

### Contents lists available at ScienceDirect

# C. R. Acad. Sci. Paris. Ser. I

www.sciencedirect.com



## Analyse numérique

# Sur une correction non linéaire et un principe du minimum local pour la discrétisation d'opérateurs de diffusion en différences finies



# A nonlinear correction and local minimum principle for diffusion operators with finite differences

### Christophe Le Potier

CEA Saclay, DEN, DANS, DM2S, STMF, LMEC, 91191 Gif-sur-Yvette, France

#### INFO ARTICLE

Historique de l'article : Recu le 7 juillet 2017 Accepté après révision le 13 décembre 2017 Disponible sur Internet le 2 janvier 2018

Présenté par Olivier Pironneau

#### RÉSUMÉ

Nous proposons une nouvelle correction non linéaire pour la discrétisation d'un opérateur de diffusion anisotrope en différences finies. De plus, nous établissons que le schéma est convergent sans hypothèses spécifiques comme dans [2] ou [6].

© 2017 Académie des sciences. Publié par Elsevier Masson SAS. Cet article est publié en Open Access sous licence CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

#### ABSTRACT

We describe a nonlinear correction that suppresses oscillations appearing in the discretization of diffusion operators. We prove that the scheme is convergent without assumptions as in [2] or [6].

© 2017 Académie des sciences. Publié par Elsevier Masson SAS. Cet article est publié en Open Access sous licence CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

#### **Abridged English version**

It is well known that classical linear methods discretizing diffusion operators do not always satisfy the maximum principle for distorted meshes or high anisotropy ratios. A few years ago, nonlinear finite volume schemes were proposed to discretize the diffusion operators [6,13]. For these schemes, we obtained a discrete maximum principle for distorted meshes or highly anisotropic diffusion tensors. In the present work, we propose a proof for a nonlinear correction that gives nonoscillating solutions.

We consider problem (1) and a cell-centered finite volume scheme described in (2) on rectangular cells. Here,  $\mathcal{M}$  and  $\mathcal{E}$  are the meshes and the edges of the grid. We denote by  $|\Delta_x \Delta_y|$  the surface of a mesh. Then, we propose the nonlinear corrections  $\mathcal{R}^1_{i,j}(u)$ ,  $\mathcal{R}^2_{i,j}(u)$ ,  $\mathcal{R}^3_{i,j}(u)$ ,  $\mathcal{R}^4_{i,j}(u)$  and  $\mathcal{R}_\eta(u)$  described in (3). In Proposition 4.1, we show that the new method satisfies a local minimum principle.

https://doi.org/10.1016/j.crma.2017.12.004

Adresse e-mail: clepotier@cea.fr.

<sup>1631-073</sup>X/© 2017 Académie des sciences. Publié par Elsevier Masson SAS. Cet article est publié en Open Access sous licence CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

In Proposition 4.2, we show that the new scheme remains coercive without specific assumptions as in [6].

In Proposition 4.3, we explain why there exists at least one solution to the system (3).

In Proposition 4.4, without specific assumptions as in [2], we show the consistency of the scheme. We finally obtain the convergence of the new algorithm.

The numerical results are presented in the final section. We consider the analytical problem (11) and use the method described in [15]. We first show the  $L^2$  error with respect to the analytical solution, the order in space, the minimum values and the percentage of negative values as a function of the discretization step h (Scheme 1). Then we present the results obtained with the method presented in section 3 (Scheme 2), whereas two variants are described in section 5 (Schemes 3 and 4). We observe that there are no oscillations for all the modified methods. The most accurate is the scheme 4. Unfortunately, we are not able to prove that it is coercive. If it is not, we use the scheme 2 or the scheme 3, both of which are less accurate but which always converge.

#### 1. Introduction

Nous nous intéressons à la discrétisation d'une équation elliptique. Il est bien connu que les méthodes classiques discrétisant des opérateurs de diffusion ne satisfont pas toujours le principe du maximum pour des mailles très déformées ou des rapports d'anisotropie très élevés [10]. Nous proposons une nouvelle correction non linéaire qui donne des solutions non oscillantes pour des schémas volumes finis centrés sur des rectangles ou des parallélépipèdes rectangles. Elle peut s'appliquer, par exemple, en imagerie [16], où il est fondamental de bien approcher les opérateurs de diffusion anisotropes sur des grilles régulières. Elle se généralise aux schémas hybrides [7] ou aux schémas DDFV [5] en suivant la méthode décrite dans [12]. Nous établissons que la méthode est convergente sans ajout d'hypothèses spécifiques, comme dans [2] ou [6]. Notons également le travail développé dans [3], où les auteurs proposent un algorithme non linéaire qui conserve les bornes physiques et une preuve complète de convergence pour des schémas de type volumes éléments finis pour des équations paraboliques. Rappelons également l'analyse proposée dans [4] sur l'équation de la chaleur en dimension 1, pour une méthode d'ordre 4.

#### 2. Présentation

Nous considérons un domaine polygonal  $\Omega$  de  $\mathcal{R}^N$ . Nous nous intéressons au problème elliptique suivant :

$$\begin{cases} -\operatorname{div}\overline{D}\nabla u = f \operatorname{sur} \Omega, \\ u = 0 \operatorname{sur} \partial\Omega, \end{cases}$$
(1)

avec : u, la solution du problème dans  $H_0^1(\Omega)$ ,  $\overline{\overline{D}}(x)$ , une matrice (N, N) symétrique définie positive continue pour  $x \in \Omega$  telle que ses valeurs propres appartiennent à  $[\underline{\lambda}, \overline{\lambda}]$   $(0 < \underline{\lambda} \leq \overline{\lambda})$ ,  $f \in L^2(\Omega)$ , le second membre.

#### 3. Correction non linéaire

Nous décrivons la méthode en deux dimensions pour simplifier la présentation. Elle se généralise facilement en trois dimensions pour des parallélépipèdes rectangles.

Nous considérons un maillage  $\mathcal M$  structuré de  $\Omega$  caractérisé par l'ensemble de ses rectangles, de ses arêtes  $\mathcal E$  et de ses points  $\mathcal{P}$ , notés  $(x_{i,j})_{i,j\in\mathcal{M}}$  et  $(x_{\sigma})_{\sigma\in\mathcal{E}}$ . Nous notons :

- $\Delta_x \Delta_y$ , l'aire des rectangles,  $h = \max(\Delta_x, \Delta_y)$ ,  $\mathcal{E}_{\text{ext}}$ , l'ensemble des arêtes appartenant à  $\partial \Omega$ ,  $\mathcal{M}_{\partial}$  les mailles adjacentes aux arêtes de  $\mathcal{E}_{\text{ext}}$ ,

• 
$$\forall i, j \in \mathcal{M}, u_{i,j}$$
 la solution calculée sur la maille  $i, j$  et  $f_{i,j} = \frac{f_{i,j} f \, dx \, dy}{\Delta_x \Delta_y}$ 

- $\forall \sigma \in \mathcal{E}_{ext}$ ,  $u_{\sigma}$  la solution calculée sur l'arête  $\sigma$ ,
- $\mathcal{H}_{\mathcal{M}}$ , l'ensemble des fonctions constantes par mailles.

Nous nous donnons un schéma  $A_{i,i}(u)$ , linéaire, centré sur des rectangles structurés avec un stencil noté V(i, j), de type VF9 (voir [11]), constitué de mailles voisines de *i*, *j* et éventuellement d'arêtes situées sur le bord de  $\Omega$ . On suppose alors que  $\mathcal{A}_{i,j}(u)$  vérifie :  $\mathcal{A}_{i,j}(u) = \sum_{z \in V(i,j)} \alpha_{z,i,j}(u_z - u_{i,j}), \alpha_{z,i,j}$  étant des réels. Le schéma  $\mathcal{A}_{i,j}(u)$  vérifie :

$$\begin{cases} \forall i, j \in \mathcal{M}, -\mathcal{A}_{i,j}(u) = \Delta_x \Delta_y f_{i,j}, \\ \forall \sigma \in \mathcal{E}_{\text{ext}}, \ u_\sigma = 0. \end{cases}$$
(2)

Nous supposons aussi que la méthode est symétrique et vérifie les hypothèses de coercivité et de consistance décrites dans les paragraphes 2.2.2 et 2.2.3 dans [2] ou dans [14]. Nous notons  $\delta$  un entier égal à zéro ou 1,

$$\alpha_* = \max_{i,j \in \mathcal{M}, z \in V(i,j), \alpha_{z,i,j} < 0} 1.1 |\alpha_{z,i,j}|, \forall i, j \in \mathcal{M}, Ter_{i,j} = \frac{|\mathcal{A}_{i,j}(u)|}{\sum_{z \in V(i,j)} |u_z - u_{i,j}| + \beta \Delta_x \Delta_y \frac{f_{i,j}}{2B}}$$
 où  $\beta$  est un entier égal à

zéro ou 1 et où B > 0 est un réel (indépendant de u) vérifiant :  $\max_{i,j \in \mathcal{M}} \frac{|\mathcal{A}_{i,j}(u)|}{\sum_{z \in V(i,j)} |u_z - u_{i,j}|} \le B, \ \alpha_{**} = \max_{i,j \in \mathcal{M}} Ter_{i,j},$ 

$$\alpha_{i,j,k,l} = \max(Ter_{i,j}, Ter_{k,l})) \text{ et } Exp_{\alpha^*,\alpha_{**}} = ((1-\beta)\alpha_* + \beta\alpha_{**}). \text{ Posons :}$$

$$\Delta_{i,j,1} = ((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta\alpha_{i,j,i+1,j})(u_{i,j} - u_{i+1,j})^+ + ((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta\alpha_{i,j,i-1,j})(u_{i,j} - u_{i-1,j})^+,$$

$$\Delta_{i,j,2} = ((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta\alpha_{i,j,i,j+1})(u_{i,j} - u_{i,j+1})^+ + ((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta\alpha_{i,j,i,j-1})(u_{i,j} - u_{i,j-1})^+,$$

$$\Delta_{i,j,3} = ((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta\alpha_{i,j,i+1,j+1})(u_{i,j} - u_{i+1,j+1})^+ + ((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta\alpha_{i,j,i-1,j-1})(u_{i,j} - u_{i-1,j-1})^+$$

et

$$\Delta_{i,j,4} = ((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta\alpha_{i,j,i-1,j+1})(u_{i,j} - u_{i-1,j+1})^+ + ((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta\alpha_{i,j,i+1,j-1}(u_{i,j} - u_{i+1,j-1})^+,$$

où  $x^+ = \max(x, 0)$ . Par convention, si  $i, j \in \mathcal{M}_{\partial}$  et si les indices z de la forme (i + 1, j), (i - 1, j), (i, j + 1), (i, j - 1), (i + 1, j - 1), (i + 1, j - 1), (i + 1, j - 1) ou (i - 1, j + 1) n'appartiennent pas à  $\mathcal{M}$ , alors on choisit  $u_z = 0$ . Soit  $\eta > 0$ . Nous proposons de modifier le schéma précédent de la manière suivante :

$$\begin{aligned} \forall i, j \in \mathcal{M}, \\ -\mathcal{A}_{i,j}(u) + \left( (\Delta_{i,j,1} - \Delta_{i+1,j,1}) + (\Delta_{i,j,1} - \Delta_{i-1,j,1}) \right) + \left( (\Delta_{i,j,2} - \Delta_{i,j+1,2}) + (\Delta_{i,j,2} - \Delta_{i,j-1,2}) \right) \\ \left( (\Delta_{i,j,3} - \Delta_{i+1,j+1,3}) + (\Delta_{i,j,3} - \Delta_{i-1,j-1,3}) \right) + \left( (\Delta_{i,j,4} - \Delta_{i+1,j-1,4}) + (\Delta_{i,j,4} - \Delta_{i-1,j+1,4}) \right) \\ + \delta\eta \sum_{z \in V(i,j)} \min \left( (\Delta_x \Delta_y)^2, \frac{(\Delta_x \Delta_y)^2}{\sum_{y \in V(i,j)} |u_y - u_{i,j}|} + \frac{(\Delta_x \Delta_y)^2}{\sum_{y \in V(z)} |u_y - u_z|} \right) (u_{i,j} - u_z) \\ = -\mathcal{A}_{i,j}(u) + \mathcal{R}^1_{i,j}(u) + \mathcal{R}^2_{i,j}(u) + \mathcal{R}^3_{i,j}(u) + \mathcal{R}^4_{i,j}(u) + \mathcal{R}_\eta(u) = \Delta_x \Delta_y f_{i,j}, \\ \forall \sigma \in \mathcal{E}_{\text{ext}}, \ u_\sigma = 0, \end{aligned}$$

avec la convention,  $\Delta_{z,n} = -u_{i,j}((1-\delta)Exp_{\alpha^*,\alpha_{**}} + \delta Ter_{i,j})$  si les indices  $z \in (i+1, j), (i-1, j), (i, j+1), (i, j-1), (i+1, j+1), (i-1, j-1), (i+1, j-1)$  ou (i-1, j+1) n'appartiennent pas à  $\mathcal{M}$ .

#### 4. Propriétés de l'algorithme

Dans cette section, nous choisissons  $\delta = 0$  et  $\beta = 0$ .

Proposition 4.1. Le schéma modifié satisfait le principe du minimum local suivant :

$$\forall i, j \in \mathcal{M} \ u_{i,j} \geq \min_{z \in V(i,j)} u_z$$

Nous remarquons que le terme  $\mathcal{R}^1_{i,i}(u)$  s'écrit sous la forme :

$$\begin{cases} \mathcal{R}_{i,j}^{1}(u) = \alpha_{*}((u_{i,j} - u_{i+1,j}) + (u_{i,j} - u_{i-1,j}) - (u_{i-1,j} - u_{i-2,j})^{+} - (u_{i+1,j} - u_{i+2,j})^{+}) \\ (si \ (i+1, j) \ ou \ (i-1, j) \in \mathcal{M}), \end{cases}$$

$$\mathcal{R}_{i,j}^{1}(u) = \alpha_{*}((u_{i,j}^{+} + u_{i,j} + (u_{i,j} - u_{i-1,j}) - (u_{i-1,j} - u_{i-2,j})^{+}) \ (si \ (i+1, j) \notin \mathcal{M}), \end{cases}$$

$$\mathcal{R}_{i,j}^{1}(u) = \alpha_{*}((u_{i,j}^{+} + u_{i,j} + (u_{i,j} - u_{i+1,j}) - (u_{i+1,j} - u_{i+2,j})^{+}) \ (si \ (i-1, j) \notin \mathcal{M}).$$

$$\tag{4}$$

En utilisant un calcul similaire pour  $\mathcal{R}_{i,j}^2(u)$ ,  $\mathcal{R}_{i,j}^3(u)$  et  $\mathcal{R}_{i,j}^4(u)$ , le système de l'équation (3) se réécrit :

$$\begin{cases} \forall i, j \in \mathcal{M}, i, j \notin \mathcal{M}_{\partial}, \sum_{z \in V(i,j) \in \mathcal{M}} (\alpha_{z,i,j} + \alpha_{*})(u_{i,j} - u_{z}) = \Delta_{x} \Delta_{y} f_{i,j} \\ + \alpha_{*} \left( (u_{i-1,j} - u_{i-2,j})^{+} + (u_{i+1,j} - u_{i+2,j})^{+} + (u_{i,j-1} - u_{i,j-2})^{+} + (u_{i,j+1} + u_{i,j+2})^{+} \\ + (u_{i-1,j-1} - u_{i-2,j-2})^{+} + (u_{i+1,j+1} - u_{i+2,j+2})^{+} + (u_{i-1,j+1} - u_{i-2,j})^{+} + (u_{i+1,j-1} + u_{i+2,j})^{+} \right) = S_{i,j}(u), \\ \forall \sigma \in \mathcal{E}_{\text{ext}}, \ u_{\sigma} = 0. \end{cases}$$

(5)

Si  $i, j \in M_{\partial}$ , les calculs sont similaires. Comme les termes  $\alpha_{z,i,j} + \alpha_*$  sont strictement positifs, et que  $S_{i,j}(u) \ge 0$ , nous obtenons la propriété souhaitée.

#### Proposition 4.2. Le schéma modifié est coercif.

Nous écrivons :

$$Exp = \sum_{i,j \in \mathcal{M}} u_{i,j} \mathcal{R}^{1}_{i,j}(u) = \alpha_{*} \sum_{\substack{i,j \in \mathcal{M}, i,j \notin \mathcal{M}_{\partial}}} (u_{i,j} - u_{i-1,j}) (\Delta_{i,j,1} - \Delta_{i-1,j,1})$$
$$+ \alpha_{*} \sum_{\substack{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}}} u_{i,j} (\Delta_{i,j,1} + u_{i,j}).$$
(6)

Premier cas :  $(u_{i,j} - u_{i-1,j}) \ge 0$ . Les deux termes à contrôler  $T_{1,i}^+ = -(u_{i,j} - u_{i-1,j})(u_{i-1,j} - u_{i-2,j})^+$  et  $T_{2,i}^+ = (u_{i,j} - u_{i+1,j})(u_{i,j} - u_{i-1,j})^+$  vérifient :  $T_{1,i}^+ \ge -\frac{1}{2}(u_{i,j} - u_{i-1,j})^2 - \frac{1}{2}(u_{i-1,j} - u_{i-2,j})^2$  et  $T_{2,i}^+ \ge -\frac{1}{2}(u_{i,j} - u_{i+1,j})^2 - \frac{1}{2}(u_{i,j} - u_{i-1,j})^2$ .

Deuxième cas :  $(u_{i,j} - u_{i-1,j}) \le 0$ . Les termes à contrôler  $T_{2,i}^- = (u_{i,j} - u_{i-1,j})(u_{i,j} - u_{i+1,j})^+$  et  $T_{1,i}^- = -(u_{i-1,j} - u_{i-2,j})(u_{i-1,j} - u_{i,j})^+$  vérifient :  $T_{1,i}^- \ge -\frac{1}{2}(u_{i,j} - u_{i-1,j})^2 - \frac{1}{2}(u_{i-1,j} - u_{i-2,j})^2$  et  $T_{2,i}^- \ge -\frac{1}{2}(u_{i,j} - u_{i+1,j})^2 - \frac{1}{2}(u_{i,j} - u_{i-1,j})^2$ . On déduit :

$$\sum_{i,j \in \mathcal{M}, i,j \notin \mathcal{M}_{\partial}} \min(T_{1,i}^{+} + T_{2,i}^{+}, T_{1,i}^{-} + T_{2,i}^{-}) \geq -\sum_{i,j \in \mathcal{M}, i,j \notin \mathcal{M}_{\partial}} (u_{i,j} - u_{i-1,j})^{2} - \sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} u_{i,j}^{2}$$

Nous écrivons alors :

$$\sum_{i,j \in \mathcal{M}, i,j \notin \mathcal{M}_{\partial}} (u_{i,j} - u_{i-1,j}) (\Delta_{i,j,1} - \Delta_{i-1,j,1}) \ge \sum_{i,j \in \mathcal{M}, i,j \notin \mathcal{M}_{\partial}} (u_{i,j} - u_{i-1,j})^{2} + \sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} u_{i,j}^{2} + \sum_{i,j \in \mathcal{M}, i,j \notin \mathcal{M}_{\partial}} \min(T_{1,i}^{+} + T_{2,i}^{+}, T_{1,i}^{-} + T_{2,i}^{-}).$$
(7)

Nous déduisons que  $Exp \ge 0$ . Les calculs sont similaires pour les expressions  $\sum_{i,j \in \mathcal{M}} u_{i,j} \mathcal{R}_{i,j}^2(u)$ ,  $\sum_{i,j \in \mathcal{M}} u_{i,j} \mathcal{R}_{i,j}^3(u)$  et  $\sum_{i,j \in \mathcal{M}} u_{i,j} \mathcal{R}_{i,j}^4(u)$ . La méthode initiale étant coercive, nous obtenons le résultat souhaité.

#### **Proposition 4.3.** Il existe au moins une solution au système (3).

En utilisant la proposition 4.2, les termes  $\mathcal{R}_{i,j}^n(u)$  étant continus de  $\mathcal{H}_{\mathcal{M}}$  dans  $\mathcal{H}_{\mathcal{M}}$ , nous déduisons l'existence d'une  $1 \le n \le 4$  solution au système (3) en appliquant la proposition 6 dans [2].

#### Proposition 4.4. Le schéma modifié est convergent.

Nous reprenons les trois étapes  $C_1$ ,  $C_2$  et  $C_3$  décrites dans [8] (section 1.2). L'étape  $C_1$  est satisfaite avec la norme 2.6 décrite dans [8] (pour laquelle on a une inégalité de Poincaré) grâce à la proposition 4.2. L'étape  $C_2$  découle de résultats de compacité classiques (mentionnés dans le lemme B.19 du livre [9]). Il existe donc  $\bar{u} \in H_0^1(\Omega)$  telle qu'une sous-suite de u converge vers  $\bar{u}$  dans  $L^2(\Omega)$ . Nous nous concentrons sur l'étape  $C_3$  de [8]. Il suffit donc de montrer que  $\bar{u}$  est une solution faible de (1) pour obtenir la convergence de la méthode.

Nous nous donnons  $\varphi \in C_c^{\infty}(\Omega)$  et nous posons  $\varphi_D = (\varphi_{i,j})_{\forall i,j \in \mathcal{M}_i}$  avec  $\varphi_{i,j} = \varphi(x_{i,j})$ . En multipliant l'équation (2) par  $\varphi_{i,j}$  et en sommant sur toutes les mailles, nous obtenons :

$$-\sum_{i,j\in\mathcal{M}}\mathcal{A}_{i,j}(u)\varphi_{i,j} + \sum_{i,j\in\mathcal{M}}(\mathcal{R}^{1}_{i,j}(u) + \mathcal{R}^{2}_{i,j}(u) + \mathcal{R}^{3}_{i,j}(u) + \mathcal{R}^{4}_{i,j}(u))\varphi_{i,j} = \int_{\Omega}f\varphi_{\mathcal{D}}\,\mathrm{d}\Omega.$$
(8)

Nous nous intéressons au terme :

$$T = \sum_{i,j \in \mathcal{M}} \mathcal{R}^{i}_{i,j}(u)\varphi_{i,j} = \sum_{i,j \in \mathcal{M}_{\partial}} \Delta_{i,j,1}(2\varphi_{i,j} - \varphi_{i-1,j} - \varphi_{i+1,j}) + \sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} \Delta_{i,j,1}(2\varphi_{i,j} - \varphi_{i-1,j} - \varphi_{i+1,j}) + \sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} \varphi_{i,j}u_{i,j} = T_{1} + T_{2} + T_{3}.$$
 (9)

D'après la régularité de  $\varphi$ , nous déduisons qu'il existe une constante  $C_4$  (indépendante de h) telle que  $|2\varphi_{i,j} - \varphi_{i-1,j} - \varphi_{i+1,j}| \le C_4 h^2$ , car le maillage est constitué de rectangles. Nous obtenons donc :

$$|T_1| \leq \sqrt{\sum_{i,j\in\mathcal{M},\ i,j\notin\mathcal{M}_{\partial}} (\Delta_{i,j,1})^2 \sum_{i,j\in\mathcal{M},\ i,j\notin\mathcal{M}_{\partial}} (2\varphi_{i,j} - \varphi_{i-1,j} - \varphi_{i+1,j})^2} = \sqrt{T_4 T_5}.$$

Par coercivité du schéma modifié, le terme  $T_4$  est borné. Nous déduisons donc qu'il existe une constante  $C_5$  telle que  $|T_1| \le C_5 h$ . Par ailleurs, d'après la proposition 4.2, il existe une constante  $C_6$  telle que  $T_2$  et  $T_3$  vérifient :

$$|T_{2}| + |T_{3}| \leq \sqrt{\sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} (\Delta_{i,j,1})^{2} \sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} (2\varphi_{i,j} - \varphi_{i-1,j} - \varphi_{i+1,j})^{2} + \sqrt{\sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} u_{i,j}^{2} \sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} (2\varphi_{i,j} - \varphi_{i-1,j} - \varphi_{i+1,j})^{2}} + \sqrt{\sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} u_{i,j}^{2} \sum_{i,j \in \mathcal{M}_{\partial}, i-1, j \notin \mathcal{M} \text{ ou } i+1, j \notin \mathcal{M}} (2\varphi_{i,j} - \varphi_{i-1,j} - \varphi_{i+1,j})^{2}} + (10)$$

Le raisonnement reste le même pour les termes  $\mathcal{R}_{i,j}^2(u)$ ,  $\mathcal{R}_{i,j}^3(u)$  et  $\mathcal{R}_{i,j}^4(u)$ . Nous déduisons donc que  $\sum_{i,j\in\mathcal{M}} (\mathcal{R}_{i,j}^1(u) + \mathcal{R}_{i,j}^2(u) + \mathcal{R}_{i,j}^3(u) + \mathcal{R}_{i,j}^4(u))\varphi_{i,j} \to 0$  lorsque  $h \to 0$ . Donc, d'après la convergence du schéma initial,  $-\sum_{i,j\in\mathcal{M}} \mathcal{A}_{i,j}(u)\varphi_{i,j} + \sum_{i,j\in\mathcal{M}} (\mathcal{R}_{i,j}^1(u) + \mathcal{R}_{i,j}^2(u) + \mathcal{R}_{i,j}^3(u) + \mathcal{R}_{i,j}^4(u))\varphi_{i,j} \to 0$ 

$$\sum_{i,j\in\mathcal{M}} (\mathcal{R}^1_{i,j}(u) + \mathcal{R}^2_{i,j}(u) + \mathcal{R}^3_{i,j}(u)) + \mathcal{R}^4_{i,j}(u))\varphi_{i,j} \to \int_{\Omega} \overline{D} \nabla \bar{u} \nabla \varphi \text{ lorsque } h \to 0.$$

**Remarque 1.** En dimension 1, les algorithmes utilisés dans [4] sont du type (avec  $\alpha \ge 0$ ) :  $(\alpha + 1)(-u_{i+1} + 2u_i + u_{i-1}) - \alpha \frac{(-u_{i+2} + 2u_i + u_{i-2})}{4} = \Delta_x^2 f_i$ , (pour  $\alpha = \frac{1}{3}$ , comme montré dans [4], la méthode est d'ordre 4). En choisissant  $\Delta_i = (u_i - u_{i+2})^+ + (u_i - u_{i-2})^+$  et en corrigeant avec l'expression  $\mathcal{R}_i(u) = \alpha \frac{1.1}{4} \left( (\Delta_i - \Delta_{i+2}) + (\Delta_i - \Delta_{i-2}) \right)$ , les quatre propositions précédentes restent vraies avec le même type de calcul et le même type de raisonnement.

**Remarque 2.** Si la méthode est de type VF25 (voir [1]) sur des rectangles, pour obtenir un principe du minimum local, on ajoute à l'équation (3) des termes correctifs du type :

$$\begin{split} \mathcal{R}_{i,j}^{5}(u) &= \alpha_{*} \bigg( (\Delta_{i,j,5} - \Delta_{i+2,j,5}) + (\Delta_{i,j,5} - \Delta_{i-2,j,5}) \bigg) \text{ avec } \Delta_{i,j,5} = (u_{i,j} - u_{i+2,j})^{+} + (u_{i,j} - u_{i-2,j})^{+}, \\ \mathcal{R}_{i,j}^{6}(u) &= \alpha_{*} \bigg( (\Delta_{i,j,6} - \Delta_{i,j+2,6}) + (\Delta_{i,j,6} - \Delta_{i,j-2,6}) \bigg) \text{ avec } \Delta_{i,j,6} = (u_{i,j} - u_{i,j+2})^{+} + (u_{i,j} - u_{i,j-2})^{+}, \\ \mathcal{R}_{i,j}^{7}(u) &= \alpha_{*} \bigg( (\Delta_{i,j,7} - \Delta_{i+2,j+2,7}) + (\Delta_{i,j,7} - \Delta_{i-2,j-2,7}) \bigg) \text{ avec } \Delta_{i,j,7} = (u_{i,j} - u_{i+2,j+2})^{+} + (u_{i,j} - u_{i-2,j-2})^{+}, \\ \mathcal{R}_{i,j}^{8}(u) &= \alpha_{*} \bigg( (\Delta_{i,j,8} - \Delta_{i-2,j+2,8}) + (\Delta_{i,j,8} - \Delta_{i+2,j-2,8}) \bigg) \text{ avec } \Delta_{i,j,8} = (u_{i,j} - u_{i+2,j-2})^{+} + (u_{i,j} - u_{i-2,j+2})^{+}, \\ \mathcal{R}_{i,j}^{9}(u) &= \alpha_{*} \bigg( (\Delta_{i,j,9} - \Delta_{i+2,j+1,9}) + (\Delta_{i,j,9} - \Delta_{i-2,j-1,9}) \bigg) \text{ avec } \Delta_{i,j,9} = (u_{i,j} - u_{i+2,j+1})^{+} + (u_{i,j} - u_{i-2,j-1})^{+}, \\ \mathcal{R}_{i,j}^{10}(u) &= \alpha_{*} \bigg( (\Delta_{i,j,10} - \Delta_{i+2,j-1,10}) + (\Delta_{i,j,10} - \Delta_{i-2,j+1,10}) \bigg) \text{ avec } \Delta_{i,j,9} = (u_{i,j} - u_{i+2,j+1})^{+} + (u_{i,j} - u_{i-2,j-1})^{+}, \\ \mathcal{R}_{i,j,10}^{10} &= (u_{i,j} - u_{i+2,j-1})^{+} + (u_{i,j} - u_{i-2,j+1})^{+}, \end{split}$$

Resultats obtenus avec les schemas 1, 2 et 3 en fonction du pas en espace.					
h	1 8	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
Erreur L <sup>2</sup> (schéma 1)	$5,21 \times 10^{-1}$	$1,96  imes 10^{-1}$	$7,14 imes10^{-2}$	$1,65  imes 10^{-2}$	$2,14 imes10^{-3}$
Ordre (schéma 1)		1,41	1,46	2,11	2,95
Val. négatives (schéma 1)	12,5	9,38	5,46	2,14	0,53
Val. min (schéma 1)	$-2,9  imes 10^{-1}$	$-2,4 imes10^{-1}$	$-1.4  imes 10^{-1}$	$-5,26  imes 10^{-2}$	$-1,33 \times 10^{-1}$
Erreur $L^2$ (schéma 2)	$3,70  imes 10^{-1}$	$3,21 \times 10^{-1}$	$2,77 \times 10^{-1}$	$2,21 \times 10^{-1}$	$1,50 \times 10^{-1}$
Ordre (schéma 2)		0,20	0,21	0,32	0,56
Erreur $L^2$ (schéma 3)	$1,84 \times 10^{-1}$	$1,57 \times 10^{-1}$	$1,05 \times 10^{-1}$	$6,00 \times 10^{-2}$	$3,28 \times 10^{-2}$
Ordre (schéma 3)		0,23	0,58	0,80	0,87
Erreur $L^2$ (schéma 4)	$1,23 \times 10^{-1}$	$4,22 \times 10^{-2}$	$1,33 \times 10^{-2}$	$3,83 \times 10^{-3}$	$1,02 \times 10^{-3}$
Ordro (schóma 4)		154	166	1 90	101

 Tableau 1

 Résultats obtenus avec les schémas 1, 2 et 3 en fonction du pas en espace.

$$\begin{aligned} \mathcal{R}_{i,j}^{11}(u) &= \alpha_* \bigg( (\Delta_{i,j,11} - \Delta_{i+1,j+2,11}) + (\Delta_{i,j,11} - \Delta_{i-1,j-2,11}) \bigg) \text{ avec} \\ \Delta_{i,j,11} &= (u_{i,j} - u_{i+1,j+2})^+ + (u_{i,j} - u_{i-1,j-2})^+, \\ \mathcal{R}_{i,j}^{12}(u) &= \alpha_* \bigg( (\Delta_{i,j,12} - \Delta_{i-1,j+2,12}) + (\Delta_{i,j,12} - \Delta_{i+1,j-2,12}) \bigg) \text{ avec} \\ \Delta_{i,j,12} &= (u_{i,j} - u_{i-1,j+2})^+ + (u_{i,j} - u_{i+1,j-2})^+. \end{aligned}$$

#### 5. Deux variantes plus précises

Nous choisissons tout d'abord  $\delta = 1$ ,  $\beta = 1$  et  $\forall i, j \in M$   $f_{i,j} > 0$ . On montre que la méthode reste continue et que toutes les propositions précédentes restent vraies.

Nous prenons maintenant  $\delta = 1$  et  $\beta = 0$ . On montre que la proposition 4.1 est encore vraie. Il n'est malheureusement pas possible d'établir que la méthode est toujours coercive. On peut juste écrire que, si la proposition 4.2 est vraie, alors les propositions 4.3 et 4.4 sont également vérifiées. Cependant, cette variante est intéressante, car elle s'avère d'ordre 2 dans nos applications numériques.

#### 6. Résultats numériques

Nous cherchons à retrouver numériquement la solution du problème suivant :

$$\begin{cases} -\operatorname{div}(\overline{\overline{D}}\nabla u) = f \operatorname{sur} \Omega = ]0, \ 0, 5[\times]0, \ 0, 5[\\ u = \sin(\pi x) \sin(\pi y) \operatorname{pour}(x, y) \in \partial \Omega \end{cases} \quad \text{avec } \overline{\overline{D}} = \frac{1}{(x^2 + y^2)} \begin{pmatrix} y^2 + \epsilon x^2 & -(1 - \epsilon)xy \\ -(1 - \epsilon)xy & x^2 + \epsilon y^2 \end{pmatrix}$$
(11)

et

$$u_{\text{ana}} = \sin(\pi x) \sin(\pi y),$$
  

$$f = -\text{div}\overline{\overline{D}}u_{\text{ana}}.$$
(12)

Le paramètre  $\epsilon$  est égal à 10<sup>-6</sup> ce qui donne un rapport d'anisotropie égal à 10<sup>6</sup>. Nous vérifions que  $f \ge 0$ . Par ailleurs, nous utilisons des maillages de carrés de surface  $h^2$ , h variant de  $\frac{1}{8}$  à  $\frac{1}{128}$ . Nous calculons la solution du problème à l'aide d'un algorithme de point fixe.

Nous montrons, tout d'abord, les résultats obtenus dans le Tableau 1 avec le schéma développé dans [15] (Schéma 1) (erreurs  $L^2$  de u par rapport à la solution analytique, ordre, pourcentage de valeurs négatives, valeur minimum). Nous présentons également les résultats obtenus avec la méthode décrite dans le paragraphe 3 (schéma 2 :  $\delta = 0$ ,  $\beta = 0$ ) et avec les variantes du paragraphe 5 (schéma 3 :  $\delta = 1$ ,  $\beta = 1$ ,  $\eta = 10^{-15}$  et schéma 4 :  $\delta = 1$ ,  $\beta = 0$ ,  $\eta = 10^{-15}$ ).

Nous observons que les schémas 2, 3 et 4 restent positifs. Nous notons que les algorithmes 2 et 3 sont d'un ordre compris entre 0, 5 et 1, alors que le schéma 4 s'approche de l'ordre 2. Nous remarquons que c'est le schéma dont nous n'avons pas établi une preuve complète de convergence qui est largement plus précis. En pratique, c'est ce dernier que nous utiliserons, et les méthodes 2 et 3 seront des garde-fous dans le cas où le schéma 4 n'est pas coercif.

#### Remerciements

L'auteur remercie Jérôme Droniou pour sa relecture détaillée, ses commentaires et ses suggestions.

#### Références

- [1] L. Agelas, D. Di Pietro, R. Masson, A symmetric and coercive finite volume scheme for multiphase porous media flow problems with applications in the oil industry, in: Finite Volumes for Complex Applications, vol. V, 2008, pp. 35–51.
- [2] C. Cancès, M. Cathala, C. Le Potier, Monotone corrections for generic cell-centered finite volume approximations of anisotropic diffusion equations, Numer. Math. 125 (2013) 387–417.
- [3] C. Cancès, C. Guichard, Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations, Math. Comput. 85 (298) (2016) 549–580.
- [4] B. Després, Non linear finite volume schemes for the heat equation in 1D, ESAIM: Math. Model. Numer. Anal. 48 (01) (2014) 107-134.
- [5] K. Domelevo, P. Omnes, A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids, ESAIM: Math. Model. Numer. Anal. 39 (6) (2005) 1203–1249.
- [6] J. Droniou, C. Le Potier, Construction and convergence study of local-maximum-principle preserving schemes for elliptic equations, SIAM J. Numer. Anal. 49 (2) (2011) 459–490.
- [7] J. Droniou, R. Eymard, T. Gallouët, R. Herbin, A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods, Math. Models Methods Appl. Sci. 20 (2) (2010) 265–295.
- [8] J. Droniou, Finite volume schemes for diffusion equations: introduction to and review of modern methods, Math. Models Methods Appl. Sci. 24 (8) (2014) 1575–1619 (special issue on Recent Techniques for PDE Discretizations on Polyhedral Meshes).
- [9] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, R. Herbin, The gradient discretisation method: a framework for the discretisation and numerical analysis of linear and nonlinear elliptic and parabolic problems, https://hal.archives-ouvertes.fr/hal-01382358, 2017.
- [10] R. Herbin, F. Hubert, Benchmark on discretization schemes for anisotropic diffusion problems on general grids, in: 5th International Symposium on Finite Volumes for Complex Applications, 8–13 June 2008, http://www.latp.univ-mrs.fr/fvca5.
- [11] I. Faille, Modélisation bidimensionnelle de la génèse et de la migration des hydrocarbures dans un bassin sédimentaire, Thèse de l'université Joseph-Fourier, Grenoble-1, Grenoble, France, 1992.
- [12] C. Le Potier, A. Mahamane, A nonlinear correction and maximum principle for diffusion operators discretized using hybrid schemes, C. R. Acad. Sci. Paris, Ser. I 350 (2012) 101–106.
- [13] C. Le Potier, A nonlinear second order in space correction and maximum principle for diffusion operators, C. R. Acad. Sci. Paris, Ser. I 352 (2014) 947–952.
- [14] C. Le Potier, Convergence of a nonlinear scheme for anisotropic diffusion equations, in: Finite Volumes for Complex Application VII, 2014, pp. 439-447.
- [15] K. Lipnikov, M. Shashkov, I. Yotov, Local flux mimetic finite difference methods, Numer. Math. 112 (2009) 115–152.
- [16] J.-M. Mirebeau, Minimal stencils for discretizations of anisotropic PDEs preserving causality or the maximum principle, SIAM J. Numer. Anal. 54 (3) (2016) 1582–1611.