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

Xavier Blanc, Francois Hermeline, Emmanuel Labourasse, Julie Patela

Arbitrary order monotonic 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 Paris Cité, 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;

June 15, 2023

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 meshes, 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 *monotonic* (that is, positivity preserving) numerical method for elliptic problems in 1D. We prove that this method converges to an arbitrary order (under reasonable assumptions on the mesh) and is indeed monotonic. 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 is bounded and 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 [14]) 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. An easy way to fix negative values is to truncate the solution to zero. However, it is not appropriate, since it breaks another very important property, which is the conservation. The standard finite volume two-point flux approximation (TPFA, see for example [15]) is positivity preserving (one also says monotonic) 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-κ-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 monotonic methods on highly deformed meshes deal with finite volume schemes. Monotonic methods can be designed in the finite-element framework (see [6, 8, 19, 20, 33] among others), but rely on restrictive conditions on the mesh we cannot afford. 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 [21] 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 such methods, one usually introduces secondary unknowns (for instance vertex-located or edge-located unknowns) in addition to the primal (cell-located) unknowns. Among others, important contributions to this field are [3, 23, 37], which propose efficient numerical schemes preserving the positivity of the primary unknowns. In [31], 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 [38, 24] explain how to build monotonic schemes without relying on secondary unknowns. In [22, 25, 30], maximum principle preserving schemes are proposed. Cancès and Guichard obtained moreover an entropy diminishing property in [5], introducing the non-linearity directly at the continuous level with a change of variables. Some concepts and proofs about the existence of solutions for these types of scheme can be found in [10, 13]. Recent advances in this field are [27, 34, 36]. All the works mentioned above concern 2D or 3D low-order (that is at most of order 2) numerical methods. Latterly, a third-order accurate monotonic method has been proposed in the Finite volume element (FVEM) context [35].

We are interested in designing a 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,

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

and we will suppose that $\Omega =]0,1[$ without loss of generality.

Although this setting is very specific, we believe it can be seen as a first step to tackle the question in higher dimensions. Let us be more precise about the 1D setting: in such a case, the TPFA scheme is actually consistent (and monotonic), 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, 7, 11, 12] arbitrary high order methods for the problem (1) is monotonic. In [10] it is shown how to use Le Potier's trick [22] to obtain monotonic 1D schemes of order greater than 2. But as this method uses a finite difference discretization on Cartesian meshes, it 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, under reasonable stability assumptions;
- it is monotonic;
- 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 monotonic behavior of the scheme is enforced using the trick described in [17], 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 [13] 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. On this aspect, our analysis is not completely satisfactory. A first approach consists in applying the fairly general analysis performed in [28], using the assumption that matrix of the scheme is coercive. This is what we do in Proposition 2.20 of Subsection 2.4.3, proving convergence at order k in L^2-norm. Unfortunately, we do not know how to prove that the matrix is coercive. Therefore, we propose a different approach, in which we replace such a coercivity assumption by a form of stability that is more general (see Assumption 2.16 of Subsection 2.4.1, and Proposition 2.17). We still do not know how to prove such an assumption, and Proposition 2.17 only gives convergence at order $k-1$ in L^1-norm. Finally in Section 3 we verify the properties previously stated on 1D test problems, showing that the method is indeed monotonic and of order k in L^2-norm for the solution and the fluxes.

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 of Ω 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

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

so that $\Omega =]x_{\frac{1}{2}} = 0, x_{n+\frac{1}{2}} = 1[$. We will also assume that the mesh is *quasi-uniform* that is there exists C such that

$$\max_{1\leq i\leq n}(h_i) < C \min_{1\leq i\leq n}(h_i). \tag{5}$$

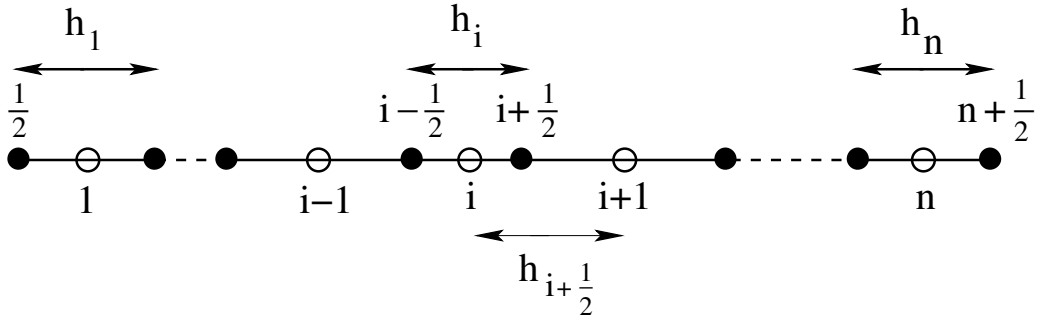

Figure 1: Definition of the mesh: i denotes the cells and $i + \frac{1}{2}$ the nodes.

We define $h = \max_{1\leq i\leq n}(h_i)$ 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

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

Finally the H^1 norm

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

Remark 1.1. *Note that* (6) *is a L^p-norm for grid function. Defining $u(x) = \sum_{i=1}^{n} u_i \mathbb{1}_{\left[i-\frac{1}{2}, i+\frac{1}{2}\right]}(x)$, we have*

$$\|\mathbf{u}\|_{L^p} = \left(\int_\Omega |u(x)|^p dx\right)^{1/p}.$$

1.1 Finite volume formulation

In this section, $\kappa(x)$ is assumed to be a continuous function. The extension to discontinuous κ is explained in Sec. 2.5. From now on we note $\kappa_{i+\frac{1}{2}} = \kappa(x_{i+\frac{1}{2}})$ and $\bar{\mathbf{u}} \in \mathbb{R}^n$ the vector defined by

$$\bar{u}_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \bar{u}(x)dx. \tag{9}$$

Let $\bar{u} \in \mathcal{C}^{k+1}(\overline{\Omega})$. The first step to design a finite volume scheme consists in integrating (3) on cell i

$$-\left[\kappa_{i+\frac{1}{2}} \left(\frac{d\bar{u}}{dx}\right)_{i+\frac{1}{2}} - \kappa_{i-\frac{1}{2}} \left(\frac{d\bar{u}}{dx}\right)_{i-\frac{1}{2}}\right] + \alpha h_i \bar{u}_i = h_i f_i,$$

with

$$f_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx. \tag{10}$$

Thus we need to define the fluxes

$$\bar{\mathcal{F}}_{i+\frac{1}{2}} = \kappa_{i+\frac{1}{2}} \left(\frac{d\bar{u}}{dx}\right)_{i+\frac{1}{2}} \quad \text{and} \quad \bar{\mathcal{F}}_{i-\frac{1}{2}} = \kappa_{i-\frac{1}{2}} \left(\frac{d\bar{u}}{dx}\right)_{i-\frac{1}{2}}.$$

First of all, the Taylor expansion at order k in the neighborhood of $x_{i+\frac{1}{2}}$ gives

$$\forall x \in \overline{\Omega}, \quad \bar{u}(x) = \bar{u}(x_{i+\frac{1}{2}}) + \sum_{\ell=1}^{k} \frac{(x - x_{i+\frac{1}{2}})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left((x - x_{i+\frac{1}{2}})^{k+1}\right). \tag{11}$$

In order to have mean values as degrees of freedom we integrate (11) from $x_{i+\frac{1}{2}}$ to $x_{i+\frac{3}{2}}$ and divide by h_{i+1}

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

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 (11) 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, using (5)

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

from which we obtain, using (5) again

$$\frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) = \frac{1}{h_{i+\frac{1}{2}}} \left(\bar{u}_{i+1} - \bar{u}_i - \sum_{\ell=2}^{k} \frac{h_{i+1}^\ell + (-1)^{\ell+1} h_i^\ell}{(\ell+1)!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}})\right) + \mathcal{O}\left(h^k\right). \tag{12}$$

Let $\mathbf{u} = (u_i)_{1\leq i\leq n}$ be the numerical solution. By mimicking the expression of the exact flux (12) the numerical flux is defined by

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) \right), \tag{13}$$

with

$$r_{i+\frac{1}{2}}(\mathbf{u}) = -\frac{1}{h_{i+\frac{1}{2}}} \sum_{\ell=2}^{k} \frac{h_{i+1}^{\ell} + (-1)^{\ell+1} h_i^{\ell}}{(\ell+1)!} \frac{d^{\ell} P}{dx^{\ell}}(x_{i+\frac{1}{2}}), \tag{14}$$

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

Remark 1.2. *For $k = 1$ (linear approximation of the fluxes), the remainder $r_{i+\frac{1}{2}}(\mathbf{u})$ vanishes, and the classical second-order accurate TPFA scheme is recovered.*

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. With a small abuse of notation, we 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

$$P(x) = a_k(u_0, ..., u_k) \left(x - x_{i+\frac{1}{2}} \right)^k + ... + a_0(u_0, ..., u_k).$$

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

$$\frac{1}{h_j} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} P(x)dx = u_j, \qquad \forall j \in [\![0, k]\!].$$

This leads to the following system

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

The matrix M_k can be rewritten

$$M_k = \begin{pmatrix} 1 & \frac{(x_{0+\frac{1}{2}} - x_{i+\frac{1}{2}})^2 - (x_{0-\frac{1}{2}} - x_{i+\frac{1}{2}})^2}{2(x_{0+\frac{1}{2}} - x_{0-\frac{1}{2}})} & \cdots & \frac{(x_{0+\frac{1}{2}} - x_{i+\frac{1}{2}})^{k+1} - (x_{0-\frac{1}{2}} - x_{i+\frac{1}{2}})^{k+1}}{(k+1)(x_{0+\frac{1}{2}} - x_{0-\frac{1}{2}})} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{(x_{k+\frac{1}{2}} - x_{i+\frac{1}{2}})^2 - (x_{k-\frac{1}{2}} - x_{i+\frac{1}{2}})^2}{2(x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}})} & \cdots & \frac{(x_{k+\frac{1}{2}} - x_{i+\frac{1}{2}})^{k+1} - (x_{k-\frac{1}{2}} - x_{i+\frac{1}{2}})^{k+1}}{(k+1)(x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}})} \end{pmatrix}. \tag{15}$$

Proposition 1.3. *Let $\{x_i\}_{1\leq i\leq n}$ be a mesh satisfying (4). Let $k \in \mathbb{N}^*$. The matrix M_k defined by (15) 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 1.4. *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 by P.*

1.3 A method to obtain monotonicity

A method borrowed from [17] and developed in the framework of 2D diffusion on arbitrary meshes can be used to make the scheme monotonic. This method has been successfully applied in a recent work [35]. The flux (13) 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

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left[\left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) u_{i+1} - \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i \right], \tag{16}$$

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

1.4 Symmetric version

Let us introduce a coefficient $s_{i+\frac{1}{2}}$ depending on $\mathbf{u}$ so that $\mathcal{F}_{i+\frac{1}{2}}$ can be rewritten as

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

To make the scheme symmetric the coefficients of u_i and u_{i+1} must be equal

$$\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} = \frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i}, \tag{18}$$

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

$$\frac{\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1} - u_i} \geq 0. \tag{19}$$

In other words, $u_{i+1} - u_i$ and $\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})$, defined by (13), 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. $\bar{u}_{i+1} \leq \bar{u}_i$).

In practice, if $\left(\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u})\right)(u_{i+1} - u_i) > 0$ we use the numerical flux (16), otherwise we use the first order approximation

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}\right). \tag{20}$$

1.5 Boundary conditions

1.5.1 Dirichlet boundary condition

In this section we only give the expression of the boundary conditions. Details are given in Appendix A. We consider problem (3) with $\beta = 1$, $\gamma = 0$. For the non-symmetric version of the scheme, application of the Dirichlet boundary condition on $x_{n+\frac{1}{2}}$ gives

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})}\right) g(x_{n+\frac{1}{2}}) - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right) u_n\right], \tag{21}$$

and for $x_{\frac{1}{2}}$,

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

For the symmetric version, we obtain

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})}\right) g(x_{n+\frac{1}{2}}) - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right) u_n\right], \tag{22}$$

and for the left boundary, similarly

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}} \left[\left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1}\right) u_1 - \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})}\right) g(x_{\frac{1}{2}})\right]. \tag{23}$$

1.5.2 Neumann boundary condition

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

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}} \left.\frac{d\bar{u}}{dx}\right|_{\frac{1}{2}} = -\kappa_{\frac{1}{2}} \left.\frac{d\bar{u}}{dn}\right|_{\frac{1}{2}} = -g(x_{\frac{1}{2}}) \tag{24}$$

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

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left.\frac{d\bar{u}}{dx}\right|_{n+\frac{1}{2}} = \kappa_{n+\frac{1}{2}} \left.\frac{d\bar{u}}{dn}\right|_{n+\frac{1}{2}} = g(x_{n+\frac{1}{2}}). \tag{25}$$

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 $i = n$

$$\bar{u}(x_{i+\frac{1}{2}}) = \frac{1}{\beta(x_{i+\frac{1}{2}})} \left(g(x_{i+\frac{1}{2}}) - \gamma(x_{i+\frac{1}{2}})\kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dn}(x_{i+\frac{1}{2}})\right). \tag{26}$$

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 (26) in (18) (see also (68) in the Appendix) yields

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \frac{\kappa_{n+\frac{1}{2}}\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u})+s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}}\right) g(x_{n+\frac{1}{2}}) - \beta(x_{n+\frac{1}{2}})\kappa_{n+\frac{1}{2}}\left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})+s_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right) u_n}{\beta(x_{n+\frac{1}{2}}) + \gamma(x_{n+\frac{1}{2}})\kappa_{n+\frac{1}{2}}\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u})+s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}}\right)}. \tag{27}$$

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

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \frac{\beta(x_{\frac{1}{2}})\kappa_{\frac{1}{2}}\left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u})+s_{\frac{1}{2}}(\mathbf{u})}{u_1}\right) u_1 - \kappa_{\frac{1}{2}}\left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u})+s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}}\right) g(x_{\frac{1}{2}})}{\beta(x_{\frac{1}{2}}) + \gamma(x_{\frac{1}{2}})\kappa_{\frac{1}{2}}\left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u})+s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}}\right)}. \tag{28}$$

Remark 1.5. *In the expression of the fluxes* (28) *and* (27)*, if we take* $\beta = 0$, $\gamma = 1$*, we obtain the same fluxes as* (24) *and* (25)*. Likewise, if we take* $\beta = 1, \gamma = 0$*, we obtain the same flux as* (23) *and* (22)*.*

1.6 Summary of the method and matrix form

The scheme reads as

$$-(\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) + \alpha h_i u_i = h_i f_i, \tag{29}$$

that is, using (17),

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

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

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

$$A_{ij} = \begin{cases}
-\kappa_{i+\frac{1}{2}}\left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}}\right) & \text{if } j = i+1, \forall i \neq n, \\
\kappa_{i+\frac{1}{2}}\left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i}\right) + \kappa_{i-\frac{1}{2}}\left(\frac{1}{h_{i-\frac{1}{2}}} + \frac{r^+_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_i}\right) + \alpha h_i & \text{if } j = i, \forall i \neq 1, n, \\
-\kappa_{i-\frac{1}{2}}\left(\frac{1}{h_{i-\frac{1}{2}}} + \frac{r^-_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_{i-1}}\right) & \text{if } j = i-1, \forall i \neq 1, \\
0 & \text{else.}
\end{cases} \tag{30}$$

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

$$b_1 = h_1 f_1 + \kappa_{\frac{1}{2}}\left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})}\right) g(x_{\frac{1}{2}}), \tag{31}$$

$$A_{1,1} = \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1, \tag{32}$$

and

$$b_n = h_n f_n + \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})} \right) g(x_{n+\frac{1}{2}}), \tag{33}$$

$$A_{n,n} = \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n. \tag{34}$$

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

$$b_1 = h_1 f_1 + g(x_{\frac{1}{2}}), \tag{35}$$

$$A_{1,1} = \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1, \tag{36}$$

and

$$b_n = h_n f_n + g(x_{n+\frac{1}{2}}), \tag{37}$$

$$A_{n,n} = \kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n. \tag{38}$$

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

$$b_1 = h_1 f_1 + \frac{\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)}{\beta(x_{\frac{1}{2}}) + \gamma(x_{\frac{1}{2}}) \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)} g(x_{\frac{1}{2}}), \tag{39}$$

$$A_{1,1} = \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \frac{\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right)}{1 + \frac{\gamma(x_{\frac{1}{2}}) \kappa_{\frac{1}{2}}}{\beta(x_{\frac{1}{2}})} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)} + \alpha h_1, \tag{40}$$

and

$$b_n = h_n f_n + \frac{\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)}{\beta(x_{n+\frac{1}{2}}) + \gamma(x_{n+\frac{1}{2}}) \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)} g(x_{n+\frac{1}{2}}), \tag{41}$$

$$A_{n,n} = \kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \frac{\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right)}{1 + \frac{\gamma(x_{n+\frac{1}{2}}) \kappa_{n+\frac{1}{2}}}{\beta(x_{n+\frac{1}{2}})} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)} + \alpha h_n. \tag{42}$$

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 1.6. *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.*

Remark 1.7. *In the case of mixed boundary condition, the right hand side of the nonlinear system depends on $\mathbf{u}$.*

1.7 A fixed point method for handling nonlinearity

The system obtained is of the form $A\mathbf{u} = \mathbf{b}$, A being a matrix dependent on the solution. So, we use a fixed point algorithm (a Picard iteration method) to solve this system as, for instance, in [3, 4, 13, 29]. 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 $\mathbf{u}^\nu$ 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[1]. To summarize, the following loop is performed

$$\begin{aligned} &\nu = 0 \\ &A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b} \\ &\text{While } \|\mathbf{u}^{\nu+1} - \mathbf{u}^\nu\|_{L_2} > \varepsilon \\ &\quad A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b} \\ &\quad \nu = \nu + 1. \end{aligned} \tag{43}$$

Unfortunately, we have no proof of convergence of this algorithm. Nevertheless, the numerical tests we have performed did not provide any situation in which the above fix-point algorithm does not converge.

Note that, in [13], the authors show that the nonlinear system has a solution. The proof is quite general and can be adapted to our case, but there is no proof of convergence of the fixed point algorithm. In some favorable cases, one can prove the convergence of the fixed point algorithm, *e.g.* if α is large enough (see [3]).

Remark 1.8. *We thus have two different schemes: the first one is linear and (expected to be) of high order, as we will see below. It is defined by the fluxes (13). Its definition does not require the unknown $\mathbf{u}$ to be positive, and its stencil is approximately of size $k+1$. The second scheme is nonlinear, and defined by the fluxes (16). We need $\mathbf{u}$ to be positive in order to define it, and its stencil is equal to 2. If it has a (positive) solution, then it is a solution of the linear scheme. Thus, two situations may occur:*

1. *the solution of the linear scheme is positive; then, it is also a solution to the nonlinear scheme;*
2. *the solution of the linear scheme has non-positive entries. Then, the nonlinear scheme* cannot *have a solution. Indeed, such a solution would be positive, hence solution to the linear scheme. We nevertheless expect the above fix-point algorithm to converge to some $\mathbf{u}$ that is non-negative, but is* not *a solution to the nonlinear scheme (nor to the linear scheme).*

However, the solution of the continuous problem (3) *satisfies a local maximum principle. Hence, assuming that the solution $\bar{u}$ is positive and that the linear scheme converges in the L^∞ norm, its solution becomes a positive vector for small enough values of h. This situation corresponds to Item 1 above, and the solution of the nonlinear scheme coincides with the solution of the linear scheme. The case of Item 2 happens only for larger values of h. In such a case, the monotonicity correction allows to recover positive values of the solution, while giving up, to some extent, the equation defining the linear scheme, at least for points at which the solution to the linear scheme is non-positive. What we observe numerically (see Section 3 below) is that the fix-point algorithm always converges, to a "solution" $\mathbf{u} \geq 0$ that is an approximation of order k to the exact solution $\bar{u}$.*

1.8 Sketch of the method

We summarize the method as follows.

Initialization

[1] In the numerical tests, we choose $\varepsilon = 10^{-12}$

- Initialize $\mathbf{u}^0 > 0$.
- Evaluate κ at the nodes: $\kappa_{i+\frac{1}{2}}, i \in [0\ n]$; and the mean value of f in each cell: $f_i, i \in [1\ n]$.

Picard iterations (ν):

Do

- Reconstruct polynomials $P_{i+\frac{1}{2}}, i \in [0\ n]$, of degree k, in each cells i using the method described in Section 1.2.
- Compute the reminder $\mathbf{r}_{i+\frac{1}{2}}(\mathbf{u}), i \in [0\ n]$ using equation (14).
- Distribute the reminder $\mathbf{r}_{i+\frac{1}{2}}(\mathbf{u})$ between cells i and $i+1$ to enforce monotonicity (see Section 1.3).
- Possibly, symmetrize the coefficients at each node, using the method of Section 1.4.
- Build the matrix $A(\mathbf{u}^\nu)$ and the right-hand side $\mathbf{b}^\nu$ (see Section 1.6).
- Solve $A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b}^\nu$.

While $\|\mathbf{u}^{\nu+1} - \mathbf{u}^\nu\|_{L_2} > \varepsilon$.

2 Properties

2.1 Conservation

Proposition 2.1. *Assume that* $\mathbf{u} > \mathbf{0}$ *and consider homogeneous Neumann boundary conditions, then the scheme defined by* (29) *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 2.2. *A matrix* $A = (a_{ij})$ *is an M-matrix if it satisfies the following inequalities*

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

and

$$\forall i, \sum_{j=1}^{n} a_{i,j} \geq 0. \tag{44}$$

Moreover, if (44) *is strict for all* $i \in [\![1, n]\!]$, *we say that* A *is a strict M-matrix.*

2.2.1 Non-symmetric version: property of the matrix

Proposition 2.3. *Assume that* $\mathbf{u} > \mathbf{0}$*, the matrix* $A(\mathbf{u})$ *defined by (30) and (31) through (34), or (35) through (38), or (39) through (42) depending on the boundary conditions, with* $s_{i+\frac{1}{2}} = 0$*, is such that* $A^T(\mathbf{u})$ *is a strict M-matrix.*

Remark 2.4. *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 2.3. The matrix satisfies

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

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

$$\sum_i A_{i,1}(\mathbf{u}) = A_{1,1}(\mathbf{u}) + A_{2,1}(\mathbf{u}),$$

which leads to

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

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

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

□

2.2.2 Strict monotonicity of the method

Proposition 2.5. *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 [32, Definition 1.15].

Definition 2.6. *An $n \times n$ matrix A is **reducible** if there exits an $n \times n$ 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 an $r \times r$ submatrix and $A_{2,2}$ is an $(n-r) \times (n-r)$ 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 [32, Theorem 1.17]).

Lemma 2.7. *To any $n \times n$ 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 [32], Corollary 3.20).

Theorem 2.8. *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 2.5.

Proof of Proposition 2.5. We argue by induction on the index ν. We assume that $\mathbf{u}^\nu > 0$. Thus $A^T(\mathbf{u}^\nu)$ is a strict M-matrix (see Proposition 2.3). 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 2.8, and consequently all the entries of $A^{-1}(\mathbf{u}^\nu)$ are positive. Using Remark 1.6, 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. We thus have

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

since all terms of this sum are non-negative, with one at least that is positive. □

Proposition 2.5 shows that the condition $\mathbf{u}^\nu > \mathbf{0}$ remains satisfied during the fixed point procedure, which allows to always define $A(\mathbf{u}^\nu)$. It shows moreover, than as long as hypothesis of the Proposition 2.5 are satisfied, all the properties requiring $\mathbf{u} > 0$ are verified for every fix point iteration.

2.2.3 Symmetric version: LMP structure

Proposition 2.9. *Assume that $\mathbf{u} > \mathbf{0}$, the matrix A defined by (30) and (31) through (34), or (35) through (38), or (39) through (42), depending on the boundary conditions, is symmetric.*

Proof. Let $x_{i+\frac{1}{2}}$, be an interior vertex of the mesh. If condition (19) is satisfied for this vertex, we use the definition of the flux (17), then symmetrization condition leads to $A_{i,i+1} = A_{i+1,i}$. Otherwise the flux is defined by (20), and once again $A_{i,i+1} = A_{i+1,i}$.

□

Proposition 2.10. *Assume that* $\mathbf{u} > \mathbf{0}$*, let* A *be defined by (30) and (31) through (34), or (35) through (38), or (39) through (42), depending on the boundary conditions, then the matrix* A *is a strict M-matrix.*

Proof. As for Proposition 2.3, 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 2.11. *This definition is taken from [13]. 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*

$$\forall i \in [\![1,n]\!] : \sum_{j=1}^{n} \lambda_{i,j}(\mathbf{u})(u_i - u_j) + \lambda_{i,\frac{1}{2}}(\mathbf{u})(u_i - u_{\frac{1}{2}}) + \lambda_{i,n+\frac{1}{2}}(\mathbf{u})(u_i - u_{n+\frac{1}{2}}) = f_i h_i, \tag{45}$$

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

$$\lambda_{1,\frac{1}{2}} > 0, \quad \lambda_{n,n+\frac{1}{2}} > 0, \quad \textit{and} \quad \forall i \in [\![1,n-1]\!] : \lambda_{i,i\pm1} > 0. \tag{46}$$

Theorem 2.12. *Assume that* $f \geq 0$*,* $g \geq 0$*, and either* $\|f\|_{L^2(\Omega)} > 0$*,* $g(0) > 0$ *or* $g(1) > 0$*. Let* A *and* $\mathbf{b}$ *be defined by (30) and (31) through (34), or (35) through (38), or (39) through (42), 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.*

Proof. For interior vertices, we consider two cases:

- if condition (19) is satisfied, then the coefficients of the fluxes are defined by (18), and we have

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

  which is positive because of (19).

- if condition (19) is not satisfied, then the coefficients of the fluxes are defined by (20), and

$$\lambda_{i+\frac{1}{2}} := \frac{\kappa_{i+\frac{1}{2}}}{h_{i+\frac{1}{2}}},$$

  which is positive.

Substituting $\lambda_{i+\frac{1}{2}}$ in equation (17) and using the definition of the scheme (29) with $\alpha = 0$ yields

$$\lambda_{i+\frac{1}{2}}(u_i - u_{i+1}) + \lambda_{i-\frac{1}{2}}(u_i - u_{i-1}) = h_i f_i.$$

In other words, we have (45), with $\lambda_{i,i\pm1} = \lambda_{i\pm\frac{1}{2}} > 0$, and $\lambda_{ij} = 0$ if $|i-j| > 1$. The proof is similar for boundary vertices, see equation (68). □

In addition to monotonicity, schemes with the LMP structure enjoy local stability properties as the nonoscillating property (Proposition 1.5 of [13]). In the present case, this reads as follows. Let $f = 0$ and $\mathbf{u}$ be a solution to the symmetric scheme; we have $\forall i \in [\![2,n-1]\!]$, $\min(u_{i-1}, u_{i+1}) \leq u_i \leq \max(u_{i-1}, u_{i+1})$, $\min(u_{\frac{1}{2}}, u_2) \leq u_1 \leq \max(u_{\frac{1}{2}}, u_2)$, and $\min(u_{n-1}, u_{n+\frac{1}{2}}) \leq u_n \leq \max(u_{n-1}, u_{n+\frac{1}{2}})$. Another very interesting property, the preservation of initial bounds (Proposition 1.6 of [13]), holds for the parabolic version of the scheme.

2.3 Consistency of the fluxes

In order to state the following result (Proposition 2.14), we need to assume that the interpolation matrix M_k defined by (15) satisfies some regularity assumption in the limit $h \to 0$. Loosely speaking, we expect column j of M_k to be of order h^j. More precisely, we assume that

$$M_k = N_k \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & h & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & h^k \end{pmatrix}, \tag{47}$$

where the matrix N_k converges as $h \to 0$, the limit N_k^0 being invertible:

$$\lim_{h\to 0} N_k = N_k^0, \quad \det(N_k^0) \neq 0. \tag{48}$$

Remark 2.13. *Assumption (47)-(48) may be seen as a regularity assumption of the mesh. It is clearly satisfied by a regular mesh, for which an explicit computation gives (47), where the matrix N_k does not depend on h.*

We have the following result:

Proposition 2.14. *Let $k \in \mathbb{N}^*$ and $\{x_i\}_{1\leq i\leq n}$ be a mesh satisfying (4), (5), (47) and (48) Let $\bar{u} \in \mathcal{C}^{k+1}(\Omega)$. The fluxes defined by (13) are consistent of order k. More precisely, the vector $\bar{\mathbf{u}}$ being defined by (9), we have*

$$\left| \mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) - \kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) \right| \leq C_1 \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^k,$$

where the constant C_1 depends only on k, on the constant C in (5) and on the norm of the matrix $\left(N_k^0\right)^{-1}$, where N_k^0 appears in (47)-(48). In particular it does not depend on $\bar{u}$ nor on i.

Proof. Since $\bar{u} \in \mathcal{C}^{k+1}(\Omega)$, a Taylor expansion gives

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

where Q is the k-th order polynomial

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

such that

$$\frac{d^\ell Q}{dx^\ell}(x_{i+\frac{1}{2}}) = \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}), \quad \forall \ell \in [\![1, k]\!]. \tag{49}$$

The remainder ρ satisfies the estimate

$$|\rho(x)| \leq \frac{1}{(k+1)!} \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} \left| x - x_{i+\frac{1}{2}} \right|^{k+1}. \tag{50}$$

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

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{Q}) + \mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}) = \kappa_{i+\frac{1}{2}} Q'(x_{i+\frac{1}{2}}) + \mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}) = \kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + \mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}),$$

where $\mathbf{Q}$ (resp. $\boldsymbol{\rho}$) is the vector defined as $\bar{\mathbf{u}}$ with the function Q (resp. ρ) instead of $\bar{u}$ (see (9)). Here, we have used first that the flux is linear, second that it is exact for polynomials of degree k (see Appendix B), and finally (49) with $\ell = 1$.

Proving the result thus amounts to show that $\left|\mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho})\right| \leq Ch^k$. To this end, we write it as follows

$$\mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}) = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \end{pmatrix} M_k^{-1} \boldsymbol{\rho},$$

and use (47)-(48)

$$\mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}) = \begin{pmatrix} 0 & h^{-1} & 0 & \dots & 0 \end{pmatrix} N_k^{-1} \boldsymbol{\rho}.$$

It is clear from estimate (50) that for each index ℓ, we have

$$|\rho_\ell| \leq C_k \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k+1},$$

where C_k depends only on k and on the constant appearing in (5). Hence,

$$\left|\mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho})\right| \leq C_k \left\| N_k^{-1} \right\| \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^k.$$

Finally, property (48) allows to prove that $\left\| N_k^{-1} \right\|$ is bounded independently of h, at least for h small enough. This concludes the proof. □

Remark 2.15. *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

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

$$\begin{cases} -\dfrac{d}{dx}\left(\kappa \dfrac{d\bar{u}}{dx}\right) + \alpha \bar{u} = f & \text{in } \Omega, \\ \kappa \dfrac{d\bar{u}}{dn} = 0 & \text{on } \partial\Omega. \end{cases} \tag{51}$$

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.

2.4.1 Convergence at the order $k-1$

The scheme reads as

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u}) + \alpha h_i u_i = h_i f_i, \qquad \forall i \in [\![1, n]\!], \tag{52}$$

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

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) \right) = \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) u_{i+1} - \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i, \tag{53}$$

and

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = 0. \tag{54}$$

In order to state our convergence result, we need the following stability property:

Assumption 2.16. *If $\mathbf{b} \geq 0$ and $A\mathbf{u} = \mathbf{b}$, with $b_i = h_i f_i, \forall i$, then $\forall i, u_i^- \leq C(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1))$, where u_i^- is the negative part of u_i and $C > 0$ a constant independent of h, $\mathbf{b}$ and $\mathbf{u}$.*

This assumption is a stability hypothesis similar to the one presented in Proposition 3.3 of [13].

Note that, if the scheme is convergent of order $\frac{1}{2}$, then Assumption 2.16 is satisfied. Let us be more precise: we assume that, denoting by $\bar{u}$ the exact solution and $\mathbf{u}$ the numerical one, we have

$$\|\mathbf{u} - \bar{\mathbf{u}}\|_{L^2} \leq C\sqrt{h}(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1)),$$

where the vector $\bar{\mathbf{u}}$ is defined by (9), the vector $\mathbf{f}$ is defined by (10), and C is a universal constant. Assuming that $f \geq 0$, we have $\bar{u} \geq 0$, and this estimate implies

$$\sum_{u_i < 0} h_i \left(u_i - \bar{u}_i\right)^2 + \sum_{u_i \geq 0} h_i \left(u_i - \bar{u}_i\right)^2 \leq Ch(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1))^2.$$

The second term in the right-hand side is non-negative, and, when $u_i < 0$, $\left(u_i - \overline{u}_i\right)^2 = \left(-u_i^- - \overline{u}_i\right)^2 \geq \left(u_i^-\right)^2$. Hence,

$$\sum_{i=1}^{n} h_i \left(u_i^-\right)^2 \leq C^2 h(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1))^2.$$

Using (5), we infer that $u_i^- \leq C(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1))$, that is, Assumption 2.16.

We now prove the following convergence result.

Proposition 2.17 (Convergence at order $k-1$ in L^1 norm)**.** *Let $k \in \mathbb{N}^*$, $\bar{u} \in \mathcal{C}^{k+1}(\Omega)$ be the exact solution of (51) 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 (52)-(53)-(54). Assume that Assumption 2.16 is satisfied. Then, we have*

$$\|\mathbf{e}\|_{L^1} \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k-1},$$

with $\|\cdot\|_{L^1}$ defined by (6), and C does not depend on h nor on $\bar{u}$, $\mathbf{u}$.

Proof. On the one hand the numerical flux defined by (53) satisfies (52) 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 (52) 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

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \bar{\mathcal{F}}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}}, \qquad \forall i \in [\![1, n,]\!] \quad \text{with } |R_{i+\frac{1}{2}}| \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^k, \text{ where } k \text{ is the order.} \tag{55}$$

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 $\Delta = \frac{1}{\alpha} \max\limits_{1 \leq i \leq n} \left(\dfrac{R_{i+\frac{1}{2}} - R_{i-\frac{1}{2}}}{h_i} \right) \in \mathbb{R}^+$, that is to say $0 \leq \Delta \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k-1}$ 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

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

with

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

Using Assumption 2.16, we can deduce that

$$(e_i + \Delta)^- \leq \left\| \frac{1}{h_i} \left(-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} \right) + \alpha \Delta \right\|_{L^2} \leq \left\| \frac{1}{h_i} \left(-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} \right) \right\|_{L^2} + \alpha |\Delta| \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k-1}. \tag{56}$$

Summing these inequalities over i, we obtain

$$\sum_{i=1}^{n} h_i \, (e_i + \Delta)^- \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k-1}. \tag{57}$$

Next, we sum the equalities $-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{e}) + \alpha h_i(e_i + \Delta) = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta$, finding

$$\left| \alpha \sum_{i=1}^{n} h_i \, (e_i + \Delta) \right| \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k-1} + \alpha \Delta \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k-1},$$

where we have used (55) and the above bound on Δ. Since $e_i + \Delta = (e_i + \Delta)^+ - (e_i + \Delta)^-$, this implies

$$\alpha \sum_{i=1}^{n} h_i \, (e_i + \Delta)^+ \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k-1} + \alpha \sum_{i=1}^{n} h_i \, (e_i + \Delta)^-$$

Using (57), we conclude that

$$\sum_{i=1}^{n} h_i \left(e_i + \Delta\right)^+ \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^{k-1}. \tag{58}$$

Collecting (57) and (58), we conclude the proof.

□

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 $\mathbf{f} \in H_E$

$$f_{\frac{1}{2}} = f_{n+\frac{1}{2}} = 0. \tag{59}$$

Let us define the scalar products

$$\begin{cases} (\mathbf{u}|\mathbf{v})_{H_M} = \displaystyle\sum_{i=1}^{n} h_i u_i v_i, \\ (\mathbf{f}|\mathbf{g})_{H_E} = \displaystyle\sum_{i=1}^{n-1} h_{i+\frac{1}{2}} f_{i+\frac{1}{2}} g_{i+\frac{1}{2}}, \end{cases} \tag{60}$$

and the operators

$$\begin{cases} D : H_M \longrightarrow H_E \text{ defined by} & (D\mathbf{u})_{i+\frac{1}{2}} = \dfrac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}, \qquad 1 \leq i \leq n-1, \\ D^* : H_E \longrightarrow H_M \text{ defined by} & (D^*\mathbf{f})_i = -\dfrac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{h_i}, \quad 1 \leq i \leq n. \end{cases} \tag{61}$$

Proposition 2.18. *If condition (59) 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 (59) 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 2.19 (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 (51) 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 (14) 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*

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

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 (53) and (54), 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

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \bar{\mathcal{F}}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}}, \qquad \text{with } |R_{i+\frac{1}{2}}| \leq Ch^k, \qquad \forall i \in [\![1, n]\!]. \tag{62}$$

Thus, we have

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

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 (59), so $\boldsymbol{\kappa}(D\mathbf{e} + \mathbf{r}(\mathbf{e})) - \mathbf{R}$ satisfies (59) too.

Using Proposition 2.18 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

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

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

$$\|\boldsymbol{\kappa}(D\mathbf{e} + \mathbf{r}(\mathbf{e}))\|_{H_E}^2 + (\alpha \mathbf{e} | \mathbf{v})_{H_M} \leq \|\mathbf{R}\|_{H_E} \|\boldsymbol{\kappa}(D\mathbf{e} + \mathbf{r}(\mathbf{e}))\|_{H_E}. \tag{63}$$

Besides, we have

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

Replacing v_i by its expression leads to

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

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 (63), 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 (62) gives

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

Proposition 2.17 gives

$$\|\boldsymbol{\kappa}(D\mathbf{e} + \mathbf{r}(\mathbf{e}))\|_{H_E} \leq C h^k + \alpha C h^{k-1}. \tag{64}$$

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

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

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

2.4.3 Convergence at order k

Proposition 2.20 (Convergence at order k)**.** *Let $k \in \mathbb{N}^*$, $\bar{u} \in \mathcal{C}^{k+1}(\Omega)$ be the exact solution of (51) 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 (53)-(54). Assume that the matrix A defining this scheme is uniformly coercive, that is, there exists a constant $C_c > 0$ independent of h such that*

$$\forall \mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T A \mathbf{x} \geq C_c \|D\mathbf{x}\|_{L^2}^2,$$

where the operator D is defined by (61). Then, we have

$$\|\mathbf{e}\|_{L^2} \leq C \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k,$$

where the constant C does not depend on $\bar{u}$, $\mathbf{u}$, h.

Proof. As in the proof of Proposition 2.17, we use the consistency of the flux to obtain that

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

with $|R_{i+\frac{1}{2}}| \leq C \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k$. The corresponding matrix system writes

$$A\mathbf{e} = \mathbf{R},$$

with

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

Taking line i of the system $A\mathbf{e} = \mathbf{R}$, we multiply it by e_i and sum over i:

$$\mathbf{e}^T A \mathbf{e} = \sum_{i=1}^{n} \left(-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}\right) e_i$$

Using a discrete integration by parts, then the Cauchy-Schwarz inequality, we have:

$$\mathbf{e}^T A \mathbf{e} = \sum_{i=0}^{n-1} R_{i+\frac{1}{2}} h_{i+\frac{1}{2}} (D\mathbf{e})_{i+\frac{1}{2}} \leq \left(\sum_{i=0}^{n-1} h_{i+\frac{1}{2}} R_{i+\frac{1}{2}}^2\right)^{1/2} \left(\sum_{i=0}^{n-1} h_{i+\frac{1}{2}} (D\mathbf{e})_{i+\frac{1}{2}}^2\right)^{1/2} \leq C \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k \|D\mathbf{e}\|_{L^2}.$$

The coercivity condition then gives

$$C_c \|D\mathbf{e}\|_{L^2} \leq C \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k.$$

A discrete mean Poincaré inequality, proved in Lemma 10.2 of [16], writes

$$\sum_{i=1}^{n} h_i e_i^2 \leq C \sum_{i=0}^{n-1} h_{i+\frac{1}{2}} (D\mathbf{e})_{i+\frac{1}{2}}^2 + \frac{1}{|\Omega|} \left(\sum_{i=1}^{n} h_i e_i\right)^2.$$

Owing to conservativity, we have $\displaystyle\sum_{i=1}^{n} h_i e_i = 0$, hence

$$\|\mathbf{e}\|_{L^2}^2 = \sum_{i=1}^{n} h_i e_i \leq C \sum_{i=0}^{n-1} h_{i+\frac{1}{2}} (D\mathbf{e})_{i+\frac{1}{2}}^2 = C \|D\mathbf{e}\|_{L^2}^2.$$

Thus, we have

$$\|\mathbf{e}\|_{L^2} \leq C \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k,$$

which concludes the proof. □

2.4.4 Asymptotic behavior of the symmetry condition

Lemma 2.21. *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 (51) and assume that $\bar{u}\geq 0$. Let $\mathbf{u}\in\mathbb{R}^n$ be the solution of (52), (53) and (54) 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 (19) is asymptotically fulfilled as $h\to 0$.*

Proof. Proposition 2.19 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 2.17 that

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

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. □

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 (see also Figure 10 for an example), 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

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

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

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

that is

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})=\tilde{\alpha}_{i+\frac{1}{2}}(u_{i+1}-u_i)+\tilde{r}_{i+\frac{1}{2}}(\mathbf{u})\tag{65}$$

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, we can achieve monotonicity as in Section 1.3 and the 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 (65) 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 2.22. *For $k = 1$, the remainders $r^L_{i+\frac{1}{2}}(\mathbf{u})$ and $r^R_{i+\frac{1}{2}}(\mathbf{u})$ vanish, and we obtain the classical harmonic mean for the equivalent diffusion coefficient.*

Remark 2.23. *In the case of a discontinuous right hand side f, we use the same type of strategy. In such a case, the second derivative of the solution $\bar{u}$ is discontinuous. Thus, the reconstruction is made on each side of the discontinuity.*

3 Numerical experiments

Before giving numerical results, we explain how we deal with possibly vanishing Dirichlet boundary conditions. The definition of the nonlinear scheme requires $\mathbf{u} > 0$ (which is enforced by construction, see Prop.2.5), and $g(x_{\frac{1}{2}}) > 0$ and $g(x_{\frac{1}{2}}) > 0$ for Dirichlet boundary conditions (see Sec. 1.5.1). However, we want to be able to deal with homogeneous Dirichlet boundary conditions. In order to circumvent this difficulty, it is possible to add a term proportional to h^k to the denominator in the flux. Let $\epsilon > 0$, the flux (21) is given by[2]

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

Same modification is made if needed for $\mathcal{F}_{\frac{1}{2}}$. We use also a correction to prevent the denominator of (19) to be zero. The condition (19) is replaced with

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

The L^2 norm of the error is computed as

$$e_{L^2} = \left(\sum_{i=1}^{n} h_i |u_i - \bar{u}_i|^2 \right)^{1/2}$$

for the solution, and

$$f_{L^2} = \left(e^2_{L^2} + \sum_{i=0}^{n} h_{i+\frac{1}{2}} |\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \bar{\mathcal{F}}(x_{i+\frac{1}{2}})|^2 \right)^{1/2} \tag{66}$$

for the flux.

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

$$\begin{cases} -\dfrac{d}{dx}\left(\kappa \dfrac{d\bar{u}}{dx}\right) = f & \text{in } \Omega, \\ \bar{u} = g & \text{on } \partial\Omega. \end{cases} \tag{67}$$

We will use three types of meshes:

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

1. Cartesian meshes,
2. deformed meshes, the deformation of which is given by: $x \to x + 0.65x(1-x)(0.5-x)\sin(0.8\pi)$,
3. random meshes, the deformation of which is given by: $x \to x + \dfrac{\eta}{n}$, with $\eta \in [-0.45, 0.45]$, and n the number of cells. Thus, $C = 19$ for inequality (5). An example of which with 8 cells being given in Figure 2.

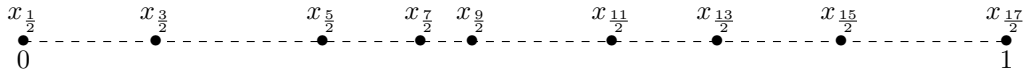

Figure 2: An example of a random mesh with 8 cells.

Figure 3 gives an example of the repartition of the cell volumes for a random mesh with 64 cells.

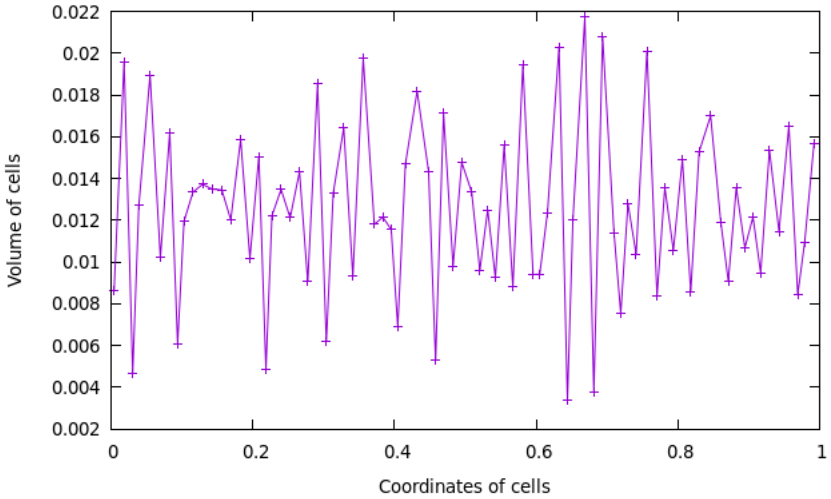

Figure 3: Example of a repartition of the volume for a random mesh with 64 cells.

For all the tests, the ε and $\mathbf{u}^0$ of the fixed-point algorithm (43) are $\varepsilon = 10^{-12}$ and $u_i^0 = 1, \forall i$. We use the linear solver GMRES with the preconditioner ILU (see [26], Chapter 7.4) and the convergence criterion is 10^{-14}.

3.1 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 (67). We perform a spectral 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.
The proof of exactness for polynomial of degree k (see appendix B) 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).

3.2 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 (67). We perform a convergence study for this problem with the non-symmetric and symmetric schemes on the deformed mesh. The L^2-error between the exact $\bar{u}$ and approximated u solutions and f_{L^2} (refer to Eq. (66)) are reported in Figures 4.

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	7.07e-09
7	4.21e-15	5.24e-10
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.

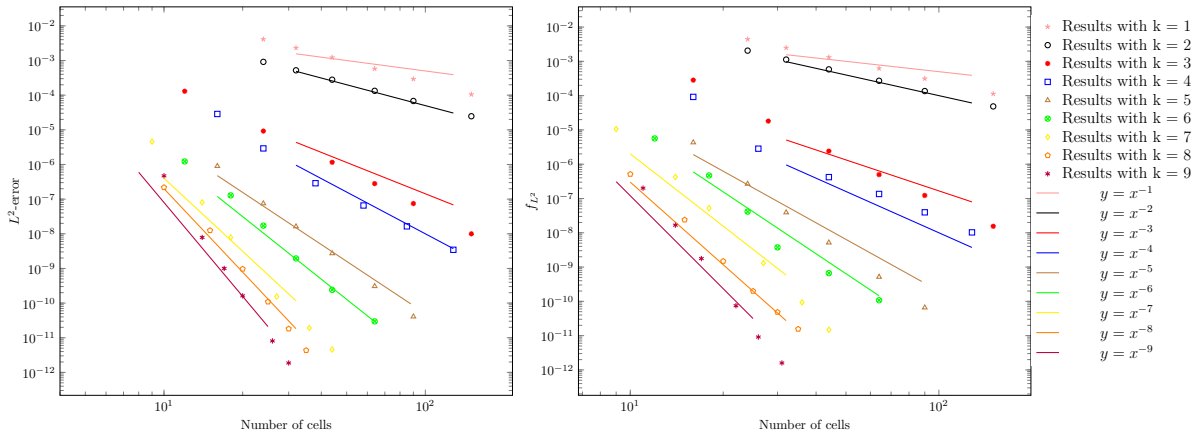

Figure 4: L^2-error, at the left, and f_{L^2} (refer to Eq. (66)), at the right, with the non-symmetric scheme for problem of Sec. 3.2.

The results show that the numerical convergence order is at worst equal to the theoretical order k (for the theoretical order 4 one obtains convergence at order 4) or better (for the theoretical order 3 one obtains the order 4). Besides, the results are qualitatively the same for the symmetric case and for the non-symmetric case (the results are only given for the non symmetric case because the figures are similar). We observe similar convergence orders for e_{L^2} and f_{L^2}.

We also perform a convergence study for the same problem on the random mesh: see Figures 5 and 6. As for the deformed mesh, the results show that the numerical convergence order is at worst equal to the theoretical order k (for the theoretical order 4 one obtains convergence at order 4) or better (for the theoretical order 3 one obtains convergence at order 4). The results are similar for the symmetric case and for the non-symmetric case (the results are only given for the non symmetric case). We observe similar convergence orders for e_{L^2} and f_{L^2}. However, the curves are slightly translated: for a given mesh size, the error is larger when the mesh is deformed. This is illustrated on the Figure 9 for the fourth-order non-symmetric scheme.

Figures 7 and 8 show that the number of iterations of the fixed-point algorithm depends weakly on the number of cells. This is especially visible in the Figure 7. Besides, for a random mesh, the number of iterations (for the fixed-point algorithm to reach stagnation) is significantly larger than for a deformed mesh.

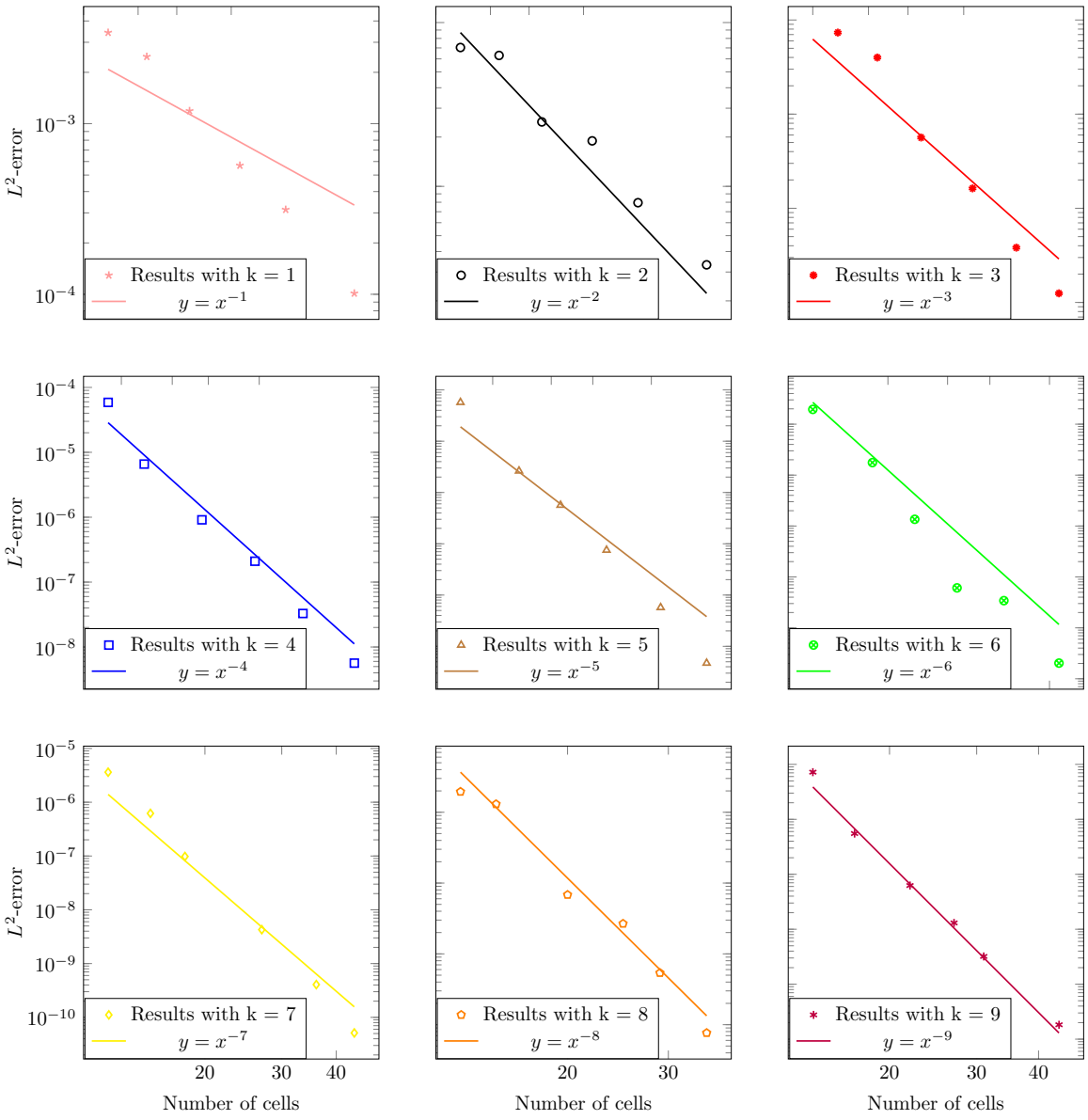

Figure 5: L^2-error with non-symmetric scheme and random mesh for problem of Sec. 3.2.

3.3 Comparison with a non-monotonic scheme

To show the effect of the monotonicity correction, we compare our scheme with a non-monotonicity preserving scheme.

Given $\kappa = 1$, $f = \pi^2 \sin(\pi x)$, $g(0) = g(1) = 0$, the function $\bar{u}(x) = \sin(\pi x)$ is solution to (67). We perform a monotonicity study for this problem on a Cartesian mesh with the third-order version for different grid sizes. Results are summarized in Table 2. Note that the non-monotonic scheme does not exhibit negative entries for all the grid resolutions, but when it happens, it is corrected with the monotonic version.

3.4 Discontinuous diffusion coefficient κ

Given κ such that

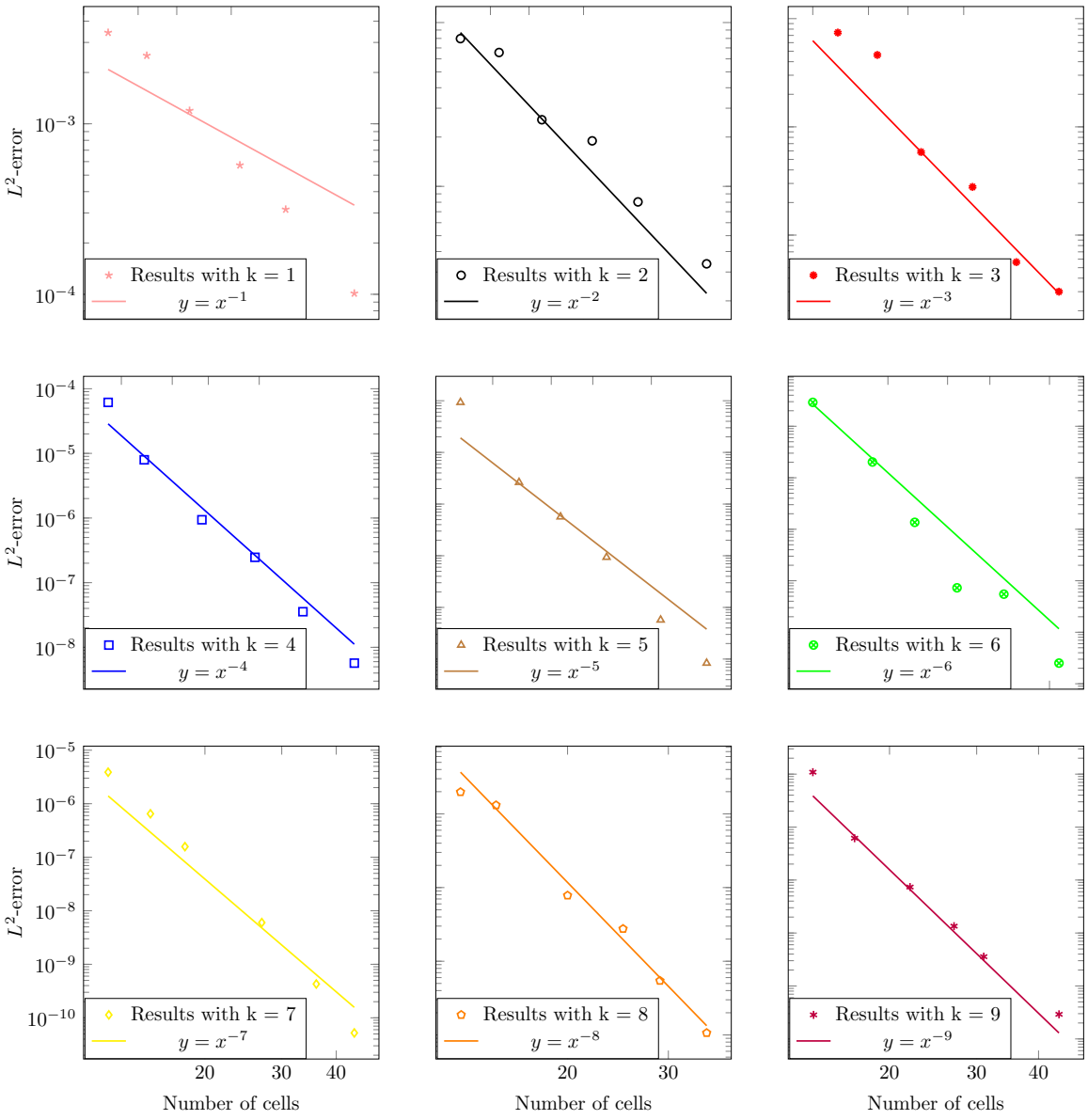

Figure 6: f_{L^2}-error with non-symmetric scheme and random mesh for problem of Sec. 3.2.

Number of cells	High order monotonic scheme	High order non-monotonic scheme
8	0	1
16	0	0
32	0	0
64	0	1
128	0	0

Table 2: Negative entries for the non-monotonic and the monotonic schemes.

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

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

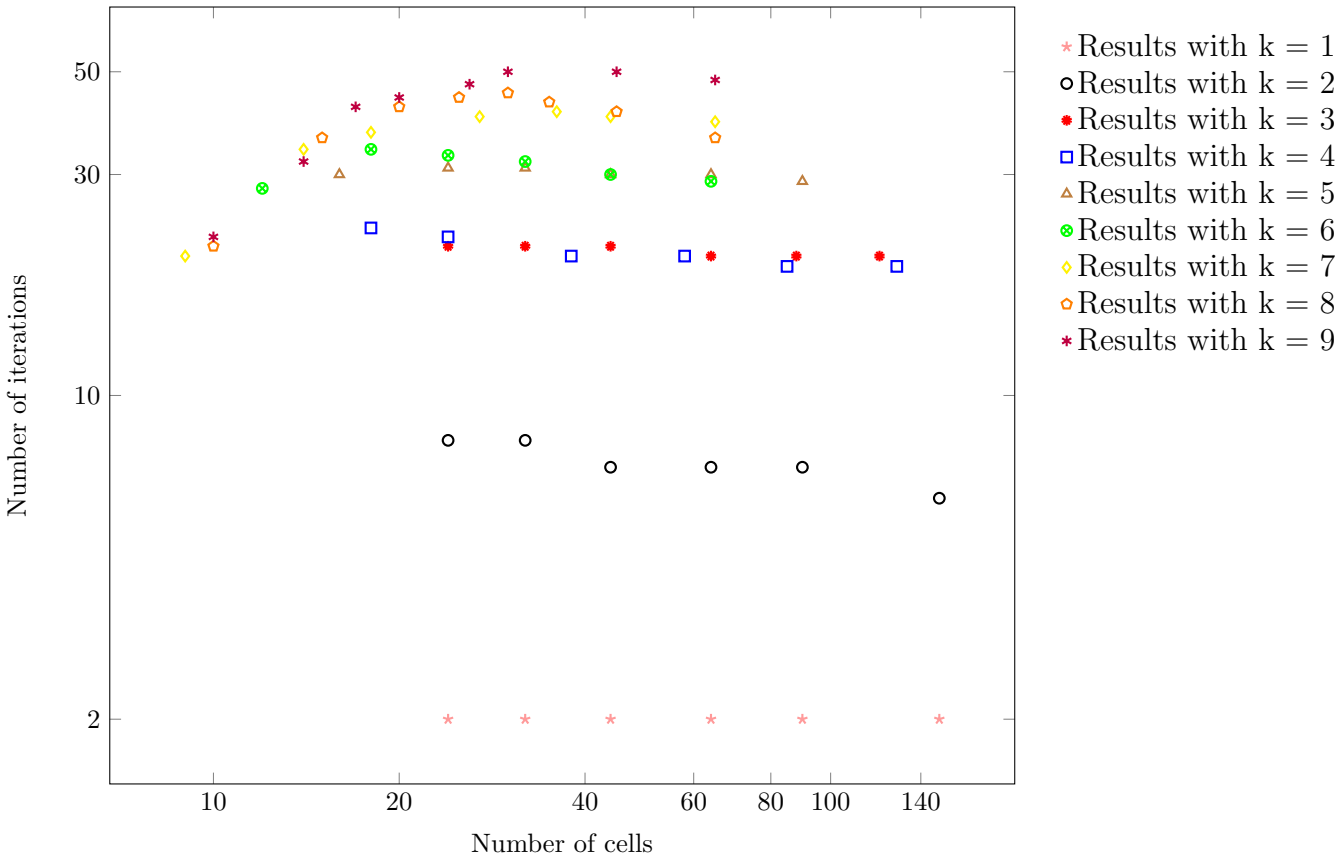

Figure 7: Number of iterations of the fixed point algorithm with the non-symmetric scheme for problem of Sec. 3.2 for a deformed mesh. The number of iteration of the fixed point algorithm increases with the order of convergence k, but is weakly affected by the mesh refinement.

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

is solution to (67). The solution of this problem is displayed on Figure 10. 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.

An even number of cells is required to have a node coinciding with the discontinuity of κ ($x = \frac{1}{2}$). Results are summarized in Figure 11. These graphs show that we achieve the expected convergence rate, even with discontinuous κ.

4 Concluding remarks

In this paper we have proposed an arbitrary-order monotonic 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 with reasonable assumptions on the mesh. We also proposed a symmetric 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. We are quite confident in the fact that our scheme can be extended to 2D because we used the same method to enforce monotonicity than Gao *et al* [17], who have applied it in the context of 2D diffusion on arbitrary meshes. To extend this method in 2D, we will need secondary unknowns. In order to compute them, several strategies are possible. Among others, one may use interpolation (see [9]), or a dual partition, in the spirit of the DDFV method (see [18]).

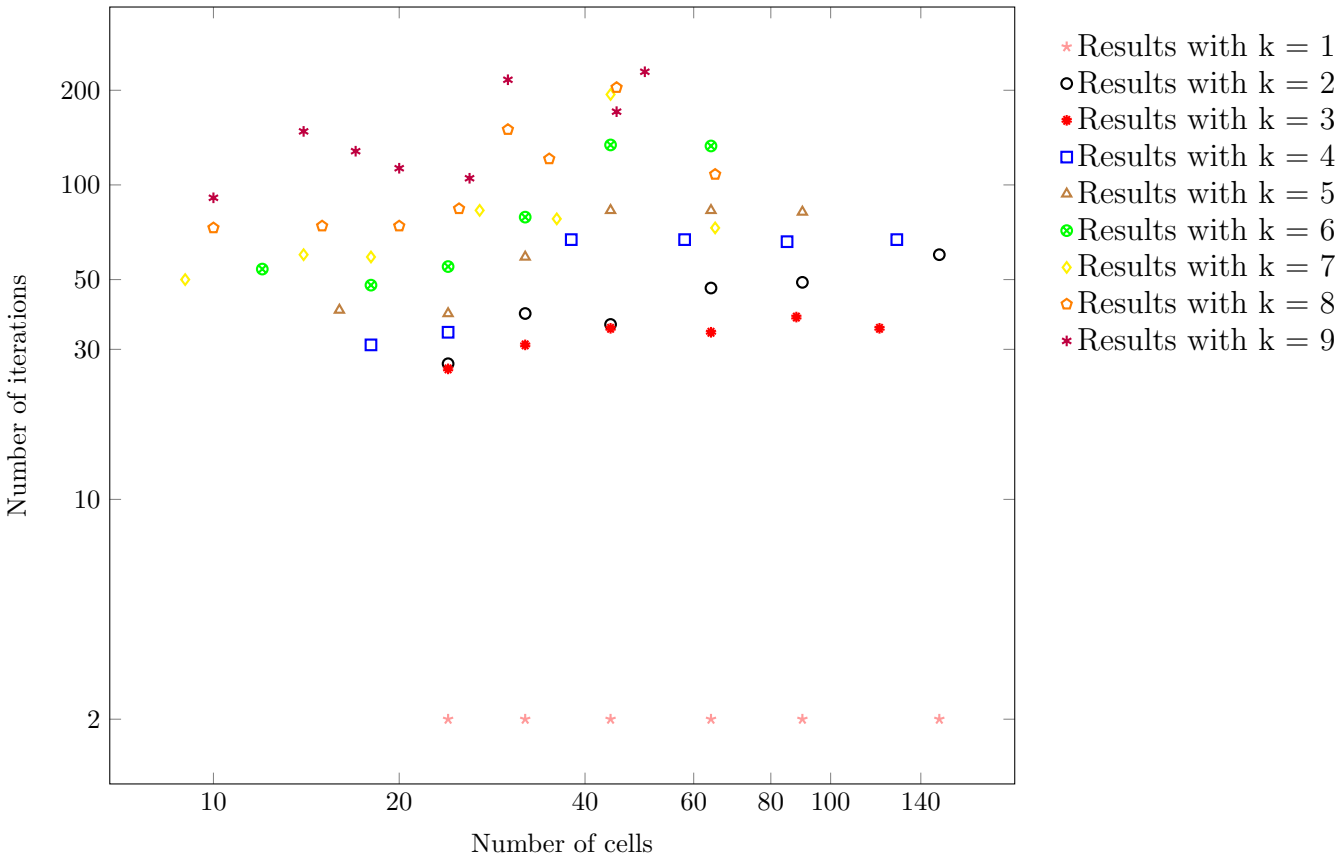

Figure 8: Number of iterations of the fixed point algorithm with the non-symmetric scheme for problem of Sec. 3.2 for a random mesh. The number of iteration of the fixed point algorithm increases with the order of convergence k, but is weakly affected by the mesh refinement.

Acknowledgement

The authors thank Christophe Buet, Clément Cancès, Stéphane Del Pino, Bruno Després 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 Dirichlet boundary conditions

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

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

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

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

that is to say

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

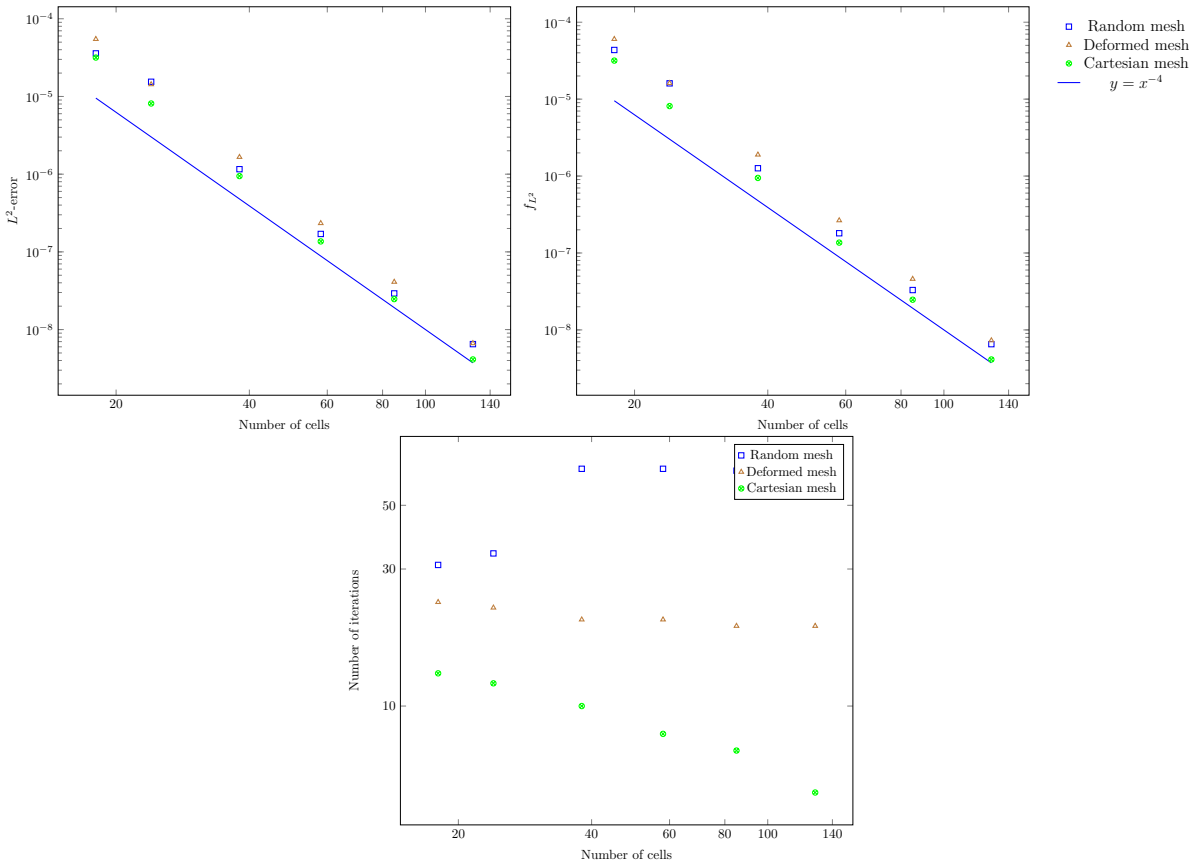

Figure 9: L^2-error, at the top left, and f_{L^2} (refer to Eq. (66)), at the top right, and number of iterations of the fixed point (bottom) with the non-symmetric scheme at order $k = 4$ for problem of Sec. 3.2. It shows that the mesh deformation impacts slightly the error, but strongly the number of fixed point iterations to achieve convergence.

from which we obtain

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

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

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right) u_{n+\frac{1}{2}} - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) u_n \right], \tag{68}$$

with

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

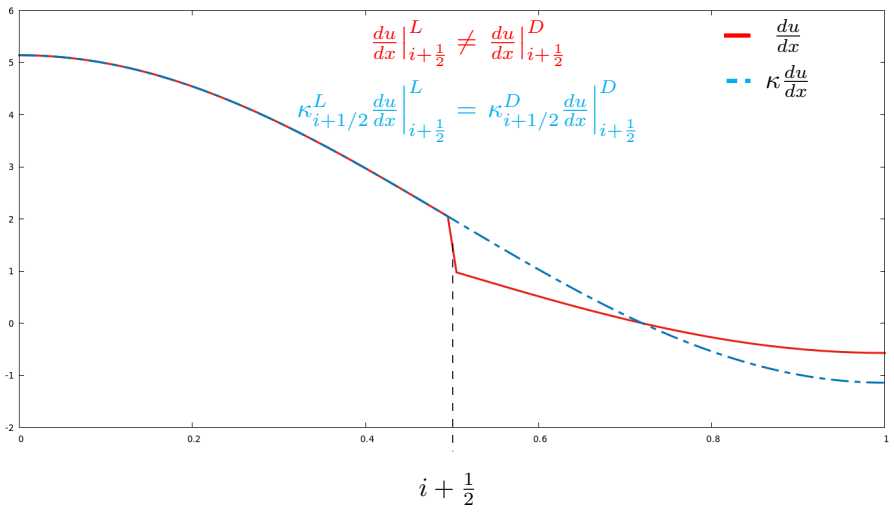

Figure 10: Illustration of problem of Sec 3.4. The diffusion being discontinuous, also is the gradient, but the flux remains continuous.

In order to preserve positivity, a condition similar to (19) 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, (68) becomes

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}}\left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})}\right) g(x_{n+\frac{1}{2}}) - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right) u_n\right]. \tag{69}$$

For the left boundary we obtain similarly

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}}\left[\left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1}\right) u_1 - \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})}\right) g(x_{\frac{1}{2}})\right]. \tag{70}$$

B 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

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

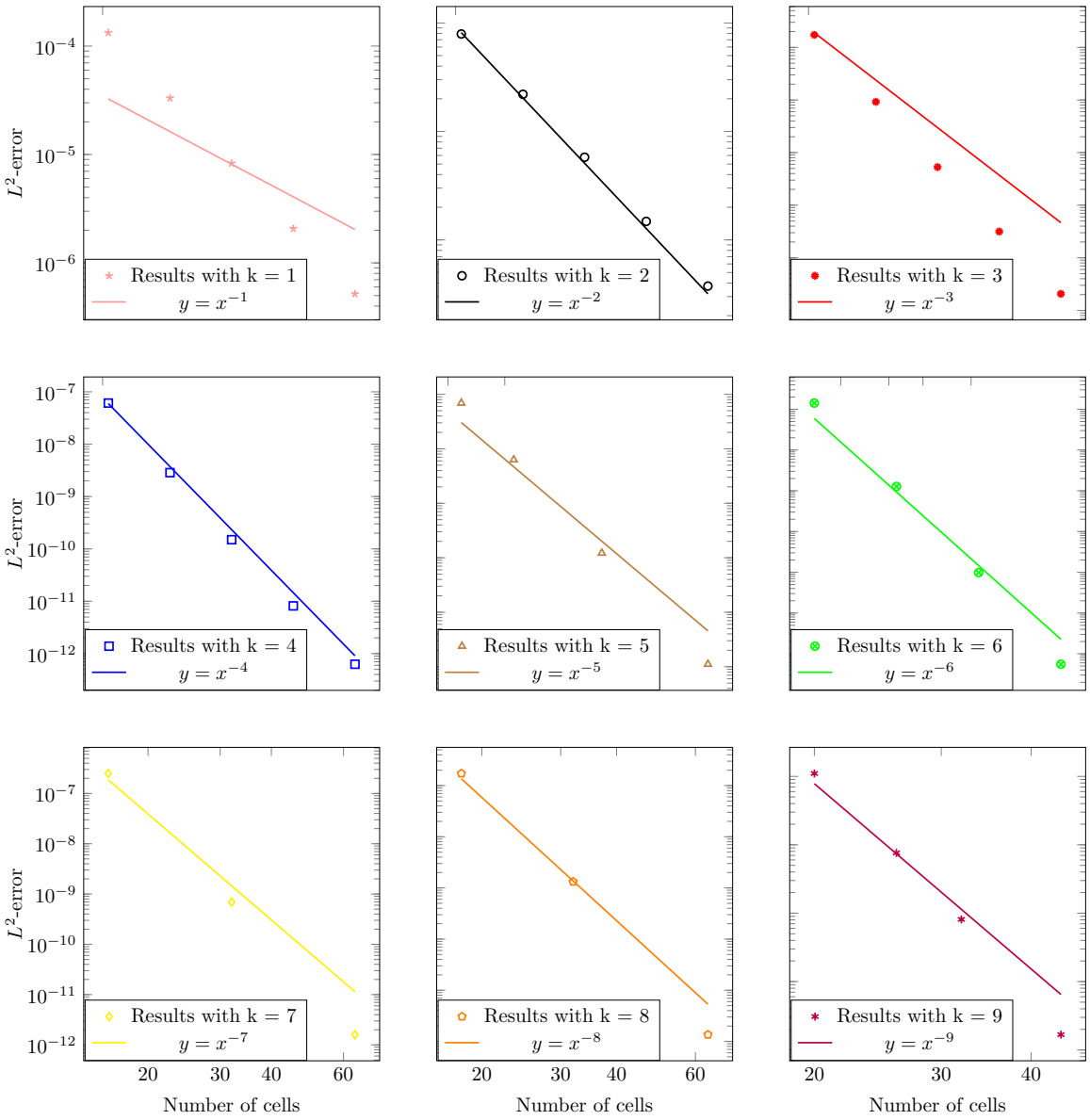

Figure 11: L^2-error with symmetric scheme and discontinuous κ for problem of Sec. 3.4.

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 is

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

where P is an interpolation polynomial of $\bar{u}$. Besides, $P = \bar{u}$ in that case since $\bar{u}$ is a polynomial of degree k and P leaves invariant polynomials of degree k. The flux 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. 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] E. Bertolazzi and G. 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] X. Blanc and E. Labourasse. A positive scheme for diffusion problems on deformed meshes. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 96(6):660–680, 2016.

[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] C. Cancès and C. Guichard. Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Found Comput Math*, 17:1525–1584, 2017.

[6] P. Ciarlet. Discrete maximum principle for finite-difference operators. *Aeq. Math.*, 4:338–352, 1970.

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

[8] P. Ciarlet and P.-A. Raviart. Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Found Comput Math*, 2:17–31, 1973.

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

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

[11] D. A. Di Pietro and J. Droniou. *The Hybrid High-Order method for polytopal meshes*, volume 19. Springer, 2020.

[12] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69. Springer, 2012.

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

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

[15] R. Eymard, T. Gallouët, C. Guichard, R. Herbin, and R. Masson. TP or not TP, that is the question. *Computational Geosciences*, 18(3-4):285–296, 2014.

[16] 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.

[17] Y. Gao, G. Yuan, S. Wang, and X. 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.

[18] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *Journal of Computational Physics*, 160(2):481–499, 2000.

[19] J. Karátson, S. Korotov, and M. Křížek. On discrete maximum principles for nonlinear elliptic problems. *Mathematics and Computers in Simulation*, 76(1):99–108, 2007. Mathematical Modelling and Computational Methods in Applied Sciences and Engineering.

[20] S. Korotov, M. Křížek, and P. Neittaanmäki. Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. *Mathematics of Computation*, 70(233):107–119, 2000.

[21] C. 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.

[22] C. 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.

[23] 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.

[24] 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.

[25] 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.

[26] Gerard Meurant. *Computer solution of large linear systems*. Elsevier, 1999.

[27] E. H. Quenjel. Enhanced positive vertex-centered finite volume scheme for anisotropic convection-diffusion equations. *ESAIM: M2AN*, 54(2):591–618, 2020.

[28] M. Schneider, L. Agélas, G. Enchéry, and B. Flemisch. Convergence of nonlinear finite volume schemes for heterogeneous anisotropic diffusion on general meshes. *Journal of Computational Physics*, 351:80–107, 2017.

[29] Z. Sheng and G. Yuan. A finite volume scheme for diffusion equations on distorted quadrilateral meshes. *Transport Theory Statist. Phys.*, 37(2-4):171–207, 2008.

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

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

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

[33] Tomáš Vejchodskỳ and Pavel Šolín. Discrete maximum principle for higher-order finite elements in 1d. *Mathematics of Computation*, 76(260):1833–1846, 2007.

[34] 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.

[35] H. Yang, B. Yu, Y. Li, and G. Yuan. Monotonicity correction for second order element finite volume methods of anisotropic diffusion problems. *Journal of Computational Physics*, 449:110759, 2022.

[36] 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.

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

[38] 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 für Angewandte Mathematik und Mechanik*, 100(5):1–25, 2020.