite volume scheme with the discrete maximum principle. *Russian J. Numer. Anal. Math. Modelling*, 27(4):369–385, 2012.

[20] Z. Sheng and G. Yuan. A new nonlinear finite volume scheme preserving positivity for diffusion equations. *Journal of Computational Physics*, 315:182–193, 2016.

[21] Zhiqiang Sheng and Guangwei Yuan. The finite volume scheme preserving extremum principle for diffusion equations on polygonal meshes. *Journal of Computational Physics*, 230(7):2588–2604, 2011.

[22] R. S. Varga. *Matrix iterative analysis*, volume 1. Prentice Hall, 1962.

[23] J. Wang, Z. Sheng, and G. Yuan. A finite volume scheme preserving maximum principle with cell-centered and vertex unknowns for diffusion equations on distorted meshes. *Applied mathematics and computation*, 398(1):1–21, 2021.

[24] Y. Yu, X. Chen, and G. Yuan. A finite volume scheme preserving maximum principle for the system of radiation diffusion equation with three temperatures. *SIAM J. Sci. Comput.*, 41(1):93–113, 2019.

[25] Guangwei Yuan and Zhiqiang Sheng. Monotone finite volume schemes for diffusion equations on polygonal meshes. *Journal of Computational Physics*, 227(12):6288–6312, June 2008.

[26] F. Zhao, Z. Sheng, and G. Yuan. A monotone combination scheme of diffusion equations on polygonal meshes. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift fr Angewandte Mathematik und Mechanik*, 100(5):1–25, 2020.