



HAL
open science

Numerical methods for nonlinear elliptic and parabolic equations. Application to flow problems in porous and fractured media

Martin Vohralík

► **To cite this version:**

Martin Vohralík. Numerical methods for nonlinear elliptic and parabolic equations. Application to flow problems in porous and fractured media. Modeling and Simulation. Université Paris Sud - Paris XI, 2004. English. NNT: . tel-00008451

HAL Id: tel-00008451

<https://theses.hal.science/tel-00008451>

Submitted on 11 Feb 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° D'ORDRE : 7741

<p style="text-align: center;">UNIVERSITE PARIS XI UFR Scientifique d'Orsay UNIVERSITE TECHNIQUE TCHEQUE A PRAGUE Faculté de Génie Nucléaire et de Sciences de l'Ingénieur</p>
--

THESE

présentée pour obtenir le grade de

Docteur en sciences de l'Université Paris XI Orsay, spécialité mathématiques
&
Docteur de l'Université Technique Tchèque à Prague, spécialité modélisation mathématique

par

Martin VOHRALÍK

Sujet :

**METHODES NUMERIQUES POUR DES EQUATIONS
ELLIPTIQUES ET PARABOLIQUES NON LINEAIRES**

Application à des problèmes d'écoulement en milieux poreux et fracturés

Soutenue le 9 décembre 2004 devant la Commission d'examen :

M.	Robert	EYMARD	examineur
M.	Miloslav	FEISTAUER	examineur
Mme	Vivette	GIRAULT	examineur
M.	Bernard	HELFFER	président
Mme	Danielle	HILHORST	directrice de thèse
M.	Jiří	MARYŠKA	directeur de thèse
M.	Marc	BONNET	membre invité

Après avis des rapporteurs : M. Miloslav FEISTAUER
M. Jérôme JAFFRÉ

Remerciements

Je voudrais tout d'abord remercier mes directeurs de thèse, Madame Danielle Hilhorst et Monsieur Jiří Maryška. Je suis très reconnaissant à Madame Hilhorst d'avoir été une directrice de thèse très responsable. Son expérience, sa disponibilité et sa confiance m'ont permis de mener à bien ce travail. Je dois aussi beaucoup à Madame Hilhorst d'un point de vue personnel pour m'avoir accueilli très chaleureusement en France. Je voudrais également remercier profondément Monsieur Maryška, pour avoir dirigé mes premiers pas scientifiques vers les méthodes numériques en milieu poreux et pour m'avoir intégré dans la collaboration franco-tchèque, qui a mené à cette thèse.

Mes remerciements sincères vont ensuite à Monsieur Robert Eymard, pour m'avoir ouvert les yeux à la beauté des principes de la méthode des volumes finis, pour m'avoir fait partager ses connaissances inestimables et pour avoir accepté d'être membre du jury. Je voudrais aussi exprimer ma reconnaissance à Monsieur Marc Bonnet de la société HydroExpert pour m'avoir donné la possibilité d'appliquer à la pratique hydrogéologique les méthodes numériques étudiées dans cette thèse d'un point de vue théorique. Je tiens finalement à remercier Monsieur Michal Beneš de ma faculté d'origine pour son soutien constant et Monsieur Christian Grossmann du Département d'Analyse Numérique de l'Université Technique à Dresde pour mon séjour dans son département.

Je suis très sensible à l'honneur que m'ont fait Monsieur Miloslav Feistauer et Monsieur Jérôme Jaffré en acceptant d'être les rapporteurs de cette thèse. J'apprécie sincèrement leur intérêt pour mes travaux ainsi que leurs commentaires avisés. Je voudrais aussi exprimer ma gratitude profonde à Monsieur Bernard Helffer pour avoir accepté d'être le président du jury et à Madame Vivette Girault et Monsieur Miloslav Feistauer pour avoir accepté d'être membres du jury.

J'apprécie énormément le soutien de la République Française, exprimée par son Ambassade en République Tchèque en m'attribuant une Bourse du Gouvernement Français, qui m'a permis de poursuivre une partie de mes études doctorales en France et de découvrir ainsi ce magnifique pays tant sur le plan scientifique que culturel. J'ai aussi bénéficié d'un soutien financier partiel de la Fondation de Marie, Zdeňka et Josef Hlávka à Prague, ce dont je suis très reconnaissant.

Je tiens à remercier toute l'équipe d'Analyse Numérique et Equations aux Dérivées Partielles de l'Université de Paris-Sud pour m'avoir accueilli très chaleureusement.

Enfin, je voudrais exprimer mes remerciements profonds à mes parents et à toute ma famille pour leur soutien permanent, leurs encouragements et leur amour. Et tout cela ne serait pas comme ça sans Martina ...

A mes parents

A Martina

METHODES NUMERIQUES POUR DES EQUATIONS ELLIPTIQUES ET PARABOLIQUES NON LINEAIRES

Application à des problèmes d'écoulement en milieux poreux et fracturés

Résumé

Les travaux de cette thèse portent sur des méthodes numériques pour la discrétisation d'équations aux dérivées partielles elliptiques et paraboliques de convection–réaction–diffusion non linéaires. Nous analysons ces méthodes et nous les appliquons à la simulation effective de l'écoulement et du transport de contaminants en milieux poreux et fracturés.

Au chapitre 1, nous proposons un schéma permettant une discrétisation efficace, robuste, conservative et stable des équations de convection–réaction–diffusion non linéaires paraboliques dégénérées sur des maillages non structurés en dimensions deux ou trois d'espace. Nous discrétisons le terme de diffusion, qui contient en général un tenseur de diffusion inhomogène et anisotrope, par la méthode des éléments finis non conformes ou mixtes-hybrides et les autres termes par la méthode des volumes finis. La partie essentielle du chapitre est ensuite consacrée à montrer l'existence et l'unicité d'une solution discrète et sa convergence vers une solution faible du problème continu. La méthode de démonstration permet en particulier d'éviter des hypothèses restrictives sur le maillage souvent présentes dans la littérature. Nous proposons finalement une variante de ce schéma pour des maillages qui ne se raccordent pas, couplant cette fois la méthode des volumes finis avec celle des éléments finis conformes, et nous l'appliquons à la simulation du transport de contaminants en milieux poreux.

Au chapitre 2, nous présentons une démonstration constructive des inégalités de Poincaré–Friedrichs discrètes pour une classe d'approximations non conformes de l'espace de Sobolev H^1 , indiquons les valeurs optimales des constantes dans ces inégalités et montrons l'inégalité de Friedrichs discrète pour des domaines bornés dans une direction uniquement. Ces résultats sont importants dans l'analyse de méthodes numériques non conformes, comme les méthodes d'éléments finis non conformes ou de Galerkin discontinu.

Au chapitre 3, nous montrons que la méthode des éléments finis mixtes de Raviart–Thomas de plus bas degré pour des problèmes elliptiques en dimension deux ou trois d'espace est équivalente à un schéma de volumes finis à plusieurs points. Après avoir étudié ce schéma, nous l'appliquons à la discrétisation d'équations de convection–réaction–diffusion paraboliques non linéaires. Cette approche permet de réduire le temps de calcul de la méthode des éléments finis mixtes, tout en conservant sa très grande précision, ce qui est confirmé par les tests numériques.

Enfin, au chapitre 4, nous proposons une version de la méthode des éléments finis mixtes de Raviart–Thomas de plus bas degré pour la résolution de problèmes elliptiques sur un système de polygones bidimensionnels placés dans l'espace tridimensionnel, démontrons qu'elle est bien posée et étudions sa relation avec la méthode des éléments finis non conformes. Ces résultats sont finalement appliqués à la simulation de l'écoulement de l'eau souterraine dans un système de polygones représentant un réseau de fractures perturbant un massif rocheux.

Mots clés : Méthodes volumes finis – Méthodes éléments finis – Méthodes éléments finis mixtes – Maillages non structurés – Inégalités de Poincaré–Friedrichs discrètes – Existence et unicité – Convergence – Simulations numériques – Equation de convection–réaction–diffusion parabolique dégénérée – Ecoulement et transport de contaminants en milieux poreux et fracturés

AMS subject classifications : 65M12, 65N30, 76M10, 76M12, 76S05, 35J20, 35K65, 46E35

NUMERICAL METHODS FOR NONLINEAR ELLIPTIC AND PARABOLIC EQUATIONS

Application to flow problems in porous and fractured media

Abstract

This thesis deals with numerical methods for the discretization of nonlinear elliptic and parabolic convection–reaction–diffusion partial differential equations. We analyze these methods and apply them to the effective simulation of flow and contaminant transport in porous and fractured media.

In Chapter 1 we propose a scheme allowing for efficient, robust, conservative, and stable discretizations of nonlinear degenerate parabolic convection–reaction–diffusion equations on unstructured grids in two or three space dimensions. We discretize the diffusion term, which generally involves an inhomogeneous and anisotropic diffusion tensor, by means of the nonconforming or mixed-hybrid finite element method and the other terms by means of the finite volume method. The essential part of this chapter is then devoted to showing the existence and uniqueness of a discrete solution and its convergence to a weak solution of the continuous problem. The proofs permit in particular to avoid restrictive hypotheses on the mesh often used in the literature. We finally propose a version of this scheme for nonmatching grids, combining this time the finite volume method with the piecewise linear conforming finite element method. We then apply this version to contaminant transport simulations in porous media.

In Chapter 2 we present a direct proof of the discrete Poincaré–Friedrichs inequalities for a class of nonconforming approximations of the Sobolev space H^1 , indicate optimal values of the constants in these inequalities, and extend the discrete Friedrichs inequality onto domains only bounded in one direction. The results are important in the analysis of nonconforming numerical methods, such as nonconforming finite element or discontinuous Galerkin methods.

In Chapter 3 we show that the lowest-order Raviart–Thomas mixed finite element method for elliptic problems on simplicial meshes in two or three space dimensions is equivalent to a particular multi-point finite volume scheme. We study this scheme and apply it to the discretization of nonlinear parabolic convection–reaction–diffusion equations. This approach allows significant reduction of the computational time of the mixed finite element method without any loss of its high precision, which is confirmed by numerical experiments.

Finally, in Chapter 4 we propose a version of the lowest-order Raviart–Thomas mixed finite element method for the approximation of elliptic problems on a system of two-dimensional polygons placed in three-dimensional space, prove that it is well-posed, and study its relation to the nonconforming finite element method. These results are finally applied to the simulation of underground water flow through a system of polygons representing a network of fractures that perturbs a rock massif.

Key words: Finite volume method – Finite element method – Mixed finite element method – Unstructured grids – Discrete Poincaré–Friedrichs inequalities – Existence and uniqueness – Convergence – Numerical simulations – Degenerate parabolic convection–reaction–diffusion equation – Flow and contaminant transport in porous and fractured media

AMS subject classifications: 65M12, 65N30, 76M10, 76M12, 76S05, 35J20, 35K65, 46E35

Table des matières

Introduction	11
1 Combined finite volume–finite element schemes for degenerate parabolic convection–reaction–diffusion problems	19
1.1 Introduction	20
1.2 The nonlinear degenerate parabolic problem	22
1.3 Combined finite volume–nonconforming/mixed-hybrid finite element scheme	24
1.3.1 Space and time discretizations	24
1.3.2 The combined scheme	26
1.4 Existence, uniqueness, and discrete properties	28
1.4.1 Discrete properties of the scheme	29
1.4.2 Existence, uniqueness, and the discrete maximum principle	33
1.5 A priori estimates	35
1.5.1 A priori estimates	36
1.5.2 Estimates on differences of time and space translates	38
1.6 Convergence	44
1.6.1 Strong convergence in $L^2(Q_T)$	45
1.6.2 Convergence to a weak solution	45
1.7 Numerical experiment	57
1.8 Appendix A: Technical lemmas	59
1.9 Appendix B: A combined finite volume–finite element scheme for contaminant transport simulation on nonmatching grids	63
1.9.1 Introduction	63
1.9.2 The contaminant transport problem	64
1.9.3 Combined finite volume–finite element scheme	65
1.9.4 Discrete properties of the scheme	68
1.9.5 Numerical simulations	71
1.9.6 Concluding remarks	76
2 Discrete Poincaré–Friedrichs inequalities	79
2.1 Introduction	80
2.2 Notation and assumptions	81
2.3 Discrete Friedrichs inequality for piecewise constant functions	84
2.4 Interpolation estimates on functions from $H^1(K)$	88
2.5 Discrete Friedrichs inequality	92
2.6 Discrete Friedrichs inequality for Crouzeix–Raviart finite elements	94
2.7 Discrete Poincaré inequality for piecewise constant functions	96
2.8 Discrete Poincaré inequality	98

3	Equivalence between lowest-order mixed finite element and multi-point finite volume methods	101
3.1	Introduction	102
3.2	The equivalence	103
3.3	Properties of the condensed mixed finite element scheme	107
3.3.1	Properties of the system matrix	107
3.3.2	Properties of the local condensation matrices	108
3.3.3	Variants, extensions, and open problems	112
3.4	Application to nonlinear parabolic problems	114
3.5	Numerical experiments	116
3.5.1	Condensed mixed finite element method for elliptic problems	118
3.5.2	Condensed mixed finite element method for nonlinear parabolic problems	118
3.5.3	Comparison of the condensed mixed finite element, finite volume, and combined finite volume–finite element methods	123
3.5.4	Conclusions	128
4	Mixed and nonconforming finite element methods on a fracture network	131
4.1	Introduction	132
4.2	Second-order elliptic problem on a system of polygons	133
4.3	Function spaces for nonconforming and mixed finite elements	134
4.3.1	Continuous function spaces	134
4.3.2	Discrete function spaces	135
4.4	Nonconforming finite element method	136
4.4.1	Weak primal solution	136
4.4.2	Nonconforming finite element approximation	136
4.5	Raviart–Thomas mixed finite element method	137
4.5.1	Weak mixed solution	137
4.5.2	Properties of the discrete velocity space	138
4.5.3	Mixed finite element approximation	141
4.5.4	Hybridization of the mixed approximation	141
4.5.5	Error estimates	142
4.6	Relation between mixed and nonconforming methods	142
4.6.1	Algebraic condensation of the mixed-hybrid approximation	142
4.6.2	Comparison of condensed mixed-hybrid and nonconforming methods . .	144
4.7	Numerical simulations	146
4.7.1	Model problem with a known analytical solution	146
4.7.2	Real problem	148
	Bibliographie	151

Introduction

Les équations aux dérivées partielles décrivent un grand nombre de phénomènes dans notre environnement. Dans la plupart des cas, il n'est pas possible d'en trouver des solutions analytiques. C'est dans le but de pouvoir calculer des solutions approchées que les méthodes numériques ont été développées. Aujourd'hui, l'impact de ces méthodes est d'autant plus essentiel que la puissance des ordinateurs ne cesse d'augmenter.

La méthode des éléments finis a été développée par des ingénieurs vers 1950 puis étudiée par des mathématiciens. On peut citer en particulier les ouvrages de Strang et Fix, Zienkiewicz et Ciarlet. La méthode des volumes finis, également développée par des ingénieurs, a été analysée d'un point de vue mathématique beaucoup plus tard, en particulier par Eymard, Gallouët et Herbin. Les idées essentielles de ces deux méthodes sont apparemment très différentes puisque la méthode des éléments finis est basée sur la minimisation d'une énergie tandis que, dans celle des volumes finis, on approche les flux à l'aide d'une formule intégrale. Pourtant, la discrétisation d'une équation elliptique du deuxième ordre à l'aide de ces méthodes peut conduire à des problèmes discrets très proches et parfois identiques. En particulier, contrairement à ce qui a été longtemps affirmé, on peut montrer, en réinterprétant les résultats, que la méthode des éléments finis conserve localement la masse tout comme la méthode des volumes finis. Tandis qu'elles sont très voisines pour la discrétisation d'un terme diffusif du deuxième ordre, ces méthodes ont un comportement différent pour celle d'un terme convectif du premier ordre, d'un terme réactif ou pour la discrétisation de la dérivée en temps. C'est ce qui motive l'introduction de schémas combinant les méthodes de volumes finis et d'éléments finis.

La méthode des éléments finis mixtes de Raviart et Thomas permet des calculs plus précis. Elle est basée sur l'approximation simultanée de la fonction scalaire inconnue et de son flux. Elle conduit néanmoins à des systèmes linéaires de type point-selle, si bien que la résolution numérique en est très onéreuse. C'est à la fois pour diminuer le temps de calcul et pour en rendre l'implémentation plus aisée que ses relations avec la méthode des éléments finis non conformes d'une part et avec les méthodes de volumes finis et de différences finies d'autre part ont été étudiées. Il y a eu de plus des retombées supplémentaires sous la forme de résultats nouveaux dans l'analyse mathématique de la méthode des éléments finis mixtes.

Dans cette thèse, nous étudions tout d'abord des schémas combinant des méthodes de volumes finis et d'éléments finis pour la discrétisation de problèmes de convection–réaction–diffusion paraboliques non linéaires avec un tenseur de diffusion inhomogène et anisotrope sur des maillages très généraux. Nous montrons ensuite que la méthode des éléments finis mixtes de Raviart–Thomas de plus bas degré est équivalente à une méthode de type volumes finis. Nous appliquons finalement ces méthodes à la résolution de problèmes intervenant en environnement, en particulier à la simulation de l'écoulement et du transport de contaminants en milieux poreux et fracturés. Nous présentons aussi des constantes optimales pour des inégalités de Poincaré–Friedrichs discrètes, nécessaires à l'analyse de ces schémas. Les quatre chapitres de cette thèse et l'appendice [1.9](#) peuvent être lus indépendamment.

Chapitre 1 : Schémas combinant les méthodes de volumes finis et d'éléments finis pour des problèmes de convection–réaction–diffusion paraboliques dégénérés

Nous proposons et étudions dans ce chapitre deux schémas qui combinent des méthodes de volumes finis et d'éléments finis et qui permettent une discrétisation efficace des équations paraboliques dégénérées sur des maillages très généraux.

Nous considérons l'équation de convection–réaction–diffusion

$$\frac{\partial \beta(c)}{\partial t} - \nabla \cdot (\mathbf{S} \nabla c) + \nabla \cdot (c \mathbf{v}) + F(c) = q, \quad (1)$$

qui décrit le transport réactif avec adsorption équilibre en milieu poreux, cf. [19, 81]. Ici, $c = c(\mathbf{x}, t)$ est la concentration d'un contaminant, $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$ est un champ de vitesses externe, $\mathbf{S} = \mathbf{S}(\mathbf{x}, \mathbf{v}, t)$ est le tenseur de dispersion–diffusion, la fonction β décrit l'évolution en temps et l'adsorption équilibre, la fonction F est liée aux réactions chimiques et $q = q(\mathbf{x}, t)$ est un terme source. On obtient un problème complet après avoir ajouté une condition aux limites et une condition initiale appropriée. L'approximation numérique de ce problème est compliquée parce que la dérivée de la fonction β n'est pas bornée, c'est-à-dire que l'équation (1) est parabolique dégénérée, que le terme de convection domine celui de diffusion et que le tenseur \mathbf{S} est inhomogène et anisotrope (il n'est pas constant et a la forme d'une matrice pleine).

La méthode des éléments finis pour des équations paraboliques dégénérées a été étudiée par exemple dans [20, 94, 109]. Elle permet une approximation très aisée des termes de diffusion dans le cas de tenseurs de diffusion anisotropes et ne nécessite aucune condition sur les maillages. Rappelons qu'elle est localement conservative, contrairement à ce qui a été longtemps affirmé, cf. [68], [61, Section III.12] ou [75]. En revanche, on sait qu'elle produit des oscillations numériques non justifiées quand le problème est dominé par la convection. La méthode des volumes finis (centrée par maille) pour des équations paraboliques dégénérées a été étudiée dans [62, 63]. Elle est très efficace, peu coûteuse et permet d'éviter les oscillations numériques quand le problème est dominé par la convection, si l'on discrétise le terme de convection à l'aide d'un schéma amont. En revanche, il y a des restrictions sur le maillage pour la discrétisation du terme de diffusion et on ne sait pas l'étendre simplement au cas où le tenseur de diffusion est anisotrope. Pour éviter les inconvénients de ces deux méthodes classiques, des schémas combinant des méthodes de volumes finis et d'éléments finis ont été proposés, cf. [10, 48, 67, 90]. Cependant, ces méthodes n'ont été proposées que pour des équations uniformément paraboliques sans terme de réaction et n'ont été étudiées qu'avec des hypothèses assez restrictives sur le maillage.

Un schéma combinant les méthodes des volumes finis et des éléments finis non conformes ou mixtes-hybrides

Nous supposons dans cette partie que le maillage est non structuré et composé de triangles en dimension deux d'espace et de tétraèdres en dimension trois d'espace. Pour discrétiser l'équation (1), nous nous sommes inspirés de la méthode proposée dans [10] pour des équations de la mécanique des fluides.

Nous discrétisons le terme de diffusion à l'aide de la méthode des éléments finis non conformes sur le maillage donné et les autres termes à l'aide de la méthode des volumes finis sur un maillage de volumes de contrôle, construits autour des faces du maillage initial. Nous proposons également de remplacer les éléments finis non conformes par des éléments finis

mixtes-hybrides, où les seules inconnues sont les multiplicateurs de Lagrange, cf. [15]. Pour discrétiser le terme de convection, nous proposons un flux numérique tenant compte du nombre de Péclet local, de façon à ajuster la quantité de décentrage amont selon la proportion locale de la convection et de la diffusion. Ainsi, on n'ajoute au schéma que le minimum de diffusion numérique nécessaire à assurer sa stabilité. Pour la discrétisation temporelle, nous utilisons une méthode de différences finies complètement implicite.

Nous montrons d'abord qu'il existe une solution discrète unique. Nous prouvons ensuite que cette solution vérifie le principe de maximum discret s'il n'y a pas d'angles obtus dans le maillage et si le tenseur de diffusion est scalaire et que le schéma conserve localement la masse. Nous démontrons ensuite des estimations a priori et des estimations sur les différences de translatées en espace et en temps, ce qui implique une propriété de compacité relative par le théorème de Fréchet–Kolmogorov. On démontre ainsi la convergence forte dans L^2 d'une sous-suite de solutions approchées vers une solution faible du problème continu. Pour les démonstrations, nous utilisons des méthodes présentées dans [61] que nous étendons au cas de transmissibilités déduites du schéma d'éléments finis et qui peuvent être négatives (auquel cas le principe de maximum discret n'est pas valide) et au cas de maillages de triangles ou de tétraèdres généraux. Cette méthode de démonstration donne aussi la possibilité de généraliser les hypothèses qui ont été nécessaires dans [10]. En particulier, nous n'imposons aucune condition sur les angles maximaux du maillage et nous permettons également son raffinement local. Nous résolvons finalement les systèmes non linéaires pour les inconnues discrètes correspondant à la fonction $\beta(c)$. On peut ainsi éviter toute régularisation parabolique (cf. [20]) ou perturbation des conditions initiales et des conditions aux limites (cf. [99]), rendant l'équation uniformément parabolique. De plus, les systèmes linéaires ainsi obtenus sont diagonaux pour la partie des inconnues correspondant aux zones où $c = 0$. Le schéma proposé permet une discrétisation efficace, robuste, conservative et stable de l'équation (1), ce qui est confirmé par les essais numériques présentés à la fin du chapitre.

Cette partie fait l'objet d'un article écrit en collaboration avec R. Eymard et D. Hilhorst, soumis pour publication dans *Numerische Mathematik*. Une version abrégée de cet article a été publiée dans des actes (avec comité de lecture) du congrès ENUMATH 2003.

Un schéma combinant les méthodes des volumes finis et des éléments finis pour la simulation du transport de contaminants sur des maillages qui ne se raccordent pas

Nous considérons dans cette partie l'équation (1) sous la forme où elle intervient en hydrologie, cf. [25, 123]. Nous supposons que le maillage est non structuré, composé de volumes de contrôle polygonaux qui ne sont pas nécessairement convexes et qui ne se raccordent pas. Nous étendons à ce type de maillages les schémas proposés dans [67, 114].

Nous construisons un maillage qui se raccorde, composé de triangles en dimension deux d'espace et de tétraèdres en dimension trois d'espace, et dont les sommets sont associés aux volumes de contrôle du maillage donné. Nous reprenons ensuite les idées du schéma précédent, en nous appuyant cette fois sur la méthode des éléments finis conformes. Nous généralisons le flux numérique tenant compte du nombre de Péclet local, montrons que le schéma reste localement conservatif et qu'il vérifie le principe de maximum discret sous certaines conditions sur le maillage et sur le tenseur de diffusion. On pourrait démontrer sa convergence comme on l'avait fait pour le premier schéma. Ce schéma est plus simple que la plupart des autres schémas présentés dans la littérature pour des maillages qui ne se raccordent pas, tout en étant très efficace. En particulier, on évite d'introduire des équations ou des inconnues supplémentaires

ou d'interpoler sur les frontières entre les mailles qui ne se raccordent pas, cf. [3, 11, 26, 55, 66]. En effet, le schéma proposé pourrait être considéré comme une version consistante du schéma proposé dans [36]. Nous présentons finalement des résultats de simulation pour un problème modèle ainsi que pour un problème réel fourni par la société HydroExpert de Paris. Le schéma proposé est implémenté dans le logiciel TALISMAN [104] de cette société, qui utilise des maillages rectangulaires raffinés localement, et donc ne se raccordant pas.

Cette partie fait l'objet d'un article écrit en collaboration avec R. Eymard et D. Hilhorst, qui sera prochainement soumis pour publication dans *Transport in Porous Media*.

Chapitre 2 : Inégalités de Poincaré–Friedrichs discrètes

Nous étudions dans ce chapitre des versions discrètes des inégalités de Poincaré–Friedrichs. Ces inégalités sont importantes dans l'analyse des méthodes numériques non conformes.

L'inégalité de Friedrichs

$$\int_{\Omega} g^2(\mathbf{x}) \, d\mathbf{x} \leq c_F \int_{\Omega} |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \quad \forall g \in H_0^1(\Omega) \quad (2)$$

et l'inégalité de Poincaré

$$\int_{\Omega} g^2(\mathbf{x}) \, d\mathbf{x} \leq c_P \int_{\Omega} |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} + \tilde{c}_P \left(\int_{\Omega} g(\mathbf{x}) \, d\mathbf{x} \right)^2 \quad \forall g \in H^1(\Omega) \quad (3)$$

(cf. [91]) sont essentielles dans la théorie des équations aux dérivées partielles. On suppose ici que $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, est un ouvert borné connexe de frontière polygonale.

Soit $\{\mathcal{T}_h\}_h$ une famille de triangulations de Ω , composée de triangles en dimension deux et de tétraèdres en dimension trois. Soit $W(\mathcal{T}_h)$ l'ensemble des fonctions de $H^1(K)$ pour tout $K \in \mathcal{T}_h$, telles que les moyennes de leurs traces sur les faces intérieures coïncident. Finalement, soit $W_0(\mathcal{T}_h) \subset W(\mathcal{T}_h)$ l'ensemble des fonctions telles que les moyennes des traces sur les faces extérieures sont nulles. Ces espaces sont des approximations non conformes des espaces continus dans le sens où $W_0(\mathcal{T}_h) \not\subset H_0^1(\Omega)$ et $W(\mathcal{T}_h) \not\subset H^1(\Omega)$. Nous étudions des versions discrètes des inégalités (2) et (3) données par

$$\int_{\Omega} g^2(\mathbf{x}) \, d\mathbf{x} \leq C_F \sum_{K \in \mathcal{T}_h} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \quad \forall g \in W_0(\mathcal{T}_h), \forall h > 0, \quad (4)$$

$$\int_{\Omega} g^2(\mathbf{x}) \, d\mathbf{x} \leq C_P \sum_{K \in \mathcal{T}_h} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} + \tilde{C}_P \left(\int_{\Omega} g(\mathbf{x}) \, d\mathbf{x} \right)^2 \quad \forall g \in W(\mathcal{T}_h), \forall h > 0. \quad (5)$$

Les inégalités (4) et (5) ont été étudiées dans [28, 51, 84, 116]. Il a été démontré dans [28, 84] que les constantes C_F et C_P dépendent seulement de Ω et de la régularité des maillages. Nous établissons les dépendances de C_F et C_P en ces paramètres et nous présentons un exemple montrant que ces dépendances sont optimales. Enfin, la dépendance de C_F de Ω nous permet d'étendre l'inégalité de Friedrichs discrète au cas de domaines bornés dans une direction uniquement. Notre démonstration de (4) et de (5) est constructive et porte sur l'extension de la démonstration des inégalités de Poincaré–Friedrichs discrètes connues pour des fonctions constantes par morceaux dans le cadre de la méthode des volumes finis, cf. [61]. Ces résultats sont importants dans l'analyse de méthodes numériques non conformes, comme les méthodes d'éléments finis non conformes ou de Galerkin discontinu.

Ce chapitre fait l'objet d'un article écrit seul, soumis pour publication dans *Numerical Functional Analysis and Optimization*.

Chapitre 3 : Equivalence entre les méthodes des éléments finis mixtes de plus bas degré et des volumes finis à plusieurs points

Nous démontrons dans ce chapitre l'équivalence entre la méthode des éléments finis mixtes et un schéma de volumes finis à plusieurs points pour des équations elliptiques et l'appliquons à la discrétisation d'équations paraboliques non linéaires. Le but est de réduire le temps de calcul de la méthode des éléments finis mixtes, tout en conservant sa très grande précision.

Nous considérons d'abord le problème elliptique

$$\mathbf{u} = -\mathbf{S}\nabla p \quad \text{dans } \Omega, \quad (6a)$$

$$\nabla \cdot \mathbf{u} = q \quad \text{dans } \Omega, \quad (6b)$$

$$p = p_D \quad \text{sur } \Gamma_D, \quad \mathbf{u} \cdot \mathbf{n} = u_N \quad \text{sur } \Gamma_N, \quad (6c)$$

où $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, est un ouvert borné connexe de frontière polygonale, \mathbf{S} est un tenseur symétrique borné et uniformément défini positif, $p_D \in H^{\frac{1}{2}}(\partial\Omega)$, $u_N \in H^{-\frac{1}{2}}(\Gamma_N)$ et $q \in L_2(\Omega)$. Soit \mathcal{T}_h un maillage de Ω composé de triangles en dimension deux d'espace et de tétraèdres en dimension trois d'espace. Dans la méthode des éléments finis mixtes de Raviart–Thomas [105], (cf. Nédélec [92] en dimension trois d'espace) de plus bas degré pour le problème (6a)–(6c), on cherche simultanément les inconnues scalaires P associées aux éléments du maillage et les flux U à travers les faces du maillage. Le système linéaire associé est de type point-selle et peut être écrit sous la forme

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^t \\ \mathbb{B} & 0 \end{pmatrix} \begin{pmatrix} U \\ P \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}. \quad (7)$$

En suivant les idées données dans [69] et en introduisant les multiplicateurs de Lagrange associés aux faces, on peut réduire le nombre d'inconnues au nombre des faces et obtenir des matrices symétriques définies positives, cf. [15]. Si \mathbf{S} est diagonal, on peut éliminer les flux U et obtenir un problème pour les seules inconnues scalaires P en utilisant l'intégration numérique approchée, cf. [8, 18, 110]. Ces idées peuvent être étendues au cas d'un tenseur \mathbf{S} général, cf. [12]. En dimension deux d'espace, la méthode de Raviart–Thomas de plus bas degré peut être reformulée à l'aide d'une nouvelle inconnue associée aux éléments du maillage, cf. [37, 121, 122]. A notre connaissance, ceci est la seule approche exacte connue pour réduire le nombre d'inconnues au nombre d'éléments.

Nous mettons au point une nouvelle méthode pour réduire le système (7) à un système pour les seules inconnues scalaires P . Nous montrons que, sous la condition que certaines matrices locales associées aux sommets soient inversibles, il est possible d'exprimer le flux à travers une face à l'aide des inconnues scalaires, des sources et des conditions sur le bord associées aux éléments qui ont un sommet commun avec cette face. Rappelons qu'exprimer le flux à travers une face à l'aide des inconnues scalaires associées aux éléments voisins de cette face est le principe des méthodes des volumes finis à plusieurs points, cf. [1, 44, 65]. La méthode de Raviart–Thomas de plus bas degré est donc dans ce cas équivalente à une méthode de type volumes finis à plusieurs points, et ceci sans aucune intégration numérique. Nous discutons ensuite les modifications à faire quand les matrices locales ne sont pas inversibles. L'élimination proposée conduit à des systèmes linéaires avec une matrice creuse mais en général non symétrique. Nous prouvons que cette matrice est définie positive sous une condition sur le maillage et sur le tenseur \mathbf{S} , permettant des éléments assez déformés si \mathbf{S} est constant par morceaux et scalaire. Cette condition implique en particulier que les matrices locales mentionnées ci-dessus sont inversibles. Enfin, l'élimination proposée s'applique de la même manière pour la discrétisation des équations de convection–réaction–diffusion paraboliques

non linéaires par les méthodes des éléments finis mixtes et des éléments finis mixtes avec un décentrage amont (cf. [14] et [46, 47, 77]).

L'idée essentielle est la suivante : étant donné un problème du deuxième ordre, faire d'abord une décomposition en des inconnues scalaires et vectorielles qui garantit que la condition inf-sup est satisfaite, puis éliminer les inconnues vectorielles ajoutées. Ceci permet d'obtenir la précision de la méthode des éléments finis mixtes au prix de la méthode des volumes finis, ce qui est confirmé par les essais numériques. En particulier, pour des problèmes paraboliques non linéaires, on peut réduire d'un facteur de 2 à 4 le temps de calcul de la méthode des éléments finis mixtes. Cette approche peut être très facilement implémentée dans des codes existants. Enfin, il s'avère que les idées présentées s'appliquent aussi à des schémas d'ordre élevé.

Ce chapitre fait l'objet d'un article écrit seul, qui sera prochainement soumis pour publication dans *M2AN. Mathematical Modelling and Numerical Analysis*. Une version abrégée de cet article a été publiée dans *Comptes Rendus de l'Académie des Sciences, Ser. I*.

Chapitre 4 : Les méthodes des éléments finis mixtes et non conformes sur un réseau de fractures

Dans ce chapitre, nous proposons la méthode des éléments finis mixtes pour des problèmes elliptiques sur un réseau de polygones, démontrons qu'elle est bien posée et étudions sa relation avec la méthode des éléments finis non conformes. Nous l'appliquons finalement à la simulation de l'écoulement dans un réseau de fractures perturbant un massif rocheux.

Nous supposons donné un système

$$\mathcal{S} := \bigcup_{\ell \in L} \alpha_\ell,$$

où α_ℓ est un polygone bidimensionnel placé dans l'espace tridimensionnel, connecté à travers ses arêtes avec d'autres polygones du réseau. La propriété importante est qu'il peut y avoir des arêtes partagées par un nombre de polygones supérieur ou égal à trois, ce qui n'est pas possible dans une décomposition polygonale d'un domaine bidimensionnel. Nous considérons le problème elliptique posé sur \mathcal{S} de trouver p et \mathbf{u} tels que

$$\mathbf{u} = -\mathbf{K}(\nabla p + \nabla z) \quad \text{dans } \alpha_\ell, \ell \in L, \quad (8a)$$

$$\nabla \cdot \mathbf{u} = q \quad \text{dans } \alpha_\ell, \ell \in L, \quad (8b)$$

$$p = p_D \quad \text{sur } \Gamma_D, \quad \mathbf{u} \cdot \mathbf{n} = u_N \quad \text{sur } \Gamma_N, \quad (8c)$$

où toutes les variables ainsi que la différentiation sont exprimées dans des coordonnées locales bidimensionnelles de α_ℓ . Soit f l'arête partagée par des polygones dont l'ensemble d'indices est I_f . On ferme le système (8a)–(8c) en imposant

$$\begin{aligned} p|_{\overline{\alpha_i}} &= p|_{\overline{\alpha_j}} \quad \text{sur } f \quad \forall i, j \in I_f, \\ \sum_{i \in I_f} \mathbf{u}|_{\overline{\alpha_i}} \cdot \mathbf{n}_{f, \alpha_i} &= 0 \quad \text{sur } f \end{aligned}$$

pour toute arête f intérieure, où \mathbf{n}_{f, α_i} est le vecteur normal unitaire extérieur à l'arête f pour le polygone α_i . Cette relation exprime la continuité de p et une continuité appropriée du flux normal de \mathbf{u} à travers f . Ce problème elliptique décrit l'écoulement de l'eau souterraine dans un réseau de fractures perturbant un massif rocheux, cf. [5].

Nous proposons la méthode des éléments finis mixtes pour le problème considéré. Après avoir défini des espaces fonctionnels qui assurent les continuités appropriées sur les arêtes intérieures, nous démontrons l'existence et l'unicité d'une solution mixte faible sur un réseau de polygones. Nous considérons ensuite une discrétisation de ce réseau à l'aide de triangles et définissons les espaces discrets, en nous appuyant sur les éléments finis mixtes de plus bas degré de Raviart–Thomas [105]. Les fonctions de base vectorielles doivent assurer le bilan de la masse pour les arêtes intérieures. Nous montrons ensuite que cette méthode est bien posée, c'est-à-dire qu'il existe une solution discrète unique. Les estimations d'erreur sont une conséquence de la théorie générale, cf. [33, 108]. Nous étudions finalement la relation de l'hybridisation de la méthode des éléments finis mixtes avec la méthode des éléments finis non conformes. Nous étendons les résultats connus dans cette direction (cf. [15, 38]) pour des tenseurs de conductivité hydraulique \mathbf{K} non constants, pour des conditions aux limites de Neumann et de Dirichlet inhomogènes et pour des réseaux de polygones. Cette relation permet en particulier une implémentation efficace de la méthode des éléments finis mixtes de plus bas degré de Raviart–Thomas sur des réseaux de polygones. Nous présentons un test numérique pour lequel la solution analytique est connue, confirmant ainsi les résultats théoriques. Nous montrons finalement l'application de la méthode développée à la simulation de l'écoulement de l'eau souterraine dans un massif rocheux à l'Ouest de la Bohême, qui devrait devenir un site de stockage de déchets radioactifs dangereux.

Ce chapitre fait l'objet d'un article écrit en collaboration avec J. Maryška et O. Severýn, soumis pour publication dans *Applied Numerical Mathematics*. Une version abrégée de cet article a été publiée dans *Contemporary Mathematics* (Current Trends in Scientific Computing). L'application de ces résultats à des simulations dans des cas réels fait l'objet d'un article écrit avec les mêmes auteurs, publié dans *Computational Geosciences*.

Chapitre 1

Combined finite volume–finite element schemes for degenerate parabolic convection–reaction–diffusion problems

We propose and analyze in this chapter a numerical scheme for nonlinear degenerate parabolic convection–reaction–diffusion equations in two or three space dimensions. We discretize the diffusion term, which generally involves an inhomogeneous and anisotropic diffusion tensor, over an unstructured simplicial mesh of the space domain by means of the piecewise linear nonconforming (Crouzeix–Raviart) finite element method, or using the stiffness matrix of the hybridization of the lowest-order Raviart–Thomas mixed finite element method. The other terms are discretized by means of a cell-centered finite volume scheme on a dual mesh, where the dual volumes are constructed around the sides of the original mesh. Checking the local Péclet number, we set up the exact necessary amount of upstream weighting to avoid spurious oscillations in the convection-dominated case. This technique also ensures the validity of the discrete maximum principle under some conditions on the mesh and the diffusion tensor. We prove the convergence of the scheme, only supposing the shape regularity condition for the original mesh. We use a priori estimates and the Kolmogorov relative compactness theorem for this purpose. The proposed scheme is robust, only 5-point (7-point in space dimension three), locally conservative, efficient, and stable, which is confirmed by a numerical experiment. We finally propose a version of this scheme for nonmatching grids, combining this time the finite volume method with the piecewise linear conforming finite element method. We then apply this version to contaminant transport simulation in porous media.

1.1 Introduction

Degenerate parabolic equations arise in many contexts, such as flow in porous media or free boundary problems. This chapter is motivated by the modeling of contaminant transport in porous media with equilibrium adsorption reaction, see [19, 25], which typically involves a convection–reaction–diffusion equation of the form

$$\frac{\partial\beta(c)}{\partial t} - \nabla \cdot (\mathbf{S}\nabla c) + \mu\nabla \cdot (c\mathbf{v}) + F(c) = q, \quad (1.1)$$

where c is the unknown concentration of the contaminant, the function $\beta(\cdot)$ represents time evolution and equilibrium adsorption reaction and is supposed to be continuous and increasing with the growth bounded from below by a positive constant, \mathbf{S} is the diffusion–dispersion tensor, \mathbf{v} is the velocity field in the convection term (given for instance by the Darcy law), the function $F(\cdot)$ represents the changes due to chemical reactions, q stands for the sources, and finally, μ is a scalar parameter. Equation (1.1) is degenerate parabolic since β' may be unbounded, generally dominated by the convection term, and involves inhomogeneous and anisotropic (nonconstant full-matrix) diffusion–dispersion tensor.

A large variety of methods have been proposed for the discretization of degenerate parabolic equations. The conforming piecewise linear finite element method has been studied e.g. in [20, 39, 53, 93, 109], the cell-centered finite volume method in [22, 62, 63], the vertex-centered finite volume method in [6, 96], the finite difference method e.g. in [82], the mixed finite element method in [14, 46, 47], characteristic or Eulerian–Lagrangian methods e.g. in [40, 81], and relaxation schemes have been proposed e.g. in [78]. We shall follow in this chapter the finite element/finite volume approach.

The finite element method allows for an easy discretization of the diffusion term with a full tensor and does not impose any restrictions on the meshes. However, it is well-known that numerical instabilities may arise in the convection-dominated case. Recall that this method is locally conservative, contrary to what has been claimed for a long time, cf. [68, 76, 112], [61, Section III.12], or a detailed analysis given in [75]. The cell-centered finite volume method with an upwind discretization of the convection term ensures the stability and is extremely robust and computationally inexpensive. However, the mesh for the discretization of the diffusion term has to fulfill the following orthogonality property: the line segment relying the emplacement of the unknowns in two neighboring volumes has to be orthogonal to the side (edge in space dimension two and face in space dimension three) between these volumes, cf. [61]. Also, there is no straightforward way to apply this finite volume method to problems with full diffusion tensors. Various “multi-point” schemes where the approximation of the flux through an edge involves several scalar unknowns have been proposed, cf. e.g. [1, 7, 44, 58, 65]. However, such schemes require using more points than the classical 4 points for triangular meshes and 5 points for quadrangular meshes in space dimension two, making the schemes less robust and more susceptible to numerical instabilities. Their extension to three-dimensional unstructured meshes is also not straightforward (with the exception of the scheme proposed in [58]).

A quite intuitive idea is hence to combine a finite element discretization of the diffusion term with a finite volume discretization of the other terms of (1.1), trying to use the “best of both worlds”. Schemes combining conforming piecewise linear finite elements on triangles for the diffusion term with $\mathbf{S} = Id$ and finite volumes on dual volumes associated with the vertices, proposed and studied in [48, 67, 90] for fluid mechanics equations, are indeed quite efficient. Our motivation is to extend these ideas to degenerate parabolic problems, to the

combination of the mixed-hybrid finite element and finite volume methods, to inhomogeneous and anisotropic diffusion–dispersion tensors, to space dimension three, and finally to meshes only satisfying the shape regularity condition. We shall also extend such schemes to the case of nonmatching grids in Appendix 1.9.

Let us now introduce the combined scheme that we analyze in this chapter. We consider a triangulation of the space domain consisting of simplices (triangles in space dimension two and tetrahedra in space dimension three). We next construct a dual mesh where the dual volumes are associated with the sides (edges or faces). To construct a dual volume, one connects the barycentres of two neighboring simplices through the vertices of their common side. We finally place the unknowns in the barycentres of the sides. For the discretization of the diffusion term of (1.1), we consider the piecewise linear nonconforming (Crouzeix–Raviart, cf. [45]) finite element method or the mixed-hybrid finite element method where the only unknowns are the Lagrange multipliers, cf. [15, 33, 108]. We recall that the elements of the obtained stiffness matrices naturally express the coefficients for the discrete diffusive fluxes between the unknowns. We obtain the combined scheme by performing a finite volume discretization of (1.1) over the dual mesh and by replacing the finite volume stiffness matrix corresponding to the diffusion term by one of the above finite element stiffness matrices. The combination of finite volumes with nonconforming finite elements was originally proposed and analyzed in [10] as a semi-implicit discretization of a convection–diffusion equation with a nonlinear convection term in space dimension two. As far as we know, the combination of the finite volume method with the mixed-hybrid method is new. However, the two finite element stiffness matrices are very close. For a piecewise constant diffusion tensor, they completely coincide (see [15, 38]), and for a general diffusion tensor, the stiffness matrix of the mixed-hybrid method is the stiffness matrix of the nonconforming method with a piecewise constant diffusion tensor, given as the elementwise harmonic average of the original one (see Lemma 1.8.1 in Appendix 1.8).

We propose the combined scheme for the equation (1.1) in combination with the backward Euler finite difference time stepping. We can mention its following advantages. The scheme is stable since we avoid spurious oscillations in the convection-dominated case by checking the local Péclet number and by adding exactly the necessary amount of upstream weighting. It inherits the diffusion properties of nonconforming/mixed-hybrid finite elements, enabling in particular the use of general meshes and the discretization of anisotropic diffusion tensors. It possesses a discrete maximum principle in the case where all transmissibilities are non-negative. This happens for instance when the diffusion tensor reduces to a scalar function and when the angles between the outward normal vectors of sides of each simplex in the triangulation are greater or equal to $\pi/2$. The scheme is next locally conservative. It is only 5-point in space dimension two and 7-point in space dimension three. It finally permits to efficiently discretize degenerate parabolic problems: when we search for the discrete unknowns corresponding to $\beta(c)$, the resulting system of nonlinear algebraic equations can be solved by the Newton method without any parabolic regularization (cf. [20]) or perturbation of initial and boundary conditions (cf. [98, 99]), which make the equation uniformly parabolic. Moreover, the resulting matrices are diagonal for the part of the unknowns corresponding to the region where the approximate solution is zero.

Our numerical scheme permits to construct approximate solutions that are piecewise constant on the dual mesh or piecewise linear on the primal simplicial mesh and continuous in the barycentres of the sides of the simplices. We prove the convergence of both these approximations to a weak solution of the continuous problem in this chapter. The methods of proof are based upon the Kolmogorov relative compactness theorem and the finite volume tools from [61]. We extend these tools onto schemes with negative transmissibilities, for cases where

the discrete maximum principle is not satisfied, and for (dual) meshes not necessarily satisfying the orthogonality property. We only need the shape regularity (minimal angle) assumption for the primal triangulation, we require neither the inverse assumption (bounded ratio between the diameters of elements in the primal mesh), nor any maximal angle condition, as it was the case in [10]. We only suppose that β is continuous with the growth bounded from below in the case where the discrete maximum principle is satisfied. In the general case we require in addition β to be bounded on some interval and Lipschitz-continuous outside this interval. There is no restriction on the maximal time step in the case where F is nondecreasing. If F does not possess this property, we impose an appropriate maximal time step condition. For the sake of simplicity, we only consider the case of a homogeneous Dirichlet boundary condition. Extensions to other types of boundary conditions and to the case where the equation (1.1) involves a nonlinear convection term are possible, using the techniques from [61] and [62].

The rest of the chapter is organized as follows. In Section 1.2 we state the assumptions on the data and present a weak formulation of the continuous problem. In Section 1.3 we define the approximation spaces and introduce the combined finite volume–nonconforming/mixed-hybrid finite element scheme. In Section 1.4 we present some properties of this scheme and prove that it possesses a unique solution, which satisfies a discrete maximum principle under the hypotheses stated above. In Section 1.5 we derive a priori estimates and estimates on differences of time and space translates for the approximate solutions. Finally, in Section 1.6, using the Kolmogorov relative compactness theorem, we prove the convergence of a subsequence of the sequence of approximate solutions to a weak solution of the continuous problem. We present the results of a numerical experiment in Section 1.7 and we give some technical lemmas in Appendix 1.8. Finally, in Appendix 1.9, we propose a version of this scheme for nonmatching grids, combining this time the finite volume method with the piecewise linear conforming finite element method. We then apply this version to contaminant transport simulation in porous media.

1.2 The nonlinear degenerate parabolic problem

We consider the equation (1.1) in a polygonal domain (open, bounded, and connected set) $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, with boundary $\partial\Omega$ on the time interval $(0, T)$, $0 < T < \infty$, and denote $Q_T := \Omega \times (0, T)$. We impose the initial condition by

$$c(\cdot, 0) = c_0 \quad \text{in } \Omega \tag{1.2}$$

and the homogeneous Dirichlet boundary condition by

$$c = 0 \quad \text{on } \partial\Omega \times (0, T). \tag{1.3}$$

Let us consider a domain $S \subset \mathbb{R}^d$. We use the standard notation $L^p(S)$ and $\mathbf{L}^p(S) = [L^p(S)]^d$ for the Lebesgue spaces on S , $(\cdot, \cdot)_{0,S}$ stands for the $L^2(S)$ or $\mathbf{L}^2(S)$ inner product, and $\|\cdot\|_{0,S}$ for the associated norm. We use $d\mathbf{x}$ as the integration symbol for the Lebesgue measure on S , $d\gamma(\mathbf{x})$ for the Lebesgue measure on a hyperplane of S , and dt for the Lebesgue measure on $(0, T)$. We denote by $|S|$ the d -dimensional Lebesgue measure of S , by $|\sigma|$ the $(d-1)$ -dimensional Lebesgue measure of σ , a part of a hyperplane in \mathbb{R}^d , and by $|\mathbf{s}|$ the length of a segment \mathbf{s} . The diameter of S is the supremum of the lengths of all the line segments \mathbf{s} such that $\mathbf{s} \subset S$. Next, $H^1(S)$ and $H_0^1(S)$ are the Sobolev spaces of functions with square-integrable weak derivatives and $\mathbf{H}(\text{div}, S)$ is the space of vector functions with square-integrable weak divergences, $\mathbf{H}(\text{div}, S) = \{\mathbf{v} \in \mathbf{L}^2(S); \nabla \cdot \mathbf{v} \in L^2(S)\}$. In the subsequent

text we will denote by C_A , c_A a constant basically dependent on a quantity A but always independent of the discretization parameters h and Δt whose definition we shall give later. We make the following assumption on the data:

Assumption (A) (Data)

(A1) $\beta \in C(\mathbb{R})$, $\beta(0) = 0$ is a strictly increasing function such that

$$|\beta(a) - \beta(b)| \geq c_\beta |a - b|, \quad c_\beta > 0$$

for all $a, b \in \mathbb{R}$

or

(A2) in addition to (A1), there exists $P \in \mathbb{R}$, $P > 0$, such that $|\beta(x)| \leq C_\beta$ in $[-P, P]$, $C_\beta > 0$, and β is Lipschitz-continuous with a constant L_β on $(-\infty, -P]$ and $[P, +\infty)$;

(A3) $\mathbf{S}_{ij} \in L^\infty(Q_T)$, $|\mathbf{S}_{ij}| \leq C_S/d$ a.e. in Q_T , $1 \leq i, j \leq d$, $C_S > 0$, \mathbf{S} is a symmetric and uniformly positive definite tensor for almost all $t \in (0, T)$ with a constant $c_S > 0$, i.e.

$$\mathbf{S}(\mathbf{x}, t) \boldsymbol{\eta} \cdot \boldsymbol{\eta} \geq c_S \boldsymbol{\eta} \cdot \boldsymbol{\eta} \quad \forall \boldsymbol{\eta} \in \mathbb{R}^d, \text{ for a.e. } (\mathbf{x}, t) \in Q_T;$$

(A4) $\mathbf{v} \in \mathbf{L}^2(0, T; \mathbf{H}(\text{div}, \Omega)) \cap \mathbf{L}^\infty(Q_T)$ satisfies $\nabla \cdot \mathbf{v} = q_S \geq 0$ a.e. in Q_T , $|\mathbf{v} \cdot \mathbf{n}| \leq C_v$, $C_v > 0$, a.e. on $l \times (0, T)$ for each hyperplane $l \subset \Omega$ with the normal vector \mathbf{n} ;

(A5) $\mu \geq 1$;

(A6) $F(0) = 0$, F is a nondecreasing, Lipschitz-continuous function with a constant L_F

or

(A7) $F(0) = 0$, F is a Lipschitz-continuous function with a constant L_F and $xF(x) \geq 0$ for $x < 0$ and $x > M$, $M > 0$;

(A8) $q \in L^2(Q_T)$, where $q = qc_S$ with $c_S \in L^\infty(Q_T)$, $0 \leq c_S \leq M$ a.e. in Q_T ;

(A9) $c_0 \in L^\infty(\Omega)$, $0 \leq c_0 \leq M$ a.e. in Ω .

Remark 1.2.1. (Hypotheses on β) In contaminant transport problems one typically has $\beta(c) = c + c^\alpha$, $\alpha \in (0, 1)$. Assumption (A1) generalizes this type of functions; we in particular do not limit the number of points where β' explodes. As we shall see, we will be able to prove the convergence of the combined scheme with this assumption only for the case where the discrete maximum principle holds. In the general case we add Assumption (A2), which is however still satisfied by all realistic functions β .

We now give the definition of a weak solution of (1.1)–(1.3), following essentially [83].

Definition 1.2.2. (Weak solution) We say that a function c is a weak solution of the problem (1.1)–(1.3) if

(i) $c \in L^2(0, T; H_0^1(\Omega))$,

(ii) $\beta(c) \in L^\infty(0, T; L^2(\Omega))$,

(iii) c satisfies the integral equality

$$\begin{aligned} & - \int_0^T \int_\Omega \beta(c) \varphi_t \, d\mathbf{x} \, dt - \int_\Omega \beta(c_0) \varphi(\cdot, 0) \, d\mathbf{x} + \int_0^T \int_\Omega \mathbf{S} \nabla c \cdot \nabla \varphi \, d\mathbf{x} \, dt - \\ & - \mu \int_0^T \int_\Omega c \mathbf{v} \cdot \nabla \varphi \, d\mathbf{x} \, dt + \int_0^T \int_\Omega F(c) \varphi \, d\mathbf{x} \, dt = \int_0^T \int_\Omega q \varphi \, d\mathbf{x} \, dt \\ & \text{for all } \varphi \in L^2(0, T; H_0^1(\Omega)) \text{ with } \varphi_t \in L^\infty(Q_T), \varphi(\cdot, T) = 0. \end{aligned}$$

Remark 1.2.3. (Existence of a weak solution) *The existence of at least one weak solution is proved in Theorem 1.6.4 below.*

Remark 1.2.4. (Uniqueness of a weak solution) *For a slightly more restrictive hypothesis on the data than that given in Assumption (A), the uniqueness of a weak solution given by Definition 1.2.2 is guaranteed by [83]. Namely, no time-dependency of the diffusion–dispersion tensor \mathbf{S} is still required in [83].*

1.3 Combined finite volume–nonconforming/mixed-hybrid finite element scheme

We will describe the space and time discretizations, define the approximation spaces, and introduce the combined finite volume–finite element scheme in this section.

1.3.1 Space and time discretizations

In order to discretize the problem (1.1)–(1.3), we perform a triangulation \mathcal{T}_h of the domain Ω , consisting of closed simplices such that $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$ and such that if $K, L \in \mathcal{T}_h$, $K \neq L$, then $K \cap L$ is either an empty set or a common face, edge, or vertex of K and L . We denote by \mathcal{E}_h the set of all sides, by $\mathcal{E}_h^{\text{int}}$ the set of all interior sides, by $\mathcal{E}_h^{\text{ext}}$ the set of all exterior sides, and by \mathcal{E}_K the set of all the sides of an element $K \in \mathcal{T}_h$. We define $h := \max_{K \in \mathcal{T}_h} \text{diam}(K)$ and make the following shape regularity assumption on the family of triangulations $\{\mathcal{T}_h\}_h$:

Assumption (B) (Shape regularity of the space mesh)

There exists a positive constant $\kappa_{\mathcal{T}}$ such that

$$\min_{K \in \mathcal{T}_h} \frac{|K|}{\text{diam}(K)^d} \geq \kappa_{\mathcal{T}} \quad \forall h > 0.$$

Assumption (B) is equivalent to the more common requirement of the existence of a constant $\theta_{\mathcal{T}} > 0$ such that

$$\max_{K \in \mathcal{T}_h} \frac{\text{diam}(K)}{\rho_K} \leq \theta_{\mathcal{T}} \quad \forall h > 0, \tag{1.4}$$

where ρ_K is the diameter of the largest ball inscribed in the simplex K .

We also use a dual partition \mathcal{D}_h of Ω such that $\overline{\Omega} = \bigcup_{D \in \mathcal{D}_h} D$. There is one dual element D associated with each side $\sigma_D \in \mathcal{E}_h$. We construct it by connecting the barycentres of every $K \in \mathcal{T}_h$ that contains σ_D through the vertices of σ_D . For $\sigma_D \in \mathcal{E}_h^{\text{ext}}$, the contour of D is completed by the side σ_D itself. We refer to Fig. 1.1 for the two-dimensional case. We denote by Q_D the barycentre of the side σ_D . As for the primal mesh, we set \mathcal{F}_h , $\mathcal{F}_h^{\text{int}}$, $\mathcal{F}_h^{\text{ext}}$, and \mathcal{F}_D for the dual mesh sides. We denote by $\mathcal{D}_h^{\text{int}}$ the set of all interior and by $\mathcal{D}_h^{\text{ext}}$ the set of all boundary dual volumes. We finally denote by $\mathcal{N}(D)$ the set of all adjacent volumes to the volume D ,

$$\mathcal{N}(D) := \{E \in \mathcal{D}_h; \exists \sigma \in \mathcal{F}_h^{\text{int}} \text{ such that } \sigma = \partial D \cap \partial E\}$$

and remark that

$$|K \cap D| = \frac{|K|}{d+1}, \tag{1.5}$$

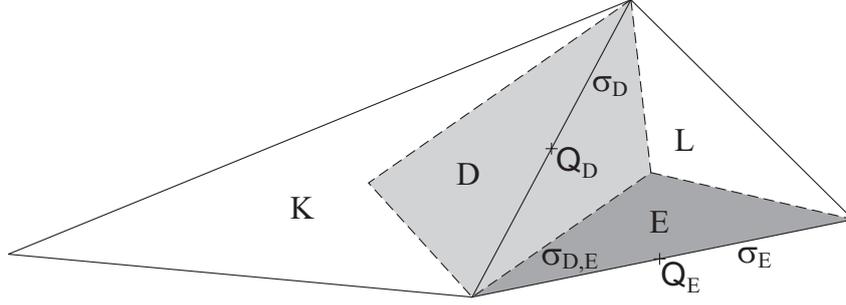


Figure 1.1: Triangles $K, L \in \mathcal{T}_h$ and dual volumes $D, E \in \mathcal{D}_h$ associated with edges $\sigma_D, \sigma_E \in \mathcal{E}_h$

for each $K \in \mathcal{T}_h$ and $D \in \mathcal{D}_h$ such that $\sigma_D \in \mathcal{E}_K$. For $E \in \mathcal{N}(D)$, we also set $d_{D,E} := |Q_E - Q_D|$, $\sigma_{D,E} := \partial D \cap \partial E$, and $K_{D,E}$ the element of \mathcal{T}_h such that $\sigma_{D,E} \subset K_{D,E}$.

We suppose the partition of the time interval $(0, T)$ such that $0 = t_0 < \dots < t_n < \dots < t_N = T$ and define $\Delta t_n := t_n - t_{n-1}$ and $\Delta t := \max_{1 \leq n \leq N} \Delta t_n$. In the case where Assumption (A6) is satisfied we do not impose any restriction on the time step. When only Assumption (A7) is satisfied, we suppose in addition:

Assumption (C) (Maximum time step for decreasing F)

The following maximum time step condition is satisfied:

$$\Delta t < \frac{c_\beta}{L_F}.$$

We define the following finite-dimensional spaces:

$$\begin{aligned} X_h &:= \{ \varphi_h \in L^2(\Omega); \varphi_h|_K \text{ is linear } \forall K \in \mathcal{T}_h, \\ &\quad \varphi_h \text{ is continuous at the points } Q_D, D \in \mathcal{D}_h^{\text{int}} \}, \\ X_h^0 &:= \{ \varphi_h \in X_h; \varphi_h(Q_D) = 0 \quad \forall D \in \mathcal{D}_h^{\text{ext}} \}. \end{aligned}$$

The basis of X_h is spanned by the shape functions φ_D , $D \in \mathcal{D}_h$, such that $\varphi_D(Q_E) = \delta_{DE}$, $E \in \mathcal{D}_h$, δ being the Kronecker delta. We recall that the approximations in these spaces are nonconforming since $X_h \not\subset H^1(\Omega)$. We equip X_h with the seminorm

$$\|c_h\|_{X_h}^2 := \sum_{K \in \mathcal{T}_h} \int_K |\nabla c_h|^2 \, d\mathbf{x},$$

which becomes a norm on X_h^0 . We have the following lemma:

Lemma 1.3.1. For all $c_h = \sum_{D \in \mathcal{D}_h} c_D \varphi_D \in X_h$, one has

$$\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \text{diam}(K_{D,E})^{d-2} (c_E - c_D)^2 \leq \frac{d+1}{2d\kappa_{\mathcal{T}}} \|c_h\|_{X_h}^2, \quad (1.6)$$

$$\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|}{d_{D,E}} (c_E - c_D)^2 \leq \frac{d+1}{2(d-1)\kappa_{\mathcal{T}}} \|c_h\|_{X_h}^2. \quad (1.7)$$

PROOF:

Obviously,

$$d_{D,E} \leq \frac{\text{diam}(K_{D,E})}{d}, \quad |\sigma_{D,E}| \leq \frac{\text{diam}(K_{D,E})^{d-1}}{d-1}. \quad (1.8)$$

Thus

$$\begin{aligned} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \text{diam}(K_{D,E})^{d-2} (c_E - c_D)^2 &\leq \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \text{diam}(K_{D,E})^{d-2} \left| \nabla c_h|_{K_{D,E}} \right|^2 d_{D,E}^2 \\ &\leq \frac{d+1}{2d} \sum_{K \in \mathcal{T}_h} \text{diam}(K)^d \left| \nabla c_h|_K \right|^2 \leq \frac{d+1}{2d\kappa_{\mathcal{T}}} \sum_{K \in \mathcal{T}_h} \left| \nabla c_h|_K \right|^2 |K| = \frac{d+1}{2d\kappa_{\mathcal{T}}} \|c_h\|_{X_h}^2, \end{aligned}$$

using the fact that the gradient of c_h is piecewise constant on \mathcal{T}_h , (1.8), the fact that each simplex $K \in \mathcal{T}_h$ contains exactly $\binom{d+1}{2} = \frac{(d+1)d}{2}$ dual sides, and Assumption (B). This proves (1.6). Similarly,

$$\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|}{d_{D,E}} (c_E - c_D)^2 \leq \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \left| \nabla c_h|_{K_{D,E}} \right|^2 d_{D,E} |\sigma_{D,E}| \leq \frac{d+1}{2(d-1)\kappa_{\mathcal{T}}} \|c_h\|_{X_h}^2. \quad \square$$

1.3.2 The combined scheme

We are now ready to present the combined scheme.

Definition 1.3.2. (Combined scheme) *The fully implicit combined finite volume–nonconforming/mixed-hybrid finite element scheme for the problem (1.1)–(1.3) reads: find the values c_D^n , $D \in \mathcal{D}_h$, $n \in \{0, 1, \dots, N\}$, such that*

$$c_D^0 = \frac{1}{|D|} \int_D c_0(\mathbf{x}) \, d\mathbf{x} \quad D \in \mathcal{D}_h^{\text{int}}, \quad (1.9a)$$

$$c_D^n = 0 \quad D \in \mathcal{D}_h^{\text{ext}}, \quad n \in \{0, 1, \dots, N\}, \quad (1.9b)$$

$$\begin{aligned} \frac{\beta(c_D^n) - \beta(c_D^{n-1})}{\Delta t_n} |D| - \sum_{E \in \mathcal{D}_h^{\text{int}}} \mathbb{S}_{D,E}^n c_E^n + \mu \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{c_{D,E}^n} + F(c_D^n) |D| &= q_D^n |D| \\ D \in \mathcal{D}_h^{\text{int}}, \quad n \in \{1, 2, \dots, N\}. \end{aligned} \quad (1.9c)$$

In (1.9a)–(1.9c) we have denoted

$$\mathbf{v}_{D,E}^n := \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \int_{\sigma_{D,E}} \mathbf{v}(\mathbf{x}, t) \cdot \mathbf{n}_{D,E} \, d\gamma(\mathbf{x}) \, dt \quad D \in \mathcal{D}_h^{\text{int}}, \quad E \in \mathcal{N}(D), \quad n \in \{1, 2, \dots, N\},$$

with $\mathbf{n}_{D,E}$ the unit normal vector of the side $\sigma_{D,E} \in \mathcal{F}_D$, outward to D , and

$$q_D^n := \frac{1}{\Delta t_n |D|} \int_{t_{n-1}}^{t_n} \int_D q(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad D \in \mathcal{D}_h, \quad n \in \{1, 2, \dots, N\}.$$

We refer to the matrix \mathbb{S}^n of the elements $\mathbb{S}_{D,E}^n$, $D, E \in \mathcal{D}_h^{\text{int}}$, at each discrete time t_n , $n \in \{1, 2, \dots, N\}$, as to the *diffusion matrix*. This matrix, the stiffness matrix of the nonconforming

or mixed-hybrid finite element method, is defined below. Finally, we define $\overline{c_{D,E}^n}$ for $D \in \mathcal{D}_h^{\text{int}}$, $E \in \mathcal{N}(D)$, and $n \in \{1, 2, \dots, N\}$ as follows:

$$\overline{c_{D,E}^n} := \begin{cases} c_D^n + \alpha_{D,E}^n (c_E^n - c_D^n) & \text{if } \mathbf{v}_{D,E}^n \geq 0 \\ c_E^n + \alpha_{D,E}^n (c_D^n - c_E^n) & \text{if } \mathbf{v}_{D,E}^n < 0 \end{cases}. \quad (1.10)$$

Here $\alpha_{D,E}^n$ is the coefficient of the amount of upstream weighting which is defined by

$$\alpha_{D,E}^n := \frac{\max \left\{ \min \left\{ \mathbb{S}_{D,E}^n, \frac{1}{2} \mu |\mathbf{v}_{D,E}^n| \right\}, 0 \right\}}{\mu |\mathbf{v}_{D,E}^n|}, \quad \mathbf{v}_{D,E}^n \neq 0. \quad (1.11)$$

We set $\alpha_{D,E}^n := 0$ if $\mathbf{v}_{D,E}^n = 0$. We remark that $\overline{c_{D,E}^n} = \widehat{c_{D,E}^n} + \text{sign}(\mathbf{v}_{D,E}^n) \alpha_{D,E}^n (c_E^n - c_D^n)$, where $\widehat{c_{D,E}^n}$ stands for full upstream weighting.

Remark 1.3.3. (Numerical flux) *We can easily see from (1.11) that $0 \leq \alpha_{D,E}^n \leq 1/2$, i.e. the numerical flux defined by (1.10) ranges from the centered scheme to the full upstream weighting. The amount of upstream weighting is set with respect to the local proportion of convection and diffusion.*

Remark 1.3.4. (Necessity to construct the dual mesh) *If we know the values of the fluxes $\mathbf{v}_{D,E}^n$, sources q_D^n , and initial conditions c_D^0 , we have no need to physically construct the dual mesh. Indeed, thanks to (1.5), expressing $|D|$ is immediate.*

We now turn to the definition of the diffusion matrix. To this purpose, we first set

$$\tilde{\mathbf{S}}^n(\mathbf{x}) := \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \mathbf{S}(\mathbf{x}, t) dt \quad \mathbf{x} \in \Omega, \quad n \in \{1, 2, \dots, N\}.$$

Diffusion matrix from the nonconforming method

The diffusion matrix \mathbb{S}^n given by the stiffness matrix \mathbb{P}^n of the nonconforming method writes in the form

$$\mathbb{S}_{D,E}^n := \mathbb{P}_{D,E}^n = - \sum_{K \in \mathcal{T}_h} (\mathbf{S}^n \nabla \varphi_E, \nabla \varphi_D)_{0,K} \quad D, E \in \mathcal{D}_h, \quad n \in \{1, 2, \dots, N\}, \quad (1.12)$$

where

$$\mathbf{S}^n(\mathbf{x}) = \tilde{\mathbf{S}}^n(\mathbf{x}) \quad n \in \{1, 2, \dots, N\}, \quad \mathbf{x} \in \Omega. \quad (1.13)$$

In fact, the members of $\mathbb{S}_{D,E}^n$ for $D \in \mathcal{D}_h^{\text{ext}}$ or $E \in \mathcal{D}_h^{\text{ext}}$ do not occur in the scheme (1.9a)–(1.9c). It will however show convenient to define these values.

Diffusion matrix from the mixed-hybrid method

Using the analytic form of the stiffness matrix \mathbb{M}^n of the mixed-hybrid method given in Lemma 1.8.1 in Appendix 1.8, we can define the diffusion matrix \mathbb{S}^n by

$$\mathbb{S}_{D,E}^n := \mathbb{M}_{D,E}^n = - \sum_{K \in \mathcal{T}_h} (\mathbf{S}^n \nabla \varphi_E, \nabla \varphi_D)_{0,K} \quad D, E \in \mathcal{D}_h, \quad n \in \{1, 2, \dots, N\}, \quad (1.14)$$

where

$$\mathbf{S}^n(\mathbf{y}) = \left(\frac{1}{|K|} \int_K [\tilde{\mathbf{S}}^n(\mathbf{x})]^{-1} d\mathbf{x} \right)^{-1} \quad \mathbf{y} \in K, \quad K \in \mathcal{T}_h, \quad n \in \{1, 2, \dots, N\}. \quad (1.15)$$

Remark 1.3.5. (Stiffness matrices of nonconforming and mixed-hybrid methods)

We remark that the stiffness matrix of the mixed-hybrid method (1.14) is the stiffness matrix of the nonconforming method (1.12) with a piecewise constant diffusion tensor, given as the inverse of the elementwise average of the inverse of the original one. In particular for an elementwise constant diffusion tensor, the stiffness matrices coincide, whereas for a general diffusion tensor, (1.12) uses its arithmetic and (1.14) its harmonic average.

Remark 1.3.6. (Comparison with a pure finite volume scheme)

Let us consider \mathcal{T}_h consisting of equilateral simplices and $\mathbf{S} = Id$. Then the segments $[Q_D, Q_E]$ are orthogonal to the dual sides $\sigma_{D,E}$ and one has $\mathbb{P}_{D,E}^n = \mathbb{M}_{D,E}^n = \frac{|\sigma_{D,E}|}{d_{D,E}}$, $E \in \mathcal{N}(D)$. Thus, in view of Corollary 1.4.2 below, the pure cell-centered finite volume scheme completely coincides in this case with the combined one. One may regard in this sense the combined scheme as an extension of the pure finite volume scheme to general triangulations and full-matrix diffusion tensors, which does not extend the original 5-point (7-point in space dimension three) stencil.

Remark 1.3.7. (Comparison of a combined finite volume–finite element scheme with pure finite volume schemes)

We recall here that for triangular meshes, the discretization of a Laplacian by the piecewise linear conforming finite element method coincides with that by the vertex-centered finite volume method [6, 96], which is also named the box scheme [17, 71], the finite volume element scheme [35, 56], or control volume finite element scheme [68, 114], see [17, Lemma 3]. Finally, for Delaunay triangulations (the sums of two opposite angles to all edges are less or equal to π), constructing the control volumes with the aid of orthogonal bisectors, these discretizations are equivalent to that by the cell-centered finite volume method, see [61, Section III.12], cf. also [76, 112]. Hence, when $\mathbf{S} = Id$ and for a Delaunay triangular mesh with the above construction of control volumes, the combined finite volume–finite element scheme [67, 90], the vertex-centered finite volume scheme [6, 96], and the cell-centered finite volume scheme [61, 62] for the discretization of (1.1) coincide.

In the sequel we shall consider apart the following special case:

Assumption (D) (Diffusion matrix)

All transmissibilities are non-negative, i.e.

$$\mathbb{S}_{D,E}^n \geq 0 \quad \forall D \in \mathcal{D}_h^{\text{int}}, E \in \mathcal{N}(D) \quad \forall n \in \{1, 2, \dots, N\}.$$

Since

$$\nabla \varphi_D|_K = \frac{|\sigma_D|}{|K|} \mathbf{n}_{\sigma_D} \quad K \in \mathcal{T}_h, \sigma_D \in \mathcal{E}_K \quad (1.16)$$

with \mathbf{n}_{σ_D} the unit normal vector of the side σ_D , outward to K , one can immediately see that Assumption (D) is satisfied e.g. when the diffusion tensor reduces to a scalar function and when the magnitude of the angles between \mathbf{n}_{σ_D} , $\sigma_D \in \mathcal{E}_K$, for all $K \in \mathcal{T}_h$ is greater or equal to $\pi/2$.

1.4 Existence, uniqueness, and discrete properties

In this section we first present some technical lemmas. We then show the conservativity of the scheme, the coercivity of the bilinear diffusion form corresponding to the diffusion term, and an a priori estimate for an extended scheme, which we shall need later in the proof of the existence of the solution of the discrete problem. Finally, we prove the uniqueness of this solution and the discrete maximum principle when Assumption (D) is satisfied.

1.4.1 Discrete properties of the scheme

Lemma 1.4.1. *For all $D \in \mathcal{D}_h$ and $n \in \{1, 2, \dots, N\}$, $\mathbb{S}_{D,D}^n = - \sum_{E \in \mathcal{N}(D)} \mathbb{S}_{D,E}^n$.*

PROOF:

We will show the assertion for $d = 2$; the case $d = 3$ is similar. We present the proof for the nonconforming method, which in view of Remark 1.3.5 implies the same result for the mixed-hybrid method. Let us consider a fixed dual volume $D \in \mathcal{D}_h$. The edge σ_D associated with D is shared by at most two triangles, which we denote by K and L . The sum over $K \in \mathcal{T}_h$ in (1.12) for $\mathbb{S}_{D,D}^n$ reduces just to these triangles, considering the definition of the basis function φ_D . We denote the dual volumes associated with the two other edges of L by E_1 and E_2 . Similarly, the sum over $K \in \mathcal{T}_h$ in (1.12) for \mathbb{S}_{D,E_1}^n and \mathbb{S}_{D,E_2}^n reduces to L . Thus it is sufficient to prove that

$$-(\mathbf{S}^n \nabla \varphi_D, \nabla \varphi_D)_{0,L} = (\mathbf{S}^n \nabla \varphi_{E_1}, \nabla \varphi_D)_{0,L} + (\mathbf{S}^n \nabla \varphi_{E_2}, \nabla \varphi_D)_{0,L},$$

since the eventual contribution of the element K is similar. However, this is immediate, since

$$-\varphi_D|_L = (\varphi_{E_1} + \varphi_{E_2})|_L - 1. \quad \square$$

Corollary 1.4.2. *Using the fact that $\mathbb{S}_{D,E}^n \neq 0$ only if $E \in \mathcal{N}(D)$ or if $E = D$ and Lemma 1.4.1, one has*

$$\sum_{E \in \mathcal{D}_h} \mathbb{S}_{D,E}^n c_E^n = \sum_{E \in \mathcal{N}(D)} \mathbb{S}_{D,E}^n c_E^n + \mathbb{S}_{D,D}^n c_D^n = \sum_{E \in \mathcal{N}(D)} \mathbb{S}_{D,E}^n (c_E^n - c_D^n).$$

Theorem 1.4.3. (Conservativity of the scheme) *The scheme (1.9a)–(1.9c) is conservative with respect to the dual mesh \mathcal{D}_h .*

PROOF:

Let us take two fixed neighboring dual volumes E and D , $D \in \mathcal{D}_h^{\text{int}}$. Using Corollary 1.4.2 and (1.9b), we can express the discrete diffusive flux from D to E as $-\mathbb{S}_{D,E}^n (c_E^n - c_D^n)$. The discrete diffusive flux from E to D is $-\mathbb{S}_{E,D}^n (c_D^n - c_E^n)$, i.e. we have their equality up to the sign, considering that $\mathbb{S}_{D,E}^n = \mathbb{S}_{E,D}^n$ for all $n \in \{1, 2, \dots, N\}$, which follows from (1.12) or (1.14) using the symmetry of the tensor \mathbf{S} .

For the discrete convective flux from D to E , we have $\mu \mathbf{v}_{D,E}^n [c_D^n + \alpha_{D,E}^n (c_E^n - c_D^n)]$, supposing $\mathbf{v}_{D,E}^n \geq 0$. For the discrete convective flux from E to D , we have $\mu \mathbf{v}_{E,D}^n [c_D^n + \alpha_{E,D}^n (c_E^n - c_D^n)]$, i.e. again the equality up to the sign, considering that $\mathbf{v}_{D,E}^n = -\mathbf{v}_{E,D}^n$ and that $\alpha_{D,E}^n = \alpha_{E,D}^n$, which follows from $\mathbb{S}_{D,E}^n = \mathbb{S}_{E,D}^n$. For $\mathbf{v}_{D,E}^n < 0$, the proof is similar. Hence the combined finite volume–finite element scheme is conservative as the pure finite volume is, cf. [61]. \square

Lemma 1.4.4. *For all $D \in \mathcal{D}_h^{\text{int}}$ and $n \in \{1, 2, \dots, N\}$,*

$$\sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \widehat{c_{D,E}^n} = \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) + (q_S)_D^n c_D^n |D|,$$

where $(\mathbf{v}_{D,E}^n)^- := \min\{\mathbf{v}_{D,E}^n, 0\}$ and

$$(q_S)_D^n := \frac{1}{\Delta t_n |D|} \int_{t_{n-1}}^{t_n} \int_D q_S(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad \forall D \in \mathcal{D}_h, \forall n \in \{1, 2, \dots, N\}.$$

PROOF:

Considering that $\mathbf{v}_{D,E}^n = (\mathbf{v}_{D,E}^n)^+ + (\mathbf{v}_{D,E}^n)^-$, where $(\mathbf{v}_{D,E}^n)^+ := \max\{\mathbf{v}_{D,E}^n, 0\}$, we have

$$\begin{aligned}
 \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \widehat{c_{D,E}^n} &= \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^+ c_D^n + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- c_E^n \\
 &= \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n c_D^n + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) \\
 &= c_D^n \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \sum_{E \in \mathcal{N}(D)} \int_{\sigma_{D,E}} \mathbf{v}(\mathbf{x}, t) \cdot \mathbf{n}_{D,E} \, d\gamma(\mathbf{x}) \, dt \\
 &\quad + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) = c_D^n \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \int_D \nabla \cdot \mathbf{v}(\mathbf{x}, t) \, dx \, dt \\
 &\quad + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) = c_D^n (q_S)_D^n |D| \\
 &\quad + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n),
 \end{aligned}$$

using Assumption (A4). \square

Lemma 1.4.5. *For all $c_h = \sum_{D \in \mathcal{D}_h} c_D \varphi_D \in X_h$ and $n \in \{1, 2, \dots, N\}$,*

$$- \sum_{D \in \mathcal{D}_h} c_D \sum_{E \in \mathcal{D}_h} \mathbb{S}_{D,E}^n c_E \geq c_S \|c_h\|_{X_h}^2.$$

PROOF:

We have

$$- \sum_{D \in \mathcal{D}_h} c_D \sum_{E \in \mathcal{D}_h} \mathbb{S}_{D,E}^n c_E = \sum_{K \in \mathcal{T}_h} (\mathbf{S}^n \nabla c_h, \nabla c_h)_{0,K} \geq c_S \|c_h\|_{X_h}^2,$$

using (1.12) or (1.14) and Assumption (A3) and the subsequent uniform positive definiteness of the diffusion tensors (1.13) and (1.15). \square

Lemma 1.4.6. *For all $c_h = \sum_{D \in \mathcal{D}_h} c_D \varphi_D \in X_h$ and $n \in \{1, 2, \dots, N\}$,*

$$\left| - \sum_{D \in \mathcal{D}_h} c_D \sum_{E \in \mathcal{D}_h} \mathbb{S}_{D,E}^n c_E \right| \leq C_S \|c_h\|_{X_h}^2. \tag{1.17}$$

Moreover, for all $D \in \mathcal{D}_h$, $E \in \mathcal{N}(D)$, and $n \in \{1, 2, \dots, N\}$,

$$|\mathbb{S}_{D,E}^n| \leq \frac{C_S \operatorname{diam}(K_{D,E})^{d-2}}{\kappa_{\mathcal{T}} (d-1)^2}. \tag{1.18}$$

PROOF:

We have

$$\left| - \sum_{D \in \mathcal{D}_h} c_D \sum_{E \in \mathcal{D}_h} \mathbb{S}_{D,E}^n c_E \right| = \left| \sum_{K \in \mathcal{T}_h} (\mathbf{S}^n \nabla c_h, \nabla c_h)_{0,K} \right| \leq C_S \|c_h\|_{X_h}^2,$$

using (1.12) or (1.14), Assumption (A3), and (1.13) or (1.15). Considering (1.12) or (1.14), where the sum reduces just to $K_{D,E} \in \mathcal{T}_h$ for $E \in \mathcal{N}(D)$, the equality (1.16), $|\sigma_D|, |\sigma_E| \leq \text{diam}(K_{D,E})^{d-1}/(d-1)$, and Assumption (B), we have

$$\begin{aligned} |\mathbb{S}_{D,E}^n| &\leq C_{\mathbf{S}} \left| \nabla \varphi_E|_{K_{D,E}} \right| \left| \nabla \varphi_D|_{K_{D,E}} \right| |K_{D,E}| = C_{\mathbf{S}} \frac{|\sigma_E|}{|K_{D,E}|} \frac{|\sigma_D|}{|K_{D,E}|} |K_{D,E}| \\ &\leq C_{\mathbf{S}} \frac{\text{diam}(K_{D,E})^{2d-2}}{(d-1)^2 |K_{D,E}|} \leq \frac{C_{\mathbf{S}}}{\kappa_{\mathcal{T}}} \frac{\text{diam}(K_{D,E})^{d-2}}{(d-1)^2}. \quad \square \end{aligned}$$

Lemma 1.4.7. *For all values c_D , $D \in \mathcal{D}_h$, such that $c_D = 0$ for all $D \in \mathcal{D}_h^{\text{ext}}$ and $n \in \{1, 2, \dots, N\}$,*

$$\sum_{D \in \mathcal{D}_h^{\text{int}}} c_D \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{c_{D,E}} \geq 0.$$

PROOF:

We can write

$$\begin{aligned} &\sum_{D \in \mathcal{D}_h^{\text{int}}} c_D \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{c_{D,E}} \\ &= \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \mathbf{v}_{D,E}^n \geq 0} \mathbf{v}_{D,E}^n \left(c_D (c_D - c_E) - \alpha_{D,E}^n (c_E - c_D)^2 \right) \\ &= \frac{1}{2} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \mathbf{v}_{D,E}^n \geq 0} \mathbf{v}_{D,E}^n (c_D^2 - c_E^2) + \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |\mathbf{v}_{D,E}^n| (c_E - c_D)^2 \left(\frac{1}{2} - \alpha_{D,E}^n \right) \\ &\geq \frac{1}{2} \sum_{D \in \mathcal{D}_h^{\text{int}}} c_D^2 \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n = \frac{1}{2} \sum_{D \in \mathcal{D}_h^{\text{int}}} c_D^2 (q_S)_D^n |D| \geq 0, \end{aligned}$$

where we have used the fact that $c_D = 0$ for all $D \in \mathcal{D}_h^{\text{ext}}$, the relation $2(a-b)a = (a-b)^2 + a^2 - b^2$, and rewritten the summation over interior dual sides with fixed denotation of the dual volumes sharing given side $\sigma_{D,E}$ such that $\mathbf{v}_{D,E}^n \geq 0$. In the last two estimates we have used, respectively, the fact that $0 \leq \alpha_{D,E}^n \leq 1/2$, which follows from (1.11), and Assumption (A4). \square

Theorem 1.4.8. (A priori estimate for an extended scheme) *Let us define an extended scheme by*

$$c_D^0 = \frac{1}{|D|} \int_D c_0(\mathbf{x}) \, d\mathbf{x} \quad D \in \mathcal{D}_h^{\text{int}}, \quad (1.19a)$$

$$c_D^n = 0 \quad D \in \mathcal{D}_h^{\text{ext}}, \, n \in \{0, 1, \dots, N\}, \quad (1.19b)$$

$$\begin{aligned} &u \frac{\beta(c_D^n) - \beta(c_D^{n-1})}{\Delta t_n} |D| - \sum_{E \in \mathcal{D}_h^{\text{int}}} \mathbb{S}_{D,E}^n c_E^n + u \mu \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{c_{D,E}^n} + u F(c_D^n) |D| \\ &= u q_D^n |D| \quad D \in \mathcal{D}_h^{\text{int}}, \, n \in \{1, 2, \dots, N\} \end{aligned} \quad (1.19c)$$

with $u \in [0, 1]$. Let $\delta > 0$ be arbitrary. Then

$$\sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D| < C_{\text{es}} \quad \forall n \in \{1, 2, \dots, N\}$$

with

$$C_{\text{es}} := \frac{4}{c_\beta} M \beta(M) |\Omega| + \frac{4T}{c_\beta^2} \|q\|_{0, Q_T}^2 + \frac{4}{c_\beta} L_F M^2 T |\Omega| + \delta.$$

PROOF:

We multiply (1.19c) by $\Delta t_n c_D^n$, sum over all $D \in \mathcal{D}_h^{\text{int}}$ and $n \in \{1, 2, \dots, k\}$, and use the fact that $u \geq 0$ and Lemmas 1.4.5 and 1.4.7. Further, for $c_D^n < 0$ or $c_D^n > M$, $F(c_D^n) c_D^n \geq 0$ follows from Assumption (A6) or (A7). When $0 \leq c_D^n \leq M$, $-F(c_D^n) c_D^n \leq |F(c_D^n)| |c_D^n| \leq L_F M^2$, which altogether yields

$$\begin{aligned} & u \sum_{n=1}^k \sum_{D \in \mathcal{D}_h^{\text{int}}} [\beta(c_D^n) - \beta(c_D^{n-1})] c_D^n |D| + c_S \sum_{n=1}^k \Delta t_n \|c_h^n\|_{X_h}^2 \\ & \leq u \sum_{n=1}^k \Delta t_n \sum_{D \in \mathcal{D}_h^{\text{int}}} c_D^n q_D^n |D| + u L_F M^2 \sum_{n=1}^k \sum_{D \in \mathcal{D}_h^{\text{int}}} \Delta t_n |D| \end{aligned} \quad (1.20)$$

with $c_h^n = \sum_{D \in \mathcal{D}_h} c_D^n \varphi_D$. Let us now introduce a function B ,

$$B(s) := \beta(s)s - \int_0^s \beta(\tau) \, d\tau \quad s \in \mathbb{R}.$$

One then can derive

$$B(c_D^n) - B(c_D^{n-1}) = [\beta(c_D^n) - \beta(c_D^{n-1})] c_D^n - \int_{c_D^{n-1}}^{c_D^n} [\beta(\tau) - \beta(c_D^{n-1})] \, d\tau.$$

Using that β is nondecreasing, one can easily show that

$$\int_{c_D^{n-1}}^{c_D^n} [\beta(\tau) - \beta(c_D^{n-1})] \, d\tau \geq 0.$$

In view of the two last expressions, one has

$$\sum_{n=1}^k \sum_{D \in \mathcal{D}_h^{\text{int}}} [B(c_D^n) - B(c_D^{n-1})] |D| \leq \sum_{n=1}^k \sum_{D \in \mathcal{D}_h^{\text{int}}} [\beta(c_D^n) - \beta(c_D^{n-1})] c_D^n |D|,$$

which yields

$$\sum_{D \in \mathcal{D}_h^{\text{int}}} B(c_D^k) |D| - \sum_{D \in \mathcal{D}_h^{\text{int}}} B(c_D^0) |D| \leq \sum_{n=1}^k \sum_{D \in \mathcal{D}_h^{\text{int}}} [\beta(c_D^n) - \beta(c_D^{n-1})] c_D^n |D|.$$

Using the growth condition on β from Assumption (A1), one can derive $B(s) \geq s^2 c_\beta / 2$ for all $s \in \mathbb{R}$, see Lemma 1.8.2 in Appendix 1.8. Thus, using in addition Assumption (A9)

$$\frac{c_\beta}{2} \sum_{D \in \mathcal{D}_h^{\text{int}}} (c_D^k)^2 |D| - M \beta(M) |\Omega| \leq \sum_{n=1}^k \sum_{D \in \mathcal{D}_h^{\text{int}}} [\beta(c_D^n) - \beta(c_D^{n-1})] c_D^n |D|.$$

We notice that

$$\sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \Delta t_n |D| (q_D^n)^2 \leq \|q\|_{0, Q_T}^2 \quad (1.21)$$

by the Cauchy–Schwarz inequality. Hence extending the summation over all $n \in \{1, 2, \dots, N\}$ and $D \in \mathcal{D}_h$ in the first term of the right-hand side of (1.20) and using the Cauchy–Schwarz and Young inequality, we have

$$\begin{aligned} \sum_{n=1}^k \Delta t_n \sum_{D \in \mathcal{D}_h^{\text{int}}} c_D^n q_D^n |D| &\leq \left(\sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D| \right)^{\frac{1}{2}} \|q\|_{0, Q_T} \\ &\leq \frac{\varepsilon}{2} \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D| + \frac{1}{2\varepsilon} \|q\|_{0, Q_T}^2. \end{aligned}$$

Hence, substituting these estimates into (1.20), we obtain

$$\begin{aligned} u \frac{c_\beta}{2} \max_{n \in \{1, 2, \dots, N\}} \sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D| + c_s \sum_{n=1}^k \Delta t_n \|c_h^n\|_{X_h}^2 &\leq uM\beta(M)|\Omega| \quad (1.22) \\ + u \frac{\varepsilon}{2} T \max_{n \in \{1, 2, \dots, N\}} \sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D| + u \frac{1}{2\varepsilon} \|q\|_{0, Q_T}^2 + uL_F M^2 T |\Omega|, \end{aligned}$$

considering also (1.19b) and the fact that k was arbitrarily chosen. We now choose $\varepsilon = c_\beta/(2T)$. When $u \neq 0$, this already yields the assertion of the lemma. When $u = 0$, it follows from (1.22) that $c_D^n = 0$ for all $D \in \mathcal{D}_h$ and all $n \in \{1, 2, \dots, N\}$, since in view of (1.19b), $\|\cdot\|_{X_h}$ is a norm on X_h . Thus the assertion of the lemma is trivially satisfied in this case. \square

1.4.2 Existence, uniqueness, and the discrete maximum principle

Theorem 1.4.9. (Existence of the solution of the discrete problem) *The problem (1.9a)–(1.9c) has at least one solution.*

PROOF:

The nonlinear system of equations given by (1.9b)–(1.9c) on each discrete time level t_n , $n \in \{1, 2, \dots, N\}$, can be written as

$$\mathcal{P}(C^n) - \mathbb{S}^n C^n + \mathbb{C}^n C^n + \mathcal{R}(C^n) = \mathcal{P}(C^{n-1}) + Q^n, \quad (1.23)$$

where $C^n \in \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$ ($|\mathcal{D}_h^{\text{int}}|$ is the cardinality of the set $\mathcal{D}_h^{\text{int}}$) is the vector of discrete unknowns c_D^n , $D \in \mathcal{D}_h^{\text{int}}$, \mathbb{S}^n , the diffusion matrix, and \mathbb{C}^n , the discretization of the convective term, are linear mappings from $\mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$ to $\mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$, \mathcal{P} and \mathcal{R} are, due to the continuity of β and F following from Assumption (A1), (A6), respectively, continuous mappings from $\mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$ to $\mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$, $Q^n \in \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$ is a constant vector, and $C^{n-1} \in \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$ is the vector of discrete values c_D^{n-1} on the previous time step. The vector C^0 is given by (1.9a). By Lemma 1.4.5 and (1.9b), $-\mathbb{S}^n$ is a positive definite and consequently invertible matrix. Thus also \mathbb{S}^n is an invertible matrix. Hence, (1.23) is equivalent to

$$C^n = G(C^n) := [\mathbb{S}^n]^{-1} [\mathcal{P}(C^n) + \mathbb{C}^n C^n + \mathcal{R}(C^n) - \mathcal{P}(C^{n-1}) - Q^n]. \quad (1.24)$$

Let $H(u, C^n) = uG(C^n)$ for $u \in [0, 1]$ and $C^n \in \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$. H is a continuous mapping from $[0, 1] \times \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$ to $\mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$. If we now consider the norm $|C^n| = \sum_{D \in \mathcal{D}_h^{\text{int}}} (c_D^n)^2 |D|$ on $\mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$, we

have from Theorem 1.4.8 that

$$|C^n| < C_{\text{es}} \quad \text{for all } (u, C^n) \in [0, 1] \times \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|} \text{ such that } C^n = H(u, C^n).$$

Therefore $C^n = H(u, C^n)$ has no solution lying on the boundary of the ball $B_{C_{\text{es}}} := \{C^n \in \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}, |C^n| < C_{\text{es}}\}$ for $u \in [0, 1]$. We thus can define for $u \in [0, 1]$ the (Brouwer) topological degree of the application $Id - H(u, \cdot)$ with respect to the ball $B_{C_{\text{es}}}$ and right-hand side 0, denoted by $d(Id - H(u, \cdot), B_{C_{\text{es}}}, 0)$. Then, the homotopy invariance of the degree ([49], Theorem 3.1 (d3)) leads to

$$d(Id - H(u, \cdot), B_{C_{\text{es}}}, 0) = d(Id - H(0, \cdot), B_{C_{\text{es}}}, 0)$$

for all $u \in [0, 1]$. Since $H(0, C^n)$ is a zero vector for all $C^n \in \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$, one has $d(Id - H(0, \cdot), B_{C_{\text{es}}}, 0) = d(Id, B_{C_{\text{es}}}, 0) = 1$, and thus there exists $C^n \in \mathbb{R}^{|\mathcal{D}_h^{\text{int}}|}$ such that $C^n - H(1, C^n) = 0$, i.e. C^n is the solution to (1.24). This proves the existence of a solution to (1.9b)–(1.9c) at each discrete time level t_n , $n \in \{1, 2, \dots, N\}$. \square

Theorem 1.4.10. (Uniqueness of the solution of the discrete problem) *The solution of the problem (1.9a)–(1.9c) is unique.*

PROOF:

We will prove the assertion by contradiction. Let us thus suppose that there exists $n \in \{1, 2, \dots, N\}$ such that $c_D^{n-1} = \tilde{c}_D^{n-1}$ for all $D \in \mathcal{D}_h^{\text{int}}$ but $c_D^n \neq \tilde{c}_D^n$ for some $D \in \mathcal{D}_h^{\text{int}}$. After subtracting the equation (1.9c) for c_D^n and \tilde{c}_D^n and denoting $s_D^n := c_D^n - \tilde{c}_D^n$, we have

$$\begin{aligned} & \frac{\beta(c_D^n) - \beta(\tilde{c}_D^n)}{\Delta t_n} |D| - \sum_{E \in \mathcal{D}_h^{\text{int}}} \mathbb{S}_{D,E}^n s_E^n + \mu \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{s_{D,E}^n} \\ & + F(c_D^n) |D| - F(\tilde{c}_D^n) |D| = 0 \quad D \in \mathcal{D}_h^{\text{int}}. \end{aligned}$$

We now multiply the above equality by s_D^n and sum the result over $D \in \mathcal{D}_h^{\text{int}}$. This yields, using Lemmas 1.4.5 and 1.4.7,

$$\sum_{D \in \mathcal{D}_h^{\text{int}}} [\beta(c_D^n) - \beta(\tilde{c}_D^n)] (c_D^n - \tilde{c}_D^n) \frac{|D|}{\Delta t_n} + \sum_{D \in \mathcal{D}_h^{\text{int}}} [F(c_D^n) - F(\tilde{c}_D^n)] (c_D^n - \tilde{c}_D^n) |D| \leq 0.$$

When Assumption (A6) is satisfied, this is already a contradiction, since from Assumption (A1), β is strictly increasing and F is nondecreasing in this case.

When only Assumption (A7) is satisfied, we have $|[F(c_D^n) - F(\tilde{c}_D^n)](c_D^n - \tilde{c}_D^n)| \leq L_F (c_D^n - \tilde{c}_D^n)^2$. In view of Assumption (A1), $[\beta(c_D^n) - \beta(\tilde{c}_D^n)](c_D^n - \tilde{c}_D^n) \geq c_\beta (c_D^n - \tilde{c}_D^n)^2$. Since

$$\sum_{D \in \mathcal{D}_h^{\text{int}}} (c_D^n - \tilde{c}_D^n)^2 |D| \neq 0,$$

$c_\beta/L_F \leq \Delta t_n$, which is a contradiction with Assumption (C) supposed in this case. \square

Theorem 1.4.11. (Discrete maximum principle) *Under Assumption (D), the solution of the problem (1.9a)–(1.9c) satisfies*

$$0 \leq c_D^n \leq M$$

for all $D \in \mathcal{D}_h$, $n \in \{1, 2, \dots, N\}$.

PROOF:

Setting $\mathbb{T}_{D,E}^n := \mathbb{S}_{D,E}^n - \mu |\mathbf{v}_{D,E}^n| \alpha_{D,E}^n$, $E \in \mathcal{N}(D)$, and using Corollary 1.4.2 and Lemma 1.4.4, we can rewrite (1.9c) as

$$\begin{aligned} & \frac{\beta(c_D^n) - \beta(c_D^{n-1})}{\Delta t_n} |D| - \sum_{E \in \mathcal{N}(D)} \mathbb{T}_{D,E}^n (c_E^n - c_D^n) + \mu \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) \\ & + \mu (q_S)_D^n c_D^n |D| + F(c_D^n) |D| = q_D^n |D| \quad D \in \mathcal{D}_h^{\text{int}}, n \in \{1, 2, \dots, N\}. \end{aligned} \quad (1.25)$$

In view of Assumption (D) and (1.11), one has $\mathbb{T}_{D,E}^n \geq 0$ for all $D \in \mathcal{D}_h^{\text{int}}$, $E \in \mathcal{N}(D)$, and $n \in \{1, 2, \dots, N\}$. We now make use of an induction argument. We remark that $0 \leq c_D^n \leq M$ is satisfied for $n = 0$ by Assumption (A9) and (1.9a) and (1.9b). Let us suppose that $0 \leq c_D^{n-1} \leq M$ for all $D \in \mathcal{D}_h^{\text{int}}$ and for a fixed $(n-1) \in \{0, 1, \dots, N-1\}$. Since $|\mathcal{D}_h|$ is finite, there exist $D_0, D_1 \in \mathcal{D}_h$ such that $c_{D_0}^n \leq c_D^n \leq c_{D_1}^n$ for all $D \in \mathcal{D}_h$. Using a contradiction argument we prove below that $c_{D_0}^n \geq 0$ and $c_{D_1}^n \leq M$. Suppose that $c_{D_0}^n < 0$. We remark that $D_0 \in \mathcal{D}_h^{\text{int}}$ because of (1.9b). Then, since $\mathbb{T}_{D_0,E}^n \geq 0$ and $-(\mathbf{v}_{D_0,E}^n)^- \geq 0$, we have

$$\sum_{E \in \mathcal{N}(D_0)} \mathbb{T}_{D_0,E}^n (c_E^n - c_{D_0}^n) + \mu \sum_{E \in \mathcal{N}(D_0)} -(\mathbf{v}_{D_0,E}^n)^- (c_E^n - c_{D_0}^n) \geq 0.$$

This yields, using (1.25)

$$\frac{\beta(c_{D_0}^n) - \beta(c_{D_0}^{n-1})}{\Delta t_n} |D_0| + \mu (q_S)_{D_0}^n c_{D_0}^n |D_0| + F(c_{D_0}^n) |D_0| - q_{D_0}^n |D_0| \geq 0.$$

Now $c_{D_0}^n < 0$ implies $\mu (q_S)_{D_0}^n c_{D_0}^n \leq 0$ and $F(c_{D_0}^n) \leq 0$ using, respectively, Assumption (A4) and (A5) and (A6) or (A7). Also $-q_{D_0}^n \leq 0$, using Assumption (A8). Thus $\beta(c_{D_0}^n) \geq \beta(c_{D_0}^{n-1})$, which is a contradiction, since β is strictly increasing from Assumption (A1).

Let us now suppose $c_{D_1}^n > M$. Again $D_1 \in \mathcal{D}_h^{\text{int}}$, because of (1.9b). Similarly as in the previous case, one comes to

$$\frac{\beta(c_{D_1}^n) - \beta(c_{D_1}^{n-1})}{\Delta t_n} |D_1| + \mu (q_S)_{D_1}^n c_{D_1}^n |D_1| + F(c_{D_1}^n) |D_1| - q_{D_1}^n |D_1| \leq 0.$$

We can estimate

$$-q_{D_1}^n |D_1| \geq -M (q_S)_{D_1}^n |D_1| \geq -\mu M (q_S)_{D_1}^n |D_1|$$

using, respectively, Assumption (A8), (A4), and (A5). It follows from (A6) or (A7) that $F(c_{D_1}^n) \geq 0$. This implies $\beta(c_{D_1}^n) \leq \beta(c_{D_1}^{n-1})$, which is again a contradiction, using Assumption (A1). \square

1.5 A priori estimates

In this section we give a priori estimates and estimates on differences of time and space translates of the approximate solutions that we shall define.

1.5.1 A priori estimates

We now give a priori estimates satisfied by the solution values c_D^n , $D \in \mathcal{D}_h$, $n \in \{0, 1, \dots, N\}$.

Theorem 1.5.1. (A priori estimates) *The solution of the combined scheme (1.9a)–(1.9c) satisfies*

$$c_\beta \max_{n \in \{1, 2, \dots, N\}} \sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D| \leq C_{\text{ae}}, \quad (1.26)$$

$$\max_{n \in \{1, 2, \dots, N\}} \sum_{D \in \mathcal{D}_h} [\beta(c_D^n)]^2 |D| \leq C_{\text{ae}\beta}, \quad (1.27)$$

$$c_S \sum_{n=1}^N \Delta t_n \|c_h^n\|_{X_h}^2 \leq C_{\text{ae}} \quad (1.28)$$

with $c_h^n = \sum_{D \in \mathcal{D}_h} c_D^n \varphi_D$,

$$C_{\text{ae}} := 4M\beta(M)|\Omega| + \frac{4T}{c_\beta} \|q\|_{0, Q_T}^2 + 4L_F M^2 T |\Omega|,$$

$$C_{\text{ae}\beta} := [\beta(M)]^2 |\Omega|$$

when Assumption (D) is satisfied and only Assumption (A1) holds and

$$C_{\text{ae}\beta} := (2C_\beta^2 + 4L_\beta^2 P^2) |\Omega| + \frac{4L_\beta^2}{c_\beta} C_{\text{ae}}$$

when Assumption (D) is not satisfied but Assumption (A2) holds.

PROOF:

The estimates (1.26) and (1.28) follow immediately from (1.22) for $\varepsilon = c_\beta/(2T)$, since for $u = 1$ the extended scheme (1.19a)–(1.19c) completely coincides with the scheme (1.9a)–(1.9c). To see the boundedness of the term on the left-hand side of (1.27) under Assumption (D) is immediate, using the discrete maximum principle stated by Theorem 1.4.11. In this case Assumption (A1) suffices. In the general case one has to use Assumption (A2) to show $[\beta(s)]^2 \leq 2C_\beta^2 + 4L_\beta^2 P^2 + 4L_\beta^2 s^2$, see Lemma 1.8.3 in Appendix 1.8. Hence,

$$\sum_{D \in \mathcal{D}_h} [\beta(c_D^n)]^2 |D| \leq (2C_\beta^2 + 4L_\beta^2 P^2) |\Omega| + 4L_\beta^2 \sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D| \quad \forall n \in \{1, 2, \dots, N\}. \quad \square$$

Remark 1.5.2. (Discrete Friedrichs inequality) *In the proof of Theorem 1.5.1, as well as throughout the whole chapter, we do not make use of the discrete Friedrichs inequality*

$$\|c_h\|_{0, \Omega} \leq C_P \|c_h\|_{X_h} \quad \forall c_h \in X_h^0, C_P > 0.$$

This is possible due to the growth condition imposed on β in Assumption (A1). However, to prove the convergence of the scheme for an elliptic–parabolic problem or in the stationary case, when Assumption (D) is not satisfied and therefore the discrete maximum principle stated by Theorem 1.4.11 is not valid, the discrete Friedrichs inequality would be necessary. We then refer to [84], [28], or Chapter 2 of this thesis for the proof of this inequality.

Using the values c_D^n , $D \in \mathcal{D}_h$, $n \in \{0, 1, \dots, N\}$, we now define two approximate solutions.

Definition 1.5.3. (Approximate solutions) *Let the values c_D^n , $D \in \mathcal{D}_h$, $n \in \{0, 1, \dots, N\}$, be the solutions to (1.9a)–(1.9c). As the approximate solutions of the problem (1.1)–(1.3) by means of the combined finite volume–nonconforming/mixed-hybrid finite element scheme, we understand:*

(i) A function $c_{h,\Delta t}$ such that

$$\begin{aligned} c_{h,\Delta t}(\mathbf{x}, 0) &= c_h^0(\mathbf{x}) \text{ for } \mathbf{x} \in \Omega, \\ c_{h,\Delta t}(\mathbf{x}, t) &= c_h^n(\mathbf{x}) \text{ for } \mathbf{x} \in \Omega, t \in (t_{n-1}, t_n] \quad n \in \{1, \dots, N\}, \end{aligned} \quad (1.29)$$

where $c_h^n = \sum_{D \in \mathcal{D}_h} c_D^n \varphi_D$;

(ii) A function $\tilde{c}_{h,\Delta t}$ such that

$$\begin{aligned} \tilde{c}_{h,\Delta t}(\mathbf{x}, 0) &= c_D^0 \text{ for } \mathbf{x} \in D^\circ, D \in \mathcal{D}_h, \\ \tilde{c}_{h,\Delta t}(\mathbf{x}, t) &= c_D^n \text{ for } \mathbf{x} \in D^\circ, D \in \mathcal{D}_h, t \in (t_{n-1}, t_n] \quad n \in \{1, \dots, N\}. \end{aligned} \quad (1.30)$$

The function $c_{h,\Delta t}$ is piecewise linear and continuous in the barycentres of the interior sides in space and piecewise constant in time; we will call it a *nonconforming finite element solution*. The function $\tilde{c}_{h,\Delta t}$ is given by the values of $c_{h,\Delta t}$ in side barycentres and is piecewise constant on the dual volumes in space and piecewise constant in time; we will call it a *finite volume solution*. The following important relation between $c_{h,\Delta t}$ and $\tilde{c}_{h,\Delta t}$ is valid:

Lemma 1.5.4. *There holds*

$$\|c_{h,\Delta t} - \tilde{c}_{h,\Delta t}\|_{0,Q_T} \longrightarrow 0 \text{ as } h \rightarrow 0.$$

PROOF:

We have

$$\begin{aligned} \|c_{h,\Delta t} - \tilde{c}_{h,\Delta t}\|_{0,Q_T}^2 &= \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \sum_{\sigma_D \in \mathcal{E}_K} \int_{K \cap D} |c_{h,\Delta t}(\mathbf{x}, t_n) - \tilde{c}_{h,\Delta t}(\mathbf{x}, t_n)|^2 d\mathbf{x} \\ &= \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \sum_{\sigma_D \in \mathcal{E}_K} \int_{K \cap D} |c_{h,\Delta t}(\mathbf{x}, t_n) - c_{h,\Delta t}(Q_D, t_n)|^2 d\mathbf{x} \\ &= \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \sum_{\sigma_D \in \mathcal{E}_K} \int_{K \cap D} |\nabla c_{h,\Delta t}(\mathbf{x}, t_n) \cdot (\mathbf{x} - Q_D)|^2 d\mathbf{x} \\ &\leq \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \sum_{\sigma_D \in \mathcal{E}_K} \left| \nabla c_h^n|_K \right|^2 [\text{diam}(D)]^2 |K \cap D| \\ &\leq h^2 \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \left| \nabla c_h^n|_K \right|^2 |K| \leq h^2 \sum_{n=1}^N \Delta t_n \|c_h^n\|_{X_h}^2 \leq h^2 \frac{C_{ae}}{c_S}, \end{aligned}$$

where we have used the definitions of $c_{h,\Delta t}$ and $\tilde{c}_{h,\Delta t}$ and the a priori estimate (1.28). \square

Remark 1.5.5. (Interpretation of the values c_D^n) *We remark that the approximate solutions $c_{h,\Delta t}$, $\tilde{c}_{h,\Delta t}$ are only an interpretation of the values c_D^n , $D \in \mathcal{D}_h$, $n \in \{0, 1, \dots, N\}$. In particular, we may work with $\tilde{c}_{h,\Delta t}$ as in the finite volume method and then use Lemma 1.5.4 to extend the convergence results also to $c_{h,\Delta t}$.*

1.5.2 Estimates on differences of time and space translates

Estimates on differences of time and space translates have been used in [63, 64] to prove the relative compactness property of the sequence of approximate solutions. We give below the time translate estimate for $\tilde{c}_{h,\Delta t}$ given by (1.30). We extend the techniques from [63, 64] to the case of negative transmissibilities (which in particular implies that the discrete maximum principle is not satisfied) and to a nonconstant time step.

Lemma 1.5.6. (Time translate estimate) *There exists a constant $C_{tt} > 0$ such that*

$$\int_0^{T-\tau} \int_{\Omega} \left(\tilde{c}_{h,\Delta t}(\mathbf{x}, t + \tau) - \tilde{c}_{h,\Delta t}(\mathbf{x}, t) \right)^2 dx dt \leq C_{tt}(\tau + \Delta t)$$

for all $\tau \in (0, T)$.

PROOF:

We set

$$T_T := \int_0^{T-\tau} \int_{\Omega} \left(\tilde{c}_{h,\Delta t}(\mathbf{x}, t + \tau) - \tilde{c}_{h,\Delta t}(\mathbf{x}, t) \right)^2 dx dt.$$

Using the definition of $\tilde{c}_{h,\Delta t}$ given by (1.30), we can rewrite T_T as

$$T_T = \int_0^{T-\tau} \sum_{D \in \mathcal{D}_h} |D| \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right)^2 dt,$$

where

$$\begin{aligned} n_1(t) &\in \{1, 2, \dots, N\} \text{ is such that } t_{n_1-1} < t + \tau \leq t_{n_1}, \\ n_2(t) &\in \{1, 2, \dots, N\} \text{ is such that } t_{n_2-1} < t \leq t_{n_2}. \end{aligned}$$

We now use (1.9b) and the growth condition imposed on β in Assumption (A1) and estimate

$$\begin{aligned} T_T &\leq \frac{1}{c_\beta} \int_0^{T-\tau} \sum_{D \in \mathcal{D}_h^{\text{int}}} |D| \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right) \left(\beta(c_D^{n_1(t)}) - \beta(c_D^{n_2(t)}) \right) dt \\ &= \frac{1}{c_\beta} \int_0^{T-\tau} \sum_{D \in \mathcal{D}_h^{\text{int}}} |D| \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right) \sum_{n=1}^N \chi(n, t) \left(\beta(c_D^n) - \beta(c_D^{n-1}) \right) dt, \end{aligned}$$

where the function $\chi(n, t)$ is defined as

$$\chi(n, t) := \begin{cases} 1 & \text{if } t \leq t_{n-1} < t + \tau \\ 0 & \text{otherwise} \end{cases}.$$

In view of the definition (1.9a)–(1.9c) of the combined scheme and of Corollary 1.4.2, we have

$$\begin{aligned} T_T &\leq \frac{1}{c_\beta} \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{D \in \mathcal{D}_h^{\text{int}}} \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right) \left(\sum_{E \in \mathcal{N}(D)} \mathbb{S}_{D,E}^n (c_E^n - c_D^n) \right. \\ &\quad \left. - \mu \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{c_{D,E}^n} - F(c_D^n) |D| + q_D^n |D| \right) dt. \end{aligned} \quad (1.31)$$

We now estimate each term separately.

Diffusion term

We set

$$T_D := \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{D \in \mathcal{D}_h} \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right) \sum_{E \in \mathcal{N}(D)} \mathbb{S}_{D,E}^n (c_E^n - c_D^n) dt,$$

where we have changed the summation over $D \in \mathcal{D}_h^{\text{int}}$ into the summation over $D \in \mathcal{D}_h$ using (1.9b). This enables us to rewrite T_D as a summation over interior dual sides, since each $\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}$ is in the original sum just twice. This gives

$$\begin{aligned} T_D &= \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \mathbb{S}_{D,E}^n \left[(c_E^n - c_D^n) \left(c_D^{n_1(t)} - c_E^{n_1(t)} \right) \right. \\ &\quad \left. + (c_E^n - c_D^n) \left(c_E^{n_2(t)} - c_D^{n_2(t)} \right) \right] dt. \end{aligned}$$

Using the inequality $cab \leq |c|a^2/2 + |c|b^2/2$ and the estimate (1.18) on $|\mathbb{S}_{D,E}^n|$, we can write

$$T_D \leq T_{D_1} + T_{D_2} + T_{D_3}$$

with

$$\begin{aligned} T_{D_1} &:= \frac{C_S}{\kappa_{\mathcal{T}}} \frac{1}{(d-1)^2} \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \text{diam}(K_{D,E})^{d-2} (c_E^n - c_D^n)^2 dt, \\ T_{D_2} &:= \frac{C_S}{2\kappa_{\mathcal{T}}} \frac{1}{(d-1)^2} \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \text{diam}(K_{D,E})^{d-2} \left(c_E^{n_1(t)} - c_D^{n_1(t)} \right)^2 dt, \\ T_{D_3} &:= \frac{C_S}{2\kappa_{\mathcal{T}}} \frac{1}{(d-1)^2} \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \text{diam}(K_{D,E})^{d-2} \left(c_E^{n_2(t)} - c_D^{n_2(t)} \right)^2 dt. \end{aligned}$$

We now notice that

$$\int_0^{T-\tau} \chi(n, t) dt \leq \tau, \quad (1.32)$$

since the function $\chi(n, t)$, for fixed n , is nonzero and equal to one just on the interval $(t_{n-1} - \tau, t_{n-1}]$ of length τ . Using this and the a priori estimate (1.28), we have

$$T_{X_1}^* := \sum_{n=1}^N \Delta t_n \|c_h^n\|_{X_h}^2 \int_0^{T-\tau} \chi(n, t) dt \leq \tau \frac{C_{\text{ae}}}{C_S}. \quad (1.33)$$

We now introduce a term $T_{X_3}^*$,

$$T_{X_3}^* := \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \|c_h^{n_2(t)}\|_{X_h}^2 dt$$

and have, using the definition of $n_2(t)$

$$T_{X_3}^* = \sum_{n=1}^N \Delta t_n \sum_{m=1}^N \int_{t_{m-1}}^{t_m} \chi(n, t) \|c_h^{n_2(t)}\|_{X_h}^2 dt = \sum_{m=1}^N \|c_h^m\|_{X_h}^2 \sum_{n=1}^N \Delta t_n \int_{t_{m-1}}^{t_m} \chi(n, t) dt. \quad (1.34)$$

Let us now consider the case where the time step is constant, i.e. $\Delta t_n = \Delta t$ for all $n \in \{1, 2, \dots, N\}$. We then have, using a simple change of variables and the fact that $t_{m-1} - t_{n-1} = t_m - t_n$,

$$\begin{aligned} \sum_{n=1}^N \Delta t_n \int_{t_{m-1}}^{t_m} \chi(n, t) dt &= \sum_{n=1}^N \Delta t \int_{t_{m-1}-t_{n-1}}^{t_m-t_{n-1}} \chi(n, s+t_{n-1}) ds \\ &= \Delta t \sum_{n=1}^N \int_{t_m-t_n}^{t_m-t_{n-1}} 1_{-\tau < s \leq 0} ds \leq \tau \Delta t, \end{aligned}$$

where the function $1_{a < x \leq b}$ is equal to 1 on the interval $(a, b]$ and zero otherwise, which we substitute back into (1.34) and use the a priori estimate (1.28) to obtain

$$T_{X_3}^* \leq \tau \frac{C_{ae}}{c_S}.$$

Next we consider a nonconstant time step. We have

$$\sum_{n=1}^N \Delta t_n \chi(n, t) \leq \tau + \Delta t,$$

considering that $\chi(n, t)$, for fixed t , is nonzero and equal to one just when $t \leq t_{n-1} < t + \tau$, i.e. an interval of length τ , and that with each such n , we add Δt_n . Using this, we have

$$T_{X_3}^* \leq (\tau + \Delta t) \sum_{m=1}^N \|c_h^m\|_{X_h}^2 \Delta t_m \leq (\tau + \Delta t) \frac{C_{ae}}{c_S}.$$

We next introduce a term $T_{X_2}^*$,

$$T_{X_2}^* := \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \|c_h^{n_1(t)}\|_{X_h}^2 dt.$$

Similarly as in the previous case, using the definition of $n_1(t)$, we have

$$\begin{aligned} T_{X_2}^* &\leq \sum_{n=1}^N \Delta t_n \sum_{m=1}^N \int_{t_{m-1}-\tau}^{t_m-\tau} \chi(n, t) \|c_h^{n_1(t)}\|_{X_h}^2 dt \\ &= \sum_{m=1}^N \|c_h^m\|_{X_h}^2 \sum_{n=1}^N \Delta t_n \int_{t_{m-1}-\tau}^{t_m-\tau} \chi(n, t) dt, \end{aligned}$$

which yields the same estimate for $T_{X_2}^*$ as for $T_{X_3}^*$. We finally introduce

$$T_{L_1}^* := \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D| \int_0^{T-\tau} \chi(n, t) dt \leq \tau T \frac{C_{ae}}{c_\beta}, \quad (1.35)$$

which we have estimated using (1.32) and the a priori estimate (1.26), and

$$T_{L_i}^* := \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{D \in \mathcal{D}_h} \left(c_D^{n_{i-1}(t)} \right)^2 |D| dt \quad i \in \{2, 3\}.$$

We shall need $T_{L_i}^*$, $i \in \{1, 2, 3\}$, for the estimates of the other terms of T_T below. Using the a priori estimate (1.26) and the same techniques as for $T_{X_i}^*$, $i = 2, 3$, we altogether come to

$$T_{X_i}^* \leq \tau \frac{C_{ae}}{c_S}, \quad T_{L_i}^* \leq \tau T \frac{C_{ae}}{c_\beta} \quad i \in \{2, 3\} \quad (1.36)$$

for a constant time step and

$$T_{X_i}^* \leq (\tau + \Delta t) \frac{C_{ae}}{c_S}, \quad T_{L_i}^* \leq (\tau + \Delta t) T \frac{C_{ae}}{c_\beta} \quad i \in \{2, 3\} \quad (1.37)$$

for a generally nonconstant time step. Now using (1.6) for T_{D_1} , T_{D_2} , and T_{D_3} , we have

$$T_D \leq \frac{C_S}{\kappa_T^2} \frac{d+1}{2d(d-1)^2} \left(T_{X_1}^* + \frac{1}{2} T_{X_2}^* + \frac{1}{2} T_{X_3}^* \right). \quad (1.38)$$

Convection term

We will write the convection term as $T_{C_1} + T_{C_2}$, with

$$T_{C_1} := -\mu \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{D \in \mathcal{D}_h} \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right) \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \widehat{c_{D,E}^n} dt$$

and

$$T_{C_2} := -\mu \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{D \in \mathcal{D}_h} \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right) \sum_{E \in \mathcal{N}(D)} |\mathbf{v}_{D,E}^n| \alpha_{D,E}^n (c_E^n - c_D^n) dt,$$

using the splitting into full upstream weighting and coefficient-centered weighting.

We again rewrite T_{C_1} as the summation over the interior dual sides; we however adjust the denotation of the dual volumes sharing a given side $\sigma_{D,E}$ such that $\mathbf{v}_{D,E}^n \geq 0$. Then, using the definition of the upstream weighting, we have

$$T_{C_1} = \mu \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \mathbf{v}_{D,E}^n \geq 0} -\mathbf{v}_{D,E}^n c_D^n \left(c_D^{n_1(t)} - c_E^{n_1(t)} + c_E^{n_2(t)} - c_D^{n_2(t)} \right) dt.$$

Using $\pm ab \leq \varepsilon a^2/2 + b^2/(2\varepsilon)$, $\varepsilon > 0$, where we put $\varepsilon = d_{D,E}$, we come to

$$T_{C_1} \leq T_{C_3} + T_{C_4} + T_{C_5}$$

with

$$\begin{aligned} T_{C_3} &:= \mu \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \mathbf{v}_{D,E}^n \geq 0} |\mathbf{v}_{D,E}^n| d_{D,E} (c_D^n)^2 dt, \\ T_{C_4} &:= \frac{\mu}{2} \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\mathbf{v}_{D,E}^n|}{d_{D,E}} \left(c_E^{n_1(t)} - c_D^{n_1(t)} \right)^2 dt, \\ T_{C_5} &:= \frac{\mu}{2} \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\mathbf{v}_{D,E}^n|}{d_{D,E}} \left(c_E^{n_2(t)} - c_D^{n_2(t)} \right)^2 dt. \end{aligned}$$

We have

$$\begin{aligned} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \mathbf{v}_{D,E}^n \geq 0} |\mathbf{v}_{D,E}^n| d_{D,E} (c_D^n)^2 &\leq C_{\mathbf{v}} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \mathbf{v}_{D,E}^n \geq 0} \frac{|K_{D,E}|}{\kappa_{\mathcal{T}} d (d-1)} (c_D^n)^2 \\ &\leq \frac{C_{\mathbf{v}}}{\kappa_{\mathcal{T}}} \frac{d+1}{d-1} \sum_{D \in \mathcal{D}_h} \left(\frac{|K_D|}{d+1} + \frac{|L_D|}{d+1} \right) (c_D^n)^2 = \frac{C_{\mathbf{v}}}{\kappa_{\mathcal{T}}} \frac{d+1}{d-1} \sum_{D \in \mathcal{D}_h} (c_D^n)^2 |D|, \end{aligned}$$

where we have used Assumption (A4), which implies $|\mathbf{v}_{D,E}^n| \leq C_{\mathbf{v}} |\sigma_{D,E}|$, (1.8), Assumption (B), (1.9b), the fact that each dual volume $D \in \mathcal{D}_h^{\text{int}}$ has d dual sides inside a simplex K_D and d dual sides inside a simplex L_D and that c_D^n can appear as an upwind value only at these sides, and (1.5). Thus, we have

$$T_{C_3} \leq \mu \frac{C_{\mathbf{v}}}{\kappa_{\mathcal{T}}} \frac{d+1}{d-1} T_{L_1}^*.$$

Using $|\mathbf{v}_{D,E}^n| \leq C_{\mathbf{v}} |\sigma_{D,E}|$ and (1.7), we have

$$T_{C_i} \leq \mu \frac{C_{\mathbf{v}}}{\kappa_{\mathcal{T}}} \frac{d+1}{4(d-1)} T_{X_{i-2}}^* \quad i \in \{4, 5\},$$

which altogether leads to

$$T_{C_1} \leq \mu \frac{C_{\mathbf{v}}}{\kappa_{\mathcal{T}}} \left(\frac{d+1}{d-1} T_{L_1}^* + \frac{d+1}{4(d-1)} (T_{X_2}^* + T_{X_3}^*) \right). \quad (1.39)$$

We now consider T_{C_2} . We can easily notice that it is almost same as the diffusion term T_D , except for $-\mu$ and the term $\mathbb{S}_{D,E}^n$, which is replaced by $|\mathbf{v}_{D,E}^n| \alpha_{D,E}^n$. Using $|\mathbf{v}_{D,E}^n| \leq C_{\mathbf{v}} |\sigma_{D,E}|$, $\alpha_{D,E}^n \leq 1/2$, and the estimates (1.6) and (1.8), we easily come to

$$T_{C_2} \leq \mu \frac{C_{\mathbf{v}}}{\kappa_{\mathcal{T}}} h \frac{d+1}{4d(d-1)} \left(T_{X_1}^* + \frac{1}{2} T_{X_2}^* + \frac{1}{2} T_{X_3}^* \right). \quad (1.40)$$

Reaction term

We denote

$$T_R := - \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{D \in \mathcal{D}_h} \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right) F(c_D^n) |D| dt.$$

We estimate

$$-F(c_D^n) (c_D^{n_1} - c_D^{n_2}) \leq \frac{(c_D^{n_1} - c_D^{n_2})^2}{2} + \frac{(F(c_D^n))^2}{2} \leq (c_D^{n_1})^2 + (c_D^{n_2})^2 + \frac{L_F^2 (c_D^n)^2}{2},$$

using the inequalities $ab \leq a^2/2 + b^2/2$, $(a-b)^2/2 \leq a^2 + b^2$, the Lipschitz continuity of F with the constant L_F , and the fact that $F(0) = 0$, following either from Assumption (A6) or (A7). This implies

$$T_R \leq \left(\frac{L_F^2}{2} T_{L_1}^* + T_{L_2}^* + T_{L_3}^* \right). \quad (1.41)$$

Sources term

We denote

$$T_S := \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{D \in \mathcal{D}_h} \left(c_D^{n_1(t)} - c_D^{n_2(t)} \right) q_D^n |D| dt.$$

Using the same estimate as for the reaction term, (1.32), and (1.21), we come to

$$T_S \leq \frac{1}{2} \tau \|q\|_{0, Q_T}^2 + T_{L_2}^* + T_{L_3}^*. \quad (1.42)$$

The assertion of the lemma follows by introducing (1.38), (1.39), (1.40), (1.41), and (1.42) into (1.31), while using the estimates (1.33), (1.35), and (1.37). \square

Remark 1.5.7. (Time translate estimate under Assumption (D)) *If Assumption (D) is valid, the transmissibilities $\mathbb{S}_{D,E}^n$ are non-negative as in the finite volume method. Hence $T_D \leq T_{D_1} + T_{D_2} + T_{D_3}$ with*

$$T_{D_1} = \sum_{n=1}^N \Delta t_n \int_0^{T-\tau} \chi(n, t) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \mathbb{S}_{D,E}^n (c_E^n - c_D^n)^2 dt$$

and similarly for T_{D_2} and T_{D_3} . Thus using

$$\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \mathbb{S}_{D,E}^n (c_E^n - c_D^n)^2 = - \sum_{D \in \mathcal{D}_h} c_D \sum_{E \in \mathcal{D}_h} \mathbb{S}_{D,E}^n c_E$$

and (1.17), $T_D \leq C_{\mathbf{S}}(T_{X_1}^* + T_{X_2}^*/2 + T_{X_3}^*/2)$ in this case instead of (1.38).

Remark 1.5.8. (Time translate estimate for a constant time step) *For a constant time step, we have indeed an $O(\tau)$ estimate, using (1.36) instead of (1.37).*

We now give the space translate estimate for $\tilde{c}_{h,\Delta t}$ given by (1.30). Lemma 1.5.9 extends the space translate estimate from [63, 64] to the case of (dual) meshes not necessarily satisfying the orthogonality property.

Lemma 1.5.9. (Space translate estimate) *Let us define $\tilde{c}_{h,\Delta t}(\mathbf{x}, t)$ by zero outside of Ω . Then there exists a constant $C_{\text{st}} > 0$ such that*

$$\int_0^T \int_{\Omega} \left(\tilde{c}_{h,\Delta t}(\mathbf{x} + \boldsymbol{\xi}, t) - \tilde{c}_{h,\Delta t}(\mathbf{x}, t) \right)^2 d\mathbf{x} dt \leq C_{\text{st}} |\boldsymbol{\xi}| (|\boldsymbol{\xi}| + h)$$

for all $\boldsymbol{\xi} \in \mathbb{R}^d$.

PROOF:

We define a function $\chi_{\sigma}(\mathbf{x})$ for each $\sigma \in \mathcal{F}_h^{\text{int}}$ by

$$\chi_{\sigma}(\mathbf{x}) := \begin{cases} 1 & \text{if } \sigma \cap [\mathbf{x}, \mathbf{x} + \boldsymbol{\xi}] \neq \emptyset \\ 0 & \text{if } \sigma \cap [\mathbf{x}, \mathbf{x} + \boldsymbol{\xi}] = \emptyset \end{cases}.$$

A simple geometrical consideration leads to

$$|\tilde{c}_{h,\Delta t}(\mathbf{x} + \boldsymbol{\xi}, t) - \tilde{c}_{h,\Delta t}(\mathbf{x}, t)| \leq \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E^n - c_D^n| \chi_{\sigma_{D,E}}(\mathbf{x})$$

for a.e. $\mathbf{x} \in \Omega$ and for $t \in (t_{n-1}, t_n]$, considering that $\tilde{c}_{h,\Delta t}$ is piecewise constant on \mathcal{D}_h , the Dirichlet boundary condition (1.9b), and the fact that $\tilde{c}_{h,\Delta t}(\mathbf{x}, t)$ is defined by zero outside of Ω . The above inequality is in particular not valid for $\mathbf{x} \in \Omega$ such that the segment $[\mathbf{x}, \mathbf{x} + \boldsymbol{\xi}]$ intersects some vertex of the dual mesh. The Cauchy–Schwarz inequality yields

$$\begin{aligned} & \left(\tilde{c}_{h,\Delta t}(\mathbf{x} + \boldsymbol{\xi}, t) - \tilde{c}_{h,\Delta t}(\mathbf{x}, t) \right)^2 \\ & \leq \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \chi_{\sigma_{D,E}}(\mathbf{x}) \text{diam}(K_{D,E}) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{(c_E^n - c_D^n)^2}{\text{diam}(K_{D,E})} \chi_{\sigma_{D,E}}(\mathbf{x}) \end{aligned} \quad (1.43)$$

for a.e. $\mathbf{x} \in \Omega$ and for $t \in (t_{n-1}, t_n]$.

The proof of Lemma 2.3.4 from Chapter 2 of this thesis gives

$$\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \chi_{\sigma_{D,E}}(\mathbf{x}) \text{diam}(K_{D,E}) \leq 4(d-1)P\theta_T^{2P}(|\boldsymbol{\xi}| + h), \quad (1.44)$$

where $P = 2^{d-1}\pi/\phi_T$. Here, the constant $\phi_T > 0$ is such that

$$\min_{K \in \mathcal{T}_h} \phi_K \geq \phi_T \quad \forall h > 0, \quad (1.45)$$

where ϕ_K is the smallest angle of the simplex K (plain angle in radians if $d = 2$ and spheric angle in steradians if $d = 3$). Notice that Assumption (B) is equivalent to (1.45). We finally integrate (1.43) over Q_T . This gives

$$\begin{aligned} & \int_0^T \int_{\Omega} \left(\tilde{c}_{h,\Delta t}(\mathbf{x} + \boldsymbol{\xi}, t) - \tilde{c}_{h,\Delta t}(\mathbf{x}, t) \right)^2 \text{d}\mathbf{x} \text{d}t \\ & \leq 4 \frac{(d-1)}{d} P\theta_T^{2P} (|\boldsymbol{\xi}| + h) \sum_{n=1}^N \Delta t_n \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{(c_E^n - c_D^n)^2}{d_{D,E}} \int_{\Omega} \chi_{\sigma_{D,E}}(\mathbf{x}) \text{d}\mathbf{x}, \end{aligned}$$

using (1.44) and (1.8). Finally, the value $\int_{\Omega} \chi_{\sigma_{D,E}}(\mathbf{x}) \text{d}\mathbf{x}$ is the measure of the set of points of Ω which are located inside a cylinder whose basis is $\sigma_{D,E}$ and generator vector is $-\boldsymbol{\xi}$. Thus

$$\int_{\Omega} \chi_{\sigma_{D,E}}(\mathbf{x}) \text{d}\mathbf{x} \leq |\sigma_{D,E}| |\boldsymbol{\xi}|.$$

Using (1.7) and the a priori estimate (1.28), this yields the assertion of the lemma with

$$C_{\text{st}} := 2 \frac{(d+1)}{d\kappa_T} P\theta_T^{2P} \frac{C_{\text{ae}}}{c_{\mathbf{S}}}. \quad \square$$

Remark 1.5.10. (Constant in the space translate estimate) *The constant C_{st} in the space translate estimate has the form $C_{\text{st}} = C(d, \kappa_T) C_{\text{ae}}/c_{\mathbf{S}}$. We recall that for the finite volume mesh satisfying the orthogonality property, this constant equals to $C_{\text{ae}}/c_{\mathbf{S}}$, cf. [63, 64]. Hence supposing general unstructured meshes only satisfying the shape regularity Assumption (B) leads to the multiplication by a factor dependent on d and κ_T . It can be shown (see Chapter 2) that this factor is of the form $C(d)/\kappa_T^2$ for $\{\mathcal{T}_h\}_h$ satisfying the inverse assumption (local refinement not allowed).*

1.6 Convergence

Using the a priori estimates of the previous section and the Kolmogorov relative compactness theorem, we show in this section a strong $L^2(Q_T)$ convergence of the approximate solutions to a function u which we prove to be a weak solution of the continuous problem.

1.6.1 Strong convergence in $L^2(Q_T)$

Theorem 1.6.1. (Strong convergence in $L^2(Q_T)$) *There exist subsequences of $\tilde{c}_{h,\Delta t}$ and $c_{h,\Delta t}$ which converge strongly in $L^2(Q_T)$ to some function $u \in L^2(0, T; H_0^1(\Omega))$.*

PROOF:

Let us consider the sequence $\tilde{c}_{h,\Delta t}$. The a priori estimate (1.26) and Lemmas 1.5.6 and 1.5.9 imply that $\tilde{c}_{h,\Delta t}$ satisfies the assumptions of Lemma 1.8.4 in Appendix 1.8. Thus $\tilde{c}_{h,\Delta t}$ verifies the assumptions of the Kolmogorov theorem ([29, Theorem IV.25], [61, Theorem 3.9]) and consequently is relatively compact in $L^2(Q_T)$. This implies the existence of a subsequence of $\tilde{c}_{h,\Delta t}$ which converges strongly to some function u in $L^2(Q_T)$. Moreover, due to the space translate estimate of Lemma 1.5.9, [61, Theorem 3.10] gives that $u \in L^2(0, T; H_0^1(\Omega))$. Finally, considering Lemma 1.5.4, $c_{h,\Delta t}$ converges to the same u . \square

Remark 1.6.2. (Relative compactness for a constant time step) *In view of Remark 1.5.8, the a priori estimate (1.26) and Lemmas 1.5.6 and 1.5.9 directly imply that $\tilde{c}_{h,\Delta t}$ verifies the assumptions of the Kolmogorov theorem for a constant time step. Hence, in this case Lemma 1.8.4 is not necessary.*

1.6.2 Convergence to a weak solution

We have shown in Theorem 1.6.1 that subsequences of $\tilde{c}_{h,\Delta t}$ and $c_{h,\Delta t}$, which we still denote by $\tilde{c}_{h,\Delta t}$ and $c_{h,\Delta t}$, converge strongly in $L^2(Q_T)$ to some function $u \in L^2(0, T; H_0^1(\Omega))$. We now show that u is a weak solution of the continuous problem. For this purpose, we introduce

$$\Psi := \{ \psi \in C^{2,1}(\bar{\Omega} \times [0, T]), \psi = 0 \text{ on } \partial\Omega \times [0, T], \psi(\cdot, T) = 0 \}. \quad (1.46)$$

We then take an arbitrary $\psi \in \Psi$, multiply (1.9c) by $\Delta t_n \psi(Q_D, t_{n-1})$, and sum the result over $D \in \mathcal{D}_h^{\text{int}}$ and $n = 1, \dots, N$. This gives

$$T_T + T_D + T_C + T_R = T_S \quad (1.47)$$

with

$$\begin{aligned} T_T &:= \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \left(\beta(c_D^n) - \beta(c_D^{n-1}) \right) \psi(Q_D, t_{n-1}) |D|, \\ T_D &:= \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} \sum_{E \in \mathcal{D}_h} c_E^n \sum_{K \in \mathcal{T}_h} (\mathbf{S}^n \nabla \varphi_E, \nabla \varphi_D)_{0,K} \psi(Q_D, t_{n-1}), \\ T_C &:= \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{c_{D,E}^n} \psi(Q_D, t_{n-1}), \\ T_R &:= \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} F(c_D^n) \psi(Q_D, t_{n-1}) |D|, \\ T_S &:= \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} q_D^n \psi(Q_D, t_{n-1}) |D|, \end{aligned}$$

using $\psi(Q_D, t_{n-1}) = 0$ for all $D \in \mathcal{D}_h^{\text{ext}}$ and $n = 1, \dots, N$. We now show that each of the above terms converges to its continuous version as h and Δt tend to zero.

Time evolution term

We use the discrete integration by parts formula and the fact that $\psi(Q_D, t_N) = 0$ for all $D \in \mathcal{D}_h$ to obtain

$$T_T = - \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \beta(c_D^n) \left(\psi(Q_D, t_n) - \psi(Q_D, t_{n-1}) \right) |D| - \sum_{D \in \mathcal{D}_h} \beta(c_D^0) \psi(Q_D, 0) |D|. \quad (1.48)$$

We would now like to show that

$$\sum_{D \in \mathcal{D}_h} \beta(c_D^0) \psi(Q_D, 0) |D| \longrightarrow \int_{\Omega} \beta(c_0(\mathbf{x})) \psi(\mathbf{x}, 0) \, d\mathbf{x} \text{ as } h \rightarrow 0. \quad (1.49)$$

For this purpose, we introduce

$$T_{T_1} := \sum_{D \in \mathcal{D}_h} \int_D \left(\beta(c_D^0) \psi(Q_D, 0) - \beta(c_0(\mathbf{x})) \psi(\mathbf{x}, 0) \right) \, d\mathbf{x}.$$

We add and subtract $\beta(c_D^0) \psi(\mathbf{x}, 0)$ to each term and have

$$T_{T_1} = \sum_{D \in \mathcal{D}_h} \int_D \beta(c_D^0) \left(\psi(Q_D, 0) - \psi(\mathbf{x}, 0) \right) \, d\mathbf{x} + \sum_{D \in \mathcal{D}_h} \int_D \left(\beta(c_D^0) - \beta(c_0(\mathbf{x})) \right) \psi(\mathbf{x}, 0) \, d\mathbf{x}.$$

Using the definition of c_D^0 given by (1.9a) for $D \in \mathcal{D}_h^{\text{int}}$ and by (1.9b) for $D \in \mathcal{D}_h^{\text{ext}}$, the fact that β is increasing by Assumption (A1), and Assumption (A9), we have that $|\beta(c_D^0)| \leq \beta(M)$ for all $D \in \mathcal{D}_h$. Due to the boundedness of $|\psi|$ by $C_{1,\psi}$, we come to

$$|T_{T_1}| \leq \beta(M) \sum_{D \in \mathcal{D}_h} \int_D |\psi(Q_D, 0) - \psi(\mathbf{x}, 0)| \, d\mathbf{x} + C_{1,\psi} \sum_{D \in \mathcal{D}_h} \int_D |\beta(c_D^0) - \beta(c_0(\mathbf{x}))| \, d\mathbf{x}. \quad (1.50)$$

Since $\psi \in C^{2,1}(\bar{\Omega} \times [0, T])$, we have

$$|\psi(Q_D, 0) - \psi(\mathbf{x}, 0)| \leq C_{2,\psi} |Q_D - \mathbf{x}| \leq C_{2,\psi} h$$

for all $\mathbf{x} \in D$, and thus the first term of (1.50) tends to 0 as $h \rightarrow 0$. We now consider the second term of (1.50). We have, for boundary dual volumes,

$$\sum_{D \in \mathcal{D}_h^{\text{ext}}} \int_D |c_D^0 - c_0(\mathbf{x})| \, d\mathbf{x} \leq M \sum_{D \in \mathcal{D}_h^{\text{ext}}} |D| \leq M |\partial\Omega| h,$$

using (1.9b) and Assumption (A9). Considering in addition the definition of $\tilde{c}_{h,\Delta t}$ by (1.30) and (1.9a) for interior dual volumes, $\tilde{c}_{h,\Delta t}(\mathbf{x}, 0)$ converges to $c_0(\mathbf{x})$ in Ω in the L^1 sense as $h \rightarrow 0$. Hence at least a subsequence of $\tilde{c}_{h,\Delta t}(\mathbf{x}, 0)$, which we still denote by $\tilde{c}_{h,\Delta t}(\mathbf{x}, 0)$, converges to $c_0(\mathbf{x})$ pointwise a.e. in Ω . Thus also $\beta(\tilde{c}_{h,\Delta t}(\mathbf{x}, 0)) \rightarrow \beta(c_0(\mathbf{x}))$ a.e. in Ω , using the continuity of β . Further, using that β is increasing from Assumption (A1), we have $|\beta(\tilde{c}_{h,\Delta t}(\mathbf{x}, 0))| \leq \beta(M)$. Hence the Lebesgue dominated convergence theorem implies

$$\sum_{D \in \mathcal{D}_h} \int_D |\beta(c_D^0) - \beta(c_0(\mathbf{x}))| \, d\mathbf{x} = \int_{\Omega} |\beta(\tilde{c}_{h,\Delta t}(\mathbf{x}, 0)) - \beta(c_0(\mathbf{x}))| \, d\mathbf{x} \longrightarrow 0 \text{ as } h \rightarrow 0,$$

which can be by repetition extended onto whole $\tilde{c}_{h,\Delta t}(\mathbf{x}, 0)$. Thus $T_{T_1} \rightarrow 0$ as $h \rightarrow 0$ and consequently (1.49) is fulfilled.

Now we intend to prove that

$$\sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \beta(c_D^n) (\psi(Q_D, t_n) - \psi(Q_D, t_{n-1})) |D| \longrightarrow \int_0^T \int_{\Omega} \beta(u(\mathbf{x}, t)) \psi_t(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad (1.51)$$

as $h, \Delta t \rightarrow 0$. We set

$$T_{T_2} := \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \left[\beta(c_D^n) (\psi(Q_D, t_n) - \psi(Q_D, t_{n-1})) |D| - \int_{t_{n-1}}^{t_n} \int_D \beta(u(\mathbf{x}, t)) \psi_t(\mathbf{x}, t) \, d\mathbf{x} \, dt \right].$$

We add and subtract $\int_{t_{n-1}}^{t_n} \int_D \beta(c_D^n) \psi_t(\mathbf{x}, t) \, d\mathbf{x} \, dt$ in each term of T_{T_2} to obtain

$$\begin{aligned} T_{T_2} &= \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \beta(c_D^n) \int_{t_{n-1}}^{t_n} \int_D \left(\frac{\partial \psi}{\partial t}(Q_D, t) - \frac{\partial \psi}{\partial t}(\mathbf{x}, t) \right) \, d\mathbf{x} \, dt \\ &\quad + \int_0^T \int_{\Omega} \left(\beta(\tilde{c}_{h, \Delta t}(\mathbf{x}, t)) - \beta(u(\mathbf{x}, t)) \right) \psi_t(\mathbf{x}, t) \, d\mathbf{x} \, dt. \end{aligned} \quad (1.52)$$

We have, for all $\mathbf{x} \in D$, for all $D \in \mathcal{D}_h$, and all $h > 0$,

$$\left| \frac{\partial \psi}{\partial t}(Q_D, t) - \frac{\partial \psi}{\partial t}(\mathbf{x}, t) \right| \leq f(h),$$

where the function f satisfies $f(h) > 0$ and $f(h) \rightarrow 0$ as $h \rightarrow 0$. This follows by the fact that $\partial \psi / \partial t \in C(\overline{\Omega})$ from (1.46) and hence is uniformly continuous on $\overline{\Omega}$. Thus the first term of (1.52) is bounded by

$$\begin{aligned} f(h) \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} |\beta(c_D^n)| \Delta t_n |D| &\leq f(h) T^{\frac{1}{2}} |\Omega|^{\frac{1}{2}} \left(\sum_{n=1}^N \sum_{D \in \mathcal{D}_h} (\beta(c_D^n))^2 \Delta t_n |D| \right)^{\frac{1}{2}} \\ &\leq f(h) T |\Omega|^{\frac{1}{2}} C_{ae\beta}^{\frac{1}{2}}, \end{aligned}$$

using the Cauchy–Schwarz inequality and the a priori estimate (1.27). Further, $|\psi_t(\mathbf{x}, t)| \leq C_{4, \psi}$, and thus we can estimate T_{T_2} by

$$|T_{T_2}| \leq f(h) T |\Omega|^{\frac{1}{2}} C_{ae\beta}^{\frac{1}{2}} + C_{4, \psi} \int_0^T \int_{\Omega} |\beta(\tilde{c}_{h, \Delta t}(\mathbf{x}, t)) - \beta(u(\mathbf{x}, t))| \, d\mathbf{x} \, dt. \quad (1.53)$$

We now use that $\tilde{c}_{h, \Delta t} \rightarrow u$ strongly in $L^2(Q_T)$ as $h, \Delta t \rightarrow 0$, due to Theorem 1.6.1. There exists at least a subsequence of $\tilde{c}_{h, \Delta t}$, which we still denote $\tilde{c}_{h, \Delta t}$, such that $\tilde{c}_{h, \Delta t}(\mathbf{x}, t) \rightarrow u(\mathbf{x}, t)$ a.e. in Q_T . Thus, using the continuity of $\beta(\cdot)$, $\beta(\tilde{c}_{h, \Delta t}(\mathbf{x}, t)) \rightarrow \beta(u(\mathbf{x}, t))$ a.e. in Q_T . Now under Assumption (D), which implies the discrete maximum principle stated by Theorem 1.4.11, and using that β is increasing, $|\beta(\tilde{c}_{h, \Delta t}(\mathbf{x}, t))| \leq \beta(M)$, and thus we can use the Lebesgue dominated convergence theorem to conclude the convergence of the second term of (1.53) and thus of (1.53) to 0 as $h, \Delta t \rightarrow 0$. In this case Assumption (A1) suffices.

In the general case we use Assumption (A2). We decompose the function β as $\beta_1 + \beta_2$,

$$\begin{aligned} \beta_1(x) &:= \beta(x) \text{ on } [-P, P], \quad \beta_1(x) := 0 \text{ on } (-\infty, -P) \cup (P, +\infty), \\ \beta_2(x) &:= 0 \text{ on } [-P, P], \quad \beta_2(x) := \beta(x) \text{ on } (-\infty, -P) \cup (P, +\infty). \end{aligned}$$

We further introduce a function y linearly connecting the points $[-P, \beta(-P)]$ and $[P, \beta(P)]$ and zero otherwise,

$$y(x) := \frac{\beta(P) - \beta(-P)}{2P}x + \frac{\beta(P) + \beta(-P)}{2} \text{ on } [-P, P], \quad y(x) := 0 \text{ on } (-\infty, -P) \cup (P, +\infty).$$

We finally define $\tilde{\beta}_1 := \beta_1 - y$ and $\tilde{\beta}_2 := \beta_2 + y$ and remark that $\beta = \tilde{\beta}_1 + \tilde{\beta}_2$. Clearly, $\tilde{\beta}_1$ is continuous on \mathbb{R} and satisfies $|\tilde{\beta}_1(x)| \leq 2C_\beta$ on \mathbb{R} and $\tilde{\beta}_2$ is Lipschitz-continuous on \mathbb{R} with $\max\{L_\beta, [\beta(P) - \beta(-P)]/(2P)\}$. We now estimate

$$\begin{aligned} \int_0^T \int_\Omega |\beta(\tilde{c}_{h,\Delta t}(\mathbf{x}, t)) - \beta(u(\mathbf{x}, t))| \, d\mathbf{x} \, dt &\leq \int_0^T \int_\Omega |\tilde{\beta}_1(\tilde{c}_{h,\Delta t}(\mathbf{x}, t)) - \tilde{\beta}_1(u(\mathbf{x}, t))| \, d\mathbf{x} \, dt \\ &\quad + \int_0^T \int_\Omega |\tilde{\beta}_2(\tilde{c}_{h,\Delta t}(\mathbf{x}, t)) - \tilde{\beta}_2(u(\mathbf{x}, t))| \, d\mathbf{x} \, dt. \end{aligned}$$

The first term of the above expression converges to zero using the Lebesgue dominated convergence theorem as in the previous case. For the second term, it suffices to use the Lipschitz continuity of $\tilde{\beta}_2$ and the strong convergence of $\tilde{c}_{h,\Delta t}$ to u in $L^2(Q_T)$. Thus (1.51) is satisfied. Combining (1.49) and (1.51), we have

$$T_T \longrightarrow - \int_0^T \int_\Omega \beta(u(\mathbf{x}, t)) \psi_t(\mathbf{x}, t) \, d\mathbf{x} \, dt - \int_\Omega \beta(c_0(\mathbf{x})) \psi(\mathbf{x}, 0) \, d\mathbf{x} \quad (1.54)$$

as $h, \Delta t \rightarrow 0$.

Diffusion term

We rewrite T_D as

$$T_D = \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \int_K \mathbf{S}^n \nabla c_h^n(\mathbf{x}) \cdot \nabla \left(\sum_{D \in \mathcal{D}_h} \psi(Q_D, t_{n-1}) \varphi_D(\mathbf{x}) \right) \, d\mathbf{x},$$

using the definition of $c_h^n \in X_h$, and define

$$\mathbf{S}_{\Delta t}(\mathbf{x}, t) := \mathbf{S}^n(\mathbf{x}) \text{ for } \mathbf{x} \in \Omega, t \in (t_{n-1}, t_n] \quad n \in \{1, \dots, N\}, \quad (1.55)$$

where \mathbf{S}^n is given by (1.13) for the nonconforming method and by (1.15) for the mixed-hybrid method. We will show the validity of two passages to the limit. We begin by showing that

$$\begin{aligned} &\sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \int_K \mathbf{S}^n \nabla c_h^n(\mathbf{x}) \cdot \nabla \left(\sum_{D \in \mathcal{D}_h} \psi(Q_D, t_{n-1}) \varphi_D(\mathbf{x}) \right) \, d\mathbf{x} \\ &- \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \int_K \mathbf{S}^n \nabla c_h^n(\mathbf{x}) \cdot \nabla \psi(\mathbf{x}, t_{n-1}) \, d\mathbf{x} \longrightarrow 0 \text{ as } h \rightarrow 0. \end{aligned} \quad (1.56)$$

We set

$$I_\psi(\cdot, t_{n-1}) := \sum_{D \in \mathcal{D}_h} \psi(Q_D, t_{n-1}) \varphi_D$$

and

$$T_{D_1} := \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \int_K \mathbf{S}^n \nabla c_h^n(\mathbf{x}) \cdot \nabla \left(I_\psi(\mathbf{x}, t_{n-1}) - \psi(\mathbf{x}, t_{n-1}) \right) \, d\mathbf{x}.$$

We then estimate

$$|T_{D_1}| \leq C_{\mathbf{S}} \sum_{n=1}^N \Delta t_n \|c_h^n\|_{X_h} \|I_\psi(\cdot, t_{n-1}) - \psi(\cdot, t_{n-1})\|_{X_h},$$

using the Cauchy–Schwarz inequality. Next we use the interpolation estimate

$$\begin{aligned} \|I_\psi(\cdot, t_{n-1}) - \psi(\cdot, t_{n-1})\|_{X_h} &= \left(\sum_{K \in \mathcal{T}_h} \int_K \left| \nabla (I_\psi(\cdot, t_{n-1}) - \psi(\cdot, t_{n-1})) \right|^2 d\mathbf{x} \right)^{\frac{1}{2}} \\ &\leq C_I \theta_{\mathcal{T}} h \left(\sum_{K \in \mathcal{T}_h} |\psi(\cdot, t_{n-1})|_{2,K}^2 \right)^{\frac{1}{2}} \leq C_I \theta_{\mathcal{T}} C_{5,\psi} h, \end{aligned}$$

where $\theta_{\mathcal{T}}$ is given by the consequence (1.4) of Assumption (B), C_I does not depend on h (nor on Δt), and $|\cdot|_{2,K}$ denotes the H^2 seminorm, see for instance [41, Theorem 15.3]. Finally, the Cauchy–Schwarz inequality yields

$$|T_{D_1}| \leq C_{\mathbf{S}} C_I \theta_{\mathcal{T}} C_{5,\psi} h \left(\sum_{n=1}^N \Delta t_n \|c_h^n\|_{X_h}^2 \right)^{\frac{1}{2}} \left(\sum_{n=1}^N \Delta t_n \right)^{\frac{1}{2}} = C_{\mathbf{S}} C_I \theta_{\mathcal{T}} C_{5,\psi} T^{\frac{1}{2}} \left(\frac{C_{\text{ae}}}{C_{\mathbf{S}}} \right)^{\frac{1}{2}} h,$$

using the a priori estimate (1.28). Hence (1.56) is fulfilled.

We next show that

$$\sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \int_K \mathbf{S}^n \nabla c_h^n(\mathbf{x}) \cdot \nabla \psi(\mathbf{x}, t_{n-1}) d\mathbf{x} \longrightarrow \int_0^T \int_{\Omega} \mathbf{S} \nabla u(\mathbf{x}, t) \cdot \nabla \psi(\mathbf{x}, t) d\mathbf{x} dt \quad (1.57)$$

as $h, \Delta t \rightarrow 0$. We see that both $c_h^n(\mathbf{x})$ and $\psi(\mathbf{x}, t_{n-1})$ are constant in time, so that we can easily introduce an integral with respect to time into the first term of (1.57). We further add

and subtract $\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \int_{\Omega} \mathbf{S}^n \nabla c_h^n(\mathbf{x}) \nabla \psi(\mathbf{x}, t) d\mathbf{x} dt$ and introduce

$$\begin{aligned} T_{D_2} &:= \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K \in \mathcal{T}_h} \int_K \mathbf{S}^n \nabla c_h^n(\mathbf{x}) \cdot \left(\nabla \psi(\mathbf{x}, t_{n-1}) - \nabla \psi(\mathbf{x}, t) \right) d\mathbf{x} dt, \\ T_{D_3} &:= \int_0^T \sum_{K \in \mathcal{T}_h} \int_K \mathbf{S}_{\Delta t} \nabla c_{h,\Delta t}(\mathbf{x}, t) \cdot \nabla \psi(\mathbf{x}, t) d\mathbf{x} dt - \int_0^T \int_{\Omega} \mathbf{S} \nabla u(\mathbf{x}, t) \cdot \nabla \psi(\mathbf{x}, t) d\mathbf{x} dt, \end{aligned}$$

where $c_{h,\Delta t}$ is given by (1.29). Clearly, (1.57) is valid when T_{D_2} and T_{D_3} tend to zero as $h, \Delta t \rightarrow 0$. We first estimate T_{D_2} . We have, for $t \in (t_{n-1}, t_n]$,

$$|\nabla \psi(\mathbf{x}, t_{n-1}) - \nabla \psi(\mathbf{x}, t)| \leq g(\Delta t),$$

where g satisfies $g(\Delta t) > 0$ and $g(\Delta t) \rightarrow 0$ as $\Delta t \rightarrow 0$. Thus

$$|T_{D_2}| \leq C_{\mathbf{S}} g(\Delta t) \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \left| \nabla c_h^n|_K \right| |K| \leq C_{\mathbf{S}} g(\Delta t) \left(\frac{C_{\text{ae}}}{C_{\mathbf{S}}} \right)^{\frac{1}{2}} T^{\frac{1}{2}} |\Omega|^{\frac{1}{2}},$$

using the Cauchy–Schwarz inequality and the a priori estimate (1.28).

We now turn to T_{D_3} . We will show later that

$$T'_{D_3} := \int_0^T \sum_{K \in \mathcal{T}_h} \int_K \left(\nabla c_{h,\Delta t}(\mathbf{x}, t) - \nabla u(\mathbf{x}, t) \right) \cdot \mathbf{w}(\mathbf{x}, t) \, d\mathbf{x} \, dt \longrightarrow 0 \quad (1.58)$$

as $h, \Delta t \rightarrow 0$ for all $\mathbf{w} \in [C^\infty(\overline{Q_T})]^d$. Using the density of the set $[C^\infty(\overline{Q_T})]^d$ in $[L^2(Q_T)]^d$, we will conclude a weak convergence of $\nabla c_{h,\Delta t}$ (piecewise constant function in space and time) to ∇u . Next, $(\mathbf{S}_{\Delta t})_{i,j}$, $1 \leq i, j \leq d$, converge strongly in $L^1(Q_T)$ to $\mathbf{S}_{i,j}$ by its definition (1.55). Using the boundedness of $\mathbf{S}_{\Delta t}$ and \mathbf{S} given by Assumption (A3), we have also strong $L^2(Q_T)$ convergence. Hence it suffices to apply Lemma 1.8.5 from Appendix 1.8 to conclude that $T_{D_3} \rightarrow 0$ as $h, \Delta t \rightarrow 0$, provided that (1.58) is satisfied.

To show (1.58), we first rewrite T'_{D_3} as

$$T'_{D_3} = \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K \in \mathcal{T}_h} \int_K \nabla c_h^n(\mathbf{x}) \cdot \mathbf{w}(\mathbf{x}, t) \, d\mathbf{x} \, dt + \int_0^T \int_\Omega u(\mathbf{x}, t) \nabla \cdot \mathbf{w}(\mathbf{x}, t) \, d\mathbf{x} \, dt,$$

where we have used the Green theorem for u (recall that $u \in L^2(0, T; H_0^1(\Omega))$ by Theorem 1.6.1) and \mathbf{w} . We easily notice that we cannot use the Green theorem for c_h^n on Ω , since $c_h^n \notin H^1(\Omega)$. We are thus forced to apply it on each $K \in \mathcal{T}_h$.

We rewrite the first term of T'_{D_3} as

$$\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K \in \mathcal{T}_h} \int_K -c_h^n(\mathbf{x}) \nabla \cdot \mathbf{w}(\mathbf{x}, t) \, d\mathbf{x} \, dt + \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K \in \mathcal{T}_h} \int_{\partial K} c_h^n(\mathbf{x}) \mathbf{w}(\mathbf{x}, t) \cdot \mathbf{n} \, d\gamma(\mathbf{x}) \, dt.$$

We next consider the term

$$T''_{D_3} := \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K \in \mathcal{T}_h} \int_{\partial K} c_h^n \mathbf{w} \cdot \mathbf{n} \, d\gamma(\mathbf{x}) \, dt. \quad (1.59)$$

Reordering the summation by sides, we come to

$$\begin{aligned} T''_{D_3} &= \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \left(\sum_{\sigma_{K,L} \in \mathcal{E}_h^{\text{int}}} \int_{\sigma_{K,L}} (c_h^n|_K - c_h^n|_L) \mathbf{w} \cdot \mathbf{n}_{K,L} \, d\gamma(\mathbf{x}) \right. \\ &\quad \left. + \sum_{\sigma_K \in \mathcal{E}_h^{\text{ext}}} \int_{\sigma_K} c_h^n|_K \mathbf{w} \cdot \mathbf{n}_K \, d\gamma(\mathbf{x}) \right) dt, \end{aligned}$$

where we have used $\mathbf{w} \cdot \mathbf{n}_{K,L} = -\mathbf{w} \cdot \mathbf{n}_{L,K}$ following from $\mathbf{w} \in [C^\infty(\overline{Q_T})]^d$. The functions $c_h^n|_K - c_h^n|_L$ or $c_h^n|_K$ restricted to a side $\sigma_{K,L} \in \mathcal{E}_h^{\text{int}}$ or $\sigma_K \in \mathcal{E}_h^{\text{ext}}$, respectively, are first-order polynomials, vanishing in the barycentre Q_D of this side. For $\sigma_{K,L} \in \mathcal{E}_h^{\text{int}}$, this follows from the continuity requirement given in the definition of X_h and for $\sigma_K \in \mathcal{E}_h^{\text{ext}}$ from the zero Dirichlet boundary condition imposed by (1.9b). Hence

$$\int_{\sigma_{K,L}} (c_h^n|_K(\mathbf{x}) - c_h^n|_L(\mathbf{x})) \, d\gamma(\mathbf{x}) = 0, \quad \int_{\sigma_K} c_h^n|_K(\mathbf{x}) \, d\gamma(\mathbf{x}) = 0 \quad (1.60)$$

for all $\sigma_{K,L} \in \mathcal{E}_h^{\text{int}}$ and $\sigma_K \in \mathcal{E}_h^{\text{ext}}$, since the quadrature formula using the value in the barycentre of a segment ($d = 2$) or a triangle ($d = 3$) is precise for linear functions. We further estimate

$$\left| c_h^n|_K(\mathbf{x}) \right| = \left| c_h^n|_K(\mathbf{x}) - c_h^n|_K(Q_D) \right| \leq \left| \nabla c_h^n|_K \right| |\mathbf{x} - Q_D| \leq \left| \nabla c_h^n|_K \right| \frac{\text{diam}(\sigma_K)}{4-d},$$

with $\mathbf{x} \in \sigma_K \in \mathcal{E}_h^{\text{ext}}$, where we have used $|\mathbf{x} - Q_D| \leq \text{diam}(\sigma_K)/2$ for $d = 2$ but only $|\mathbf{x} - Q_D| \leq \text{diam}(\sigma_K)$ for $d = 3$. Similarly,

$$\left| c_h^n|_K(\mathbf{x}) - c_h^n|_L(\mathbf{x}) \right| \leq \left| \nabla c_h^n|_K \right| \frac{\text{diam}(\sigma_{K,L})}{4-d} + \left| \nabla c_h^n|_L \right| \frac{\text{diam}(\sigma_{K,L})}{4-d} \quad \mathbf{x} \in \sigma_{K,L} \in \mathcal{E}_h^{\text{int}}.$$

We have from the smoothness of \mathbf{w}

$$\mathbf{w} \cdot \mathbf{n}_{\sigma_D}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{n}_{\sigma_D}(Q_D) + f(\boldsymbol{\xi})|Q_D - \mathbf{x}| \quad \mathbf{x} \in \sigma_D \in \mathcal{E}_h, \boldsymbol{\xi} \in [Q_D, \mathbf{x}]$$

with $|f(\boldsymbol{\xi})| \leq C_{\mathbf{w}}$. Thus

$$\int_{\sigma_K} c_h^n|_K(\mathbf{x}) f(\boldsymbol{\xi}) |Q_D - \mathbf{x}| d\gamma(\mathbf{x}) \leq C_{\mathbf{w}} \left(\frac{\text{diam}(\sigma_K)}{4-d} \right)^2 \left| \nabla c_h^n|_K \right| |\sigma_K|$$

for an exterior side σ_K and similarly

$$\int_{\sigma_{K,L}} \left(c_h^n|_K(\mathbf{x}) - c_h^n|_L(\mathbf{x}) \right) f(\boldsymbol{\xi}) |Q_D - \mathbf{x}| d\gamma(\mathbf{x}) \leq C_{\mathbf{w}} \left(\frac{\text{diam}(\sigma_{K,L})}{4-d} \right)^2 \left(\left| \nabla c_h^n|_K \right| + \left| \nabla c_h^n|_L \right| \right) |\sigma_{K,L}|$$

for an interior side $\sigma_{K,L}$. Using these estimates, we immediately come to

$$\begin{aligned} |T_{D_3}''| &\leq C_{\mathbf{w}} \frac{h}{(4-d)^2} \frac{d+1}{d-1} \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K \in \mathcal{T}_h} \left| \nabla c_h^n|_K \right| \text{diam}(K)^d dt \\ &\leq \frac{C_{\mathbf{w}}}{\kappa_{\mathcal{T}}} \frac{h}{(4-d)^2} \frac{d+1}{d-1} \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \left| \nabla c_h^n|_K \right| |K| \leq \frac{C_{\mathbf{w}}}{\kappa_{\mathcal{T}}} \frac{h}{(4-d)^2} \frac{d+1}{d-1} \left(\frac{C_{\text{ae}}}{C_{\text{S}}} \right)^{\frac{1}{2}} T^{\frac{1}{2}} |\Omega|^{\frac{1}{2}}, \end{aligned}$$

using the fact that each $\nabla c_h^n|_K$ is in the summation over all sides just $(d+1)$ -times, $|\sigma_D| \leq \text{diam}(K)^{d-1}/(d-1)$ and $\text{diam}(\sigma_D) \leq \text{diam}(K) \leq h$ for all $\sigma_D \in \mathcal{E}_K$, Assumption (B), the Cauchy–Schwarz inequality, and the a priori estimate (1.28). Thus $T_{D_3}'' \rightarrow 0$ as $h \rightarrow 0$.

To conclude that $T_{D_3}' \rightarrow 0$ as $h, \Delta t \rightarrow 0$, it remains to show that

$$-\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K \in \mathcal{T}_h} \int_K c_h^n(\mathbf{x}) \nabla \cdot \mathbf{w}(\mathbf{x}, t) d\mathbf{x} dt + \int_0^T \int_{\Omega} u(\mathbf{x}, t) \nabla \cdot \mathbf{w}(\mathbf{x}, t) d\mathbf{x} dt \rightarrow 0.$$

This is however immediate, since we can rewrite it as

$$\int_0^T \int_{\Omega} (u(\mathbf{x}, t) - c_{h,\Delta t}(\mathbf{x}, t)) \nabla \cdot \mathbf{w}(\mathbf{x}, t) d\mathbf{x} dt \rightarrow 0,$$

which is a consequence of the strong $L^2(Q_T)$ convergence of $c_{h,\Delta t}$ to u . Thus $T_{D_3}' \rightarrow 0$ and consequently $T_{D_3} \rightarrow 0$ as $h, \Delta t \rightarrow 0$.

Using (1.56) and (1.57), we see that

$$T_D \rightarrow \int_0^T \int_{\Omega} \mathbf{S} \nabla u(\mathbf{x}, t) \cdot \nabla \psi(\mathbf{x}, t) d\mathbf{x} dt \quad \text{as } h, \Delta t \rightarrow 0. \quad (1.61)$$

Remark 1.6.3. (Nonconforming approximation) *The fact that T_{D_3}'' given by (1.59) is not immediately equal to zero is the consequence of the nonconforming-type approximation. However, since the approximation is continuous in the barycentres of interior sides and equal to zero in the barycentres of exterior sides, (1.60) is fulfilled and consequently T_{D_3}'' is of order h , which suffices for the convergence.*

Convection term

We begin by denoting

$$\mathbf{v}^n(\mathbf{x}) := \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \mathbf{v}(\mathbf{x}, t) dt \quad n \in \{1, 2, \dots, N\}, \mathbf{x} \in \Omega. \quad (1.62)$$

We now show the validity of two passages to the limit. First, we intend to show that

$$\begin{aligned} & \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{c_{D,E}^n} \psi(Q_D, t_{n-1}) + \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} c_D^n \sum_{E \in \mathcal{N}(D)} \\ & \int_{\sigma_{D,E}} \mathbf{v}^n(\mathbf{x}) \cdot \mathbf{n}_{D,E} \psi(\mathbf{x}, t_{n-1}) d\gamma(\mathbf{x}) - \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} c_D^n \int_D \nabla \cdot \mathbf{v}^n(\mathbf{x}) \psi(\mathbf{x}, t_{n-1}) dx \longrightarrow 0 \end{aligned} \quad (1.63)$$

as $h \rightarrow 0$. We add and subtract the terms $c_D^n \psi(Q_D, t_{n-1}) \mathbf{v}_{D,E}^n$ and $\overline{c_{D,E}^n} \int_{\sigma_{D,E}} \mathbf{v}^n(\mathbf{x}) \cdot \mathbf{n}_{D,E} \psi(\mathbf{x}, t_{n-1}) d\gamma(\mathbf{x})$ to each term of the left-hand side of (1.63). We denote

$$\begin{aligned} T_{C_1} &:= \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} \sum_{E \in \mathcal{N}(D)} (\overline{c_{D,E}^n} - c_D^n) \left(\psi(Q_D, t_{n-1}) \mathbf{v}_{D,E}^n \right. \\ & \quad \left. - \int_{\sigma_{D,E}} \mathbf{v}^n(\mathbf{x}) \cdot \mathbf{n}_{D,E} \psi(\mathbf{x}, t_{n-1}) d\gamma(\mathbf{x}) \right), \\ T_{C_2} &:= \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} \sum_{E \in \mathcal{N}(D)} \overline{c_{D,E}^n} \int_{\sigma_{D,E}} \mathbf{v}^n(\mathbf{x}) \cdot \mathbf{n}_{D,E} \psi(\mathbf{x}, t_{n-1}) d\gamma(\mathbf{x}), \\ T_{C_3} &:= \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} c_D^n \psi(Q_D, t_{n-1}) \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n, \\ T_{C_4} &:= \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} c_D^n \int_D \nabla \cdot \mathbf{v}^n(\mathbf{x}) \psi(\mathbf{x}, t_{n-1}) dx. \end{aligned}$$

One can easily verify that (1.63) is satisfied when $T_{C_1} \rightarrow 0$, $T_{C_2} \rightarrow 0$, and $(T_{C_3} - T_{C_4}) \rightarrow 0$ as $h \rightarrow 0$.

We begin with T_{C_2} . We denote

$$\mathbf{v}_{\psi;D,E}^n := \int_{\sigma_{D,E}} \mathbf{v}^n(\mathbf{x}) \cdot \mathbf{n}_{D,E} \psi(\mathbf{x}, t_{n-1}) d\gamma(\mathbf{x}).$$

Since the summation in T_{C_2} is over all $D \in \mathcal{D}_h$ and all its neighbors, each interior dual side is in the summation just twice. We consider one fixed interior dual side $\sigma_{D,E}$, where we have denoted D and E such that $\mathbf{v}_{D,E}^n \geq 0$, and have

$$\left(c_D^n + \alpha_{D,E}^n (c_E^n - c_D^n) \right) \mathbf{v}_{\psi;D,E}^n + \left(c_D^n + \alpha_{D,E}^n (c_E^n - c_D^n) \right) \mathbf{v}_{\psi;E,D}^n = 0,$$

considering the definition of the local Péclet upstream weighting (1.10) and $\mathbf{v}_{\psi;D,E}^n = -\mathbf{v}_{\psi;E,D}^n$. Thus $T_{C_2} = 0$.

Next we consider T_{C_3} and T_{C_4} . We immediately have that

$$\sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n = \int_D \nabla \cdot \mathbf{v}^n(\mathbf{x}) dx \quad \forall D \in \mathcal{D}_h^{\text{int}},$$

using the definition of $\mathbf{v}_{D,E}^n$. We further estimate

$$\begin{aligned}
|T_{C_3} - T_{C_4}| &= \left| \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h^{\text{int}}} c_D^n \int_D \nabla \cdot \mathbf{v}^n(\mathbf{x}) \left(\psi(Q_D, t_{n-1}) - \psi(\mathbf{x}, t_{n-1}) \right) d\mathbf{x} \right| \\
&\leq C_{2,\psi} h \mu \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} |c_D^n| \int_{t_{n-1}}^{t_n} \int_D q_S(\mathbf{x}, t) d\mathbf{x} dt \\
&\leq C_{2,\psi} h \mu \left(\sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \Delta t_n |D| (c_D^n)^2 \right)^{\frac{1}{2}} \left(\sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \frac{\left(\int_{t_{n-1}}^{t_n} \int_D q_S(\mathbf{x}, t) d\mathbf{x} dt \right)^2}{\Delta t_n |D|} \right)^{\frac{1}{2}} \\
&\leq C_{2,\psi} h \mu \left(\frac{C_{\text{ae}}}{c_\beta} T \right)^{\frac{1}{2}} \|q_S\|_{0, Q_T},
\end{aligned} \tag{1.64}$$

considering the boundary condition $c_D^n = 0$ for all $D \in \mathcal{D}_h^{\text{ext}}$,

$$|\psi(Q_D, t_{n-1}) - \psi(\mathbf{x}, t_{n-1})| \leq C_{2,\psi} h \tag{1.65}$$

for all $\mathbf{x} \in D$,

$$\int_D |\nabla \cdot \mathbf{v}^n(\mathbf{x})| d\mathbf{x} = \frac{1}{\Delta t_n} \int_D \int_{t_{n-1}}^{t_n} \nabla \cdot \mathbf{v}(\mathbf{x}, t) dt d\mathbf{x} = \frac{1}{\Delta t_n} \int_D \int_{t_{n-1}}^{t_n} q_S(\mathbf{x}, t) dt d\mathbf{x},$$

which follows from Assumption (A4), the Cauchy–Schwarz inequality, and the a priori estimate (1.26). Thus $(T_{C_3} - T_{C_4}) \rightarrow 0$ as $h \rightarrow 0$.

We finally turn to T_{C_1} . We first define

$$T_{C_5} := \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} \sum_{E \in \mathcal{N}(D)} \text{diam}(K_{D,E})^{d-2} (\overline{c_{D,E}^n} - c_D^n)^2.$$

We have

$$(\overline{c_{D,E}^n} - c_D^n)^2 = \left(\alpha_{D,E}^n (c_E^n - c_D^n) \right)^2 \leq \frac{1}{4} (c_E^n - c_D^n)^2$$

when $\mathbf{v}_{D,E}^n \geq 0$, considering the definition of the local Péclet upstream weighting (1.10) and Remark 1.3.3, which gives $0 \leq \alpha_{D,E}^n \leq 1/2$. Similarly, when $\mathbf{v}_{D,E}^n < 0$, we come to

$$(\overline{c_{D,E}^n} - c_D^n)^2 = \left((c_E^n - c_D^n)(1 - \alpha_{D,E}^n) \right)^2 \leq (c_E^n - c_D^n)^2.$$

We have

$$\begin{aligned}
T_{C_5} &\leq 2 \sum_{n=1}^N \Delta t_n \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \text{diam}(K_{D,E})^{d-2} (c_E^n - c_D^n)^2 \\
&\leq \frac{d+1}{d\kappa_{\mathcal{T}}} \sum_{n=1}^N \Delta t_n \|c_h^n\|_{X_h}^2 \leq \frac{d+1}{d\kappa_{\mathcal{T}}} \frac{C_{\text{ae}}}{c_{\mathbf{S}}},
\end{aligned}$$

noticing that each interior dual side is in the original summation just twice, using the estimate (1.6) and the a priori estimate (1.28). We next define

$$\begin{aligned}
T_{C_6} &:= \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} \sum_{E \in \mathcal{N}(D)} \frac{1}{\text{diam}(K_{D,E})^{d-2}} \\
&\quad \left(\int_{\sigma_{D,E}} \mathbf{v}^n(\mathbf{x}) \cdot \mathbf{n}_{D,E} \left(\psi(Q_D, t_{n-1}) - \psi(\mathbf{x}, t_{n-1}) \right) d\gamma(\mathbf{x}) \right)^2
\end{aligned}$$

and estimate

$$\begin{aligned} |T_{C_6}| &\leq C_{2,\psi}^2 h^2 C_v^2 \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} \sum_{E \in \mathcal{N}(D)} \frac{1}{\text{diam}(K_{D,E})^{d-2}} |\sigma_{D,E}|^2 \\ &\leq C_{2,\psi}^2 h^2 C_v^2 \frac{(d+1)d}{(d-1)^2} \sum_{n=1}^N \Delta t_n \sum_{K \in \mathcal{T}_h} \text{diam}(K)^d \leq C_{2,\psi}^2 h^2 \frac{C_v^2}{\kappa_T} \frac{(d+1)d}{(d-1)^2} |\Omega| T, \end{aligned}$$

using (1.65), $|\mathbf{v}_{D,E}^n| \leq C_v |\sigma_{D,E}|$ following from Assumption (A4), (1.8), noticing that each interior dual side is in the original summation just twice and that each $K \in \mathcal{T}_h$ contains exactly $\binom{d+1}{2} = \frac{(d+1)d}{2}$ dual sides, and finally Assumption (B). We now notice that

$$T_{C_1}^2 \leq \mu^2 T_{C_5} T_{C_6},$$

using the Cauchy–Schwarz inequality, and hence $T_{C_1} \rightarrow 0$ as $h \rightarrow 0$. Thus (1.63) is satisfied.

Using the Green theorem and considering $c_D^n = 0$ for all $D \in \mathcal{D}_h^{\text{ext}}$, we easily come to

$$\begin{aligned} &\mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} c_D^n \sum_{E \in \mathcal{N}(D)} \int_{\sigma_{D,E}} \mathbf{v}^n(\mathbf{x}) \cdot \mathbf{n}_{D,E} \psi(\mathbf{x}, t_{n-1}) \, d\gamma(\mathbf{x}) \Delta t_n \quad (1.66) \\ &= \mu \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} c_D^n \int_D \mathbf{v}^n(\mathbf{x}) \nabla \psi(\mathbf{x}, t_{n-1}) \, d\mathbf{x} + \mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} c_D^n \int_D \nabla \cdot \mathbf{v}^n(\mathbf{x}) \psi(\mathbf{x}, t_{n-1}) \, d\mathbf{x}. \end{aligned}$$

We will now demonstrate that

$$\begin{aligned} &\mu \sum_{n=1}^N \Delta t_n \sum_{D \in \mathcal{D}_h} c_D^n \int_D \mathbf{v}^n(\mathbf{x}) \cdot \nabla \psi(\mathbf{x}, t_{n-1}) \, d\mathbf{x} \quad (1.67) \\ &\longrightarrow \mu \int_0^T \int_{\Omega} u(\mathbf{x}, t) \mathbf{v}(\mathbf{x}, t) \cdot \nabla \psi(\mathbf{x}, t) \, d\mathbf{x} \, dt \text{ as } h, \Delta t \rightarrow 0. \end{aligned}$$

We introduce

$$\begin{aligned} T_{C_7} &:= \mu \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \int_{\Omega} \tilde{c}_{h,\Delta t}(\mathbf{x}, t) \mathbf{v}^n(\mathbf{x}) \cdot \left(\nabla \psi(\mathbf{x}, t_{n-1}) - \nabla \psi(\mathbf{x}, t) \right) \, d\mathbf{x} \, dt, \\ T_{C_8} &:= \mu \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \int_{\Omega} \left(\tilde{c}_{h,\Delta t}(\mathbf{x}, t) - u(\mathbf{x}, t) \right) \mathbf{v}^n(\mathbf{x}) \cdot \nabla \psi(\mathbf{x}, t) \, d\mathbf{x} \, dt, \\ T_{C_9} &:= \mu \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \int_{\Omega} u(\mathbf{x}, t) \left(\mathbf{v}^n(\mathbf{x}) - \mathbf{v}(\mathbf{x}, t) \right) \cdot \nabla \psi(\mathbf{x}, t) \, d\mathbf{x} \, dt. \end{aligned}$$

We have

$$|\nabla \psi(\mathbf{x}, t_{n-1}) - \nabla \psi(\mathbf{x}, t)| \leq g(\Delta t)$$

for $t \in (t_{n-1}, t_n]$ and thus

$$|T_{C_7}| \leq g(\Delta t) \mu \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} |c_D^n| \int_D \int_{t_{n-1}}^{t_n} |\mathbf{v}(\mathbf{x}, t)| \, d\mathbf{x} \, dt \leq g(\Delta t) \mu \left(\frac{C_{\text{ae}}}{c_\beta} T \right)^{\frac{1}{2}} \|\mathbf{v}\|_{0, Q_T},$$

using the same estimate as in (1.64). Thus $T_{C_7} \rightarrow 0$ as $\Delta t \rightarrow 0$. It is immediate that $T_{C_8} \rightarrow 0$ as $h, \Delta t \rightarrow 0$, using the strong (and consequently weak) convergence of $\tilde{c}_{h,\Delta t}$ to u . By Assumption (A4) and (1.62) \mathbf{v} and \mathbf{v}^n are bounded, and hence the piecewise constant in time approximation given by \mathbf{v}^n converges strongly in $\mathbf{L}^2(Q_T)$ to \mathbf{v} as $\Delta t \rightarrow 0$. Since $|\nabla\psi| \leq C_{2,\psi}$ and $u \in L^2(Q_T)$, it suffices to use the Cauchy–Schwarz inequality to conclude that $T_{C_9} \rightarrow 0$ as $\Delta t \rightarrow 0$. Thus (1.67) is fulfilled. Finally, using (1.63), (1.66), and (1.67), we see that

$$T_C \longrightarrow -\mu \int_0^T \int_{\Omega} u(\mathbf{x}, t) \mathbf{v}(\mathbf{x}, t) \cdot \nabla \psi(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad \text{as } h, \Delta t \rightarrow 0. \quad (1.68)$$

Reaction term

We would now like to show that

$$T_R \longrightarrow \int_0^T \int_{\Omega} F(u(\mathbf{x}, t)) \psi(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad \text{as } h, \Delta t \rightarrow 0. \quad (1.69)$$

For this purpose, we introduce

$$\begin{aligned} T_{R_1} &:= \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} F(c_D^n) \int_{t_{n-1}}^{t_n} \int_D \left(\psi(Q_D, t_{n-1}) - \psi(\mathbf{x}, t) \right) \, d\mathbf{x} \, dt, \\ T_{R_2} &:= \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \int_{t_{n-1}}^{t_n} \int_D \left(F(c_D^n) - F(u(\mathbf{x}, t)) \right) \psi(\mathbf{x}, t) \, d\mathbf{x} \, dt. \end{aligned}$$

We have

$$|\psi(Q_D, t_{n-1}) - \psi(\mathbf{x}, t)| \leq C_{3,\psi}(h + \Delta t) \quad (1.70)$$

for all $\mathbf{x} \in D$ and $t \in (t_{n-1}, t_n]$, and thus

$$|T_{R_1}| \leq C_{3,\psi} L_F (h + \Delta t) \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \Delta t_n |D| |c_D^n| \leq C_{3,\psi} L_F (h + \Delta t) \left(\frac{C_{ae}}{c_\beta} T \right)^{\frac{1}{2}} |\Omega|^{\frac{1}{2}} T^{\frac{1}{2}},$$

using the Lipschitz continuity of F , following either from Assumption (A6) or (A7), the Cauchy–Schwarz inequality, and the a priori estimate (1.26). Hence, $T_{R_1} \rightarrow 0$ as $h, \Delta t \rightarrow 0$.

We have

$$|T_{R_2}| \leq C_{1,\psi} L_F \int_0^T \int_{\Omega} |\tilde{c}_{h,\Delta t}(\mathbf{x}, t) - u(\mathbf{x}, t)| \, d\mathbf{x} \, dt,$$

which tends to 0 because of the strong $L^2(Q_T)$ convergence of $\tilde{c}_{h,\Delta t}$ to u . Thus, (1.69) is fulfilled.

Sources term

We finally show that

$$T_S \longrightarrow \int_0^T \int_{\Omega} q(\mathbf{x}, t) \psi(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad \text{as } h, \Delta t \rightarrow 0. \quad (1.71)$$

We set

$$\begin{aligned} T_{S_1} &:= \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} q_D^n \int_{t_{n-1}}^{t_n} \int_D \left(\psi(Q_D, t_{n-1}) - \psi(\mathbf{x}, t) \right) \mathrm{d}\mathbf{x} \mathrm{d}t, \\ T_{S_2} &:= \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \int_{t_{n-1}}^{t_n} \int_D \left(q_D^n - q(\mathbf{x}, t) \right) \psi(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t. \end{aligned}$$

We have

$$|T_{S_1}| \leq C_{3,\psi}(h + \Delta t) \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \int_{t_{n-1}}^{t_n} \int_D |q(\mathbf{x}, t)| \mathrm{d}\mathbf{x} \mathrm{d}t \leq C_{3,\psi}(h + \Delta t) \|q\|_{0, Q_T} |\Omega|^{\frac{1}{2}} T^{\frac{1}{2}},$$

using (1.70) and the Cauchy–Schwarz inequality. Finally,

$$|T_{S_2}| \leq C_{1,\psi} \sum_{n=1}^N \sum_{D \in \mathcal{D}_h} \int_{t_{n-1}}^{t_n} \int_D |q_D^n - q(\mathbf{x}, t)| \mathrm{d}\mathbf{x} \mathrm{d}t,$$

which tends to 0 as $h, \Delta t \rightarrow 0$ because of the L^1 convergence of the piecewise constant approximation q_D^n to q . Thus (1.71) is satisfied.

We are now ready to give the final theorem of this chapter:

Theorem 1.6.4. (Convergence to a weak solution) *There exist subsequences of $\tilde{c}_{h,\Delta t}$ and $c_{h,\Delta t}$, the approximate solutions of the problem (1.1)–(1.3) by means of the combined finite volume–nonconforming/mixed-hybrid finite element scheme (1.9a)–(1.9c) given by Definition 1.5.3, which converge strongly in $L^2(Q_T)$ to a weak solution of the problem (1.1)–(1.3) given by Definition 1.2.2. If the weak solution is unique, then the whole sequences $\tilde{c}_{h,\Delta t}$, $c_{h,\Delta t}$ converge to the weak solution.*

PROOF:

We have from Theorem 1.6.1 that subsequences of $\tilde{c}_{h,\Delta t}$ and $c_{h,\Delta t}$ converge strongly in $L^2(Q_T)$ to some function $u \in L^2(0, T; H_0^1(\Omega))$. The function u satisfies

$$\begin{aligned} & - \int_0^T \int_{\Omega} \beta(u(\mathbf{x}, t)) \psi_t(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t - \int_{\Omega} \beta(c_0(\mathbf{x})) \psi(\mathbf{x}, 0) \mathrm{d}\mathbf{x} \\ & + \int_0^T \int_{\Omega} \mathbf{S}(\mathbf{x}, t) \nabla u(\mathbf{x}, t) \cdot \nabla \psi(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t - \mu \int_0^T \int_{\Omega} u(\mathbf{x}, t) \mathbf{v}(\mathbf{x}, t) \cdot \nabla \psi(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t \\ & + \int_0^T \int_{\Omega} F(u(\mathbf{x}, t)) \psi(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t = \int_0^T \int_{\Omega} q(\mathbf{x}, t) \psi(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t \end{aligned}$$

for all test functions $\psi \in \Psi$, given by (1.46). This follows from (1.54), (1.61), (1.68), (1.69), (1.71), and (1.47). In addition, $\beta(u) \in L^\infty(0, T; L^2(\Omega))$, which follows from (1.27). Thus u is a weak solution of the problem (1.1)–(1.3), because of the density of the set Ψ in the set $\{\varphi; \varphi \in L^2(0, T; H_0^1(\Omega)), \varphi_t \in L^\infty(Q_T), \varphi(\cdot, T) = 0\}$. If the weak solution given by Definition 1.2.2 is unique, then by contradiction the convergence statement is valid for the whole sequences $\tilde{c}_{h,\Delta t}$ and $c_{h,\Delta t}$. \square

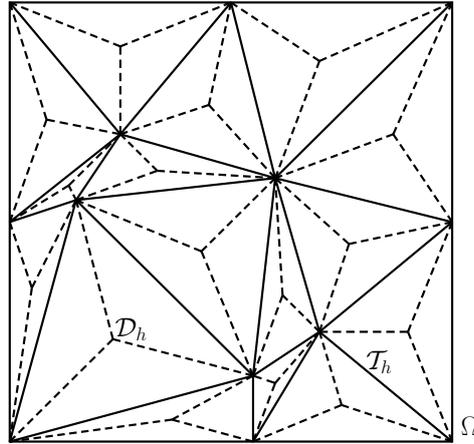


Figure 1.2: Initial space mesh \mathcal{T}_h (solid) and its dual mesh \mathcal{D}_h (dashed)

1.7 Numerical experiment

We present the results of a numerical experiment in this section. The computations were done in double precision on a notebook with Intel Pentium 4-M 1.8 GHz processor and MS Windows XP operating system. Machine precision was in power of 10^{-16} .

We test a model degenerate parabolic convection–diffusion problem with a known traveling wave solution (cf. [81]). In particular, we consider the equation (1.1) for $\Omega = (0, 1) \times (0, 1)$ and $T = 1$ with

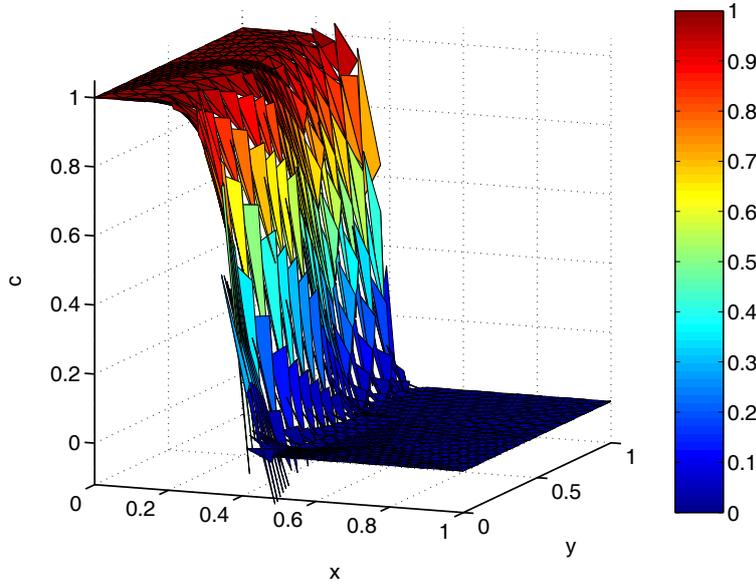
$$\begin{aligned}\beta(c) &= c^{\frac{1}{2}} \text{ for } c \geq 0, \\ \mathbf{S} &= \delta \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ \mathbf{v} &= (v, 0), \\ \mu &= 1, F(c) = 0, q = 0.\end{aligned}$$

Here, $\delta > 0$ and $v > 0$ are parameters. We fix v to 0.8 and let δ vary: for large values of δ , diffusion dominates over convection and conversely for small values of δ . The initial and Dirichlet boundary conditions are given by the exact solution

$$c(x, y, t) = \left(1 - e^{\frac{v}{2\delta}(x-vt-p)}\right)^2 \text{ for } x \leq vt + p, \quad c(x, y, t) = 0 \text{ for } x \geq vt + p.$$

The shift p defines the position of the front of the wave at $t = 0$ and is set to 0.2. Note that the problem is degenerate parabolic since $\beta'(0) = +\infty$ and the solution takes the value of 0.

We perform the simulations on an unstructured triangular mesh of the space domain; the initial mesh is given in Figure 1.2. The initial time step is $T/2$. We refine the space mesh by dividing each triangle regularly into four subtriangles. Each time the space mesh is refined, the time step is divided by two. We define the Péclet number by $\text{Pe} := hv/\delta$. The initial conditions are the values of the exact solution for $t = 0$ at the midpoints of triangle edges. The boundary conditions are given similarly. The simulated problem is only one-dimensional. We use this fact to test the performance of the numerical scheme that we propose for strongly irregular two-dimensional meshes. The case where the triangular mesh contains angles greater than $\pi/2$ is similar to the case where the diffusion tensor is anisotropic: in both cases the

Figure 1.3: Approximate solution at $t = 0.25$, $\delta = 0.01$, $r = 3$

discrete maximum principle is not necessarily satisfied (recall that this principle holds under Assumption (D) , cf. Theorem 1.4.11). Hence we need to define the function $\beta(c)$ for $c < 0$. To fulfill Assumptions $(A1)$ and $(A2)$, we set $\beta(c) := -\beta(-c)$ for $c < 0$.

At each discrete time level, we have to solve the nonlinear system of algebraic equations given by (1.9a)–(1.9c). Since $\beta'(0) = +\infty$ and since the solution takes the value of 0, we cannot directly apply the Newton method for this purpose. The traditional finite element technique to overcome this difficulty consists in regularization (approximation of β by functions with bounded slope), cf. [20]. Another method, applicable however only when the discrete maximum principle holds, consists in perturbing the initial and boundary conditions so that all the values that the scheme works with were strictly positive (the problem is no more degenerate parabolic), see [98, 99]. We use here a method which has been used in the field of the finite volume method (cf. [61]): we introduce new unknowns $u_D^n = \beta(c_D^n)$ and rewrite the system of equations (1.9a)–(1.9c) for these new unknowns. We believe that this approach is advantageous for the following reasons: (i) There is no need to regularize the problem or to perturb the data (now $[\beta^{-1}]'(0) = 0$); (ii) One can directly apply the Newton method to linearize the problem; (iii) The resulting matrices are diagonal for the part of the unknowns corresponding to the region where the concentration is zero. Indeed, on the step k of the linearization at time t_n , we approximate $c_E^{n,k} = \beta^{-1}(u_E^{n,k}) \approx \beta^{-1}(u_E^{n,k-1}) + (\beta^{-1})'(u_E^{n,k-1})(u_E^{n,k} - u_E^{n,k-1})$, which vanishes in view of $\beta^{-1}(0) = (\beta^{-1})'(0) = 0$. Let $\{u_D^{n,k}\}_{D \in \mathcal{D}_h^{\text{int}}}$ be the solution vector on the step k . The linearization is terminated whenever

$$\left(\sum_{D \in \mathcal{D}_h^{\text{int}}} (u_D^{n,k} - u_D^{n,k-1})^2 \right)^{\frac{1}{2}} / \left(\sum_{D \in \mathcal{D}_h^{\text{int}}} (u_D^{n,k})^2 \right)^{\frac{1}{2}} \leq 1e-10.$$

The bi-conjugate gradients stabilized method (Bi-CGStab) [103, 117], preconditioned by the LU incomplete factorization with drop tolerance 1e-3, cf. [111], is used for the solution of the associated linear systems. The iterations were stopped whenever the relative residual decreased below 1e-10.

Ref.	T. st.	Unkn.	$Pe_{\delta=0.05}$	$t_{\delta=0.05}$	$Pe_{\delta=0.01}$	$t_{\delta=0.01}$	$Pe_{\delta=0.0001}$	$t_{\delta=0.0001}$
1	4	88	4.56	0:01	22.80	0:01	2280.0	0:01
3	16	1504	1.14	0:16	5.70	0:15	570.0	0:11
5	64	24448	0.29	19:11	1.43	17:49	142.5	9:51

Table 1.1: Number of refinements, number of time steps, number of unknowns, Péclet number, and computational times in min:sec for $\delta = 0.05, 0.01,$ and $0.0001,$ respectively

We consider three values of δ : 0.05, 0.01, and 0.0001. The number of refinements is $r = 1, 3,$ and 5 ($r = 0$ corresponds to the initial mesh). We refer to Table 1.1 for the number of unknowns, Péclet numbers, and computational times. For the finest meshes, there were up to 15 Newton steps necessary in the first iteration. This number then decreased to approx. 7 per time step. We can see the approximate solution for $\delta = 0.01$ and $r = 3$ at $t = 0.25$ in Figure 1.3. We give the profiles of approximate solutions in $y = 0.5$ for the different values of δ and r in Figures 1.4 and 1.5. The profile in $y = 0.5$ is defined by all the calculated values c_D such that Q_D (the midpoint of the edge σ_D associated with the dual volume D) satisfies $|Q_D - l_{0.5}| \leq 0.25$ for $r = 1,$ $|Q_D - l_{0.5}| \leq 0.08$ for $r = 3,$ and $|Q_D - l_{0.5}| \leq 0.02$ for $r = 5,$ where $l_{0.5}$ is the line $y = 0.5$.

We finally give some comments on the results. First, the scheme works easily for the given irregular mesh, which would not be possible with the standard finite volume method, cf. [61]. This irregularity (angles greater than $\pi/2$) on the other hand causes the violation of the discrete maximum principle. However, this violation is only noticeable for the coarsest meshes ($r = 0, 1,$ in power of $1e-3$) and disappears with the refinement of the meshes. The scheme naturally works with negative values due to the appropriate definition of $\beta(c)$ for $c < 0$. We remark that the negative values of the approximation that are visible in Figure 1.3 have no relation to the discrete maximum principle; they are only a consequence of a piecewise linear interpretation of the (positive) values c_D^n . The influence of unsuitable shapes of the elements is also visible in Figures 1.4 and 1.5—notice the local fluctuations in the profiles for $r = 1$ and 3 . This influence is however only because of the finite volume part of the scheme, which can be easily verified by considering a pure hyperbolic problem. Next, the local Péclet upstream weighting reduces the numerical diffusion of full upstream weighting to the amount exactly necessary to ensure the stability of the scheme. In particular, the coefficients $\alpha_{D,E}^n$ given by (1.11) automatically increase with r . Moreover, the different values of these parameters for different dual sides of the mesh reflect the local ratio of the diffusion and convection fluxes (recall that e.g. for a dual side parallel with \mathbf{v} , the flux of \mathbf{v} through this side is zero). This numerical flux would be still more efficient for a problem where the ratio of v and δ is not uniform over Ω . Finally, precise approximation of realistic convection-dominated problems on fixed grids with the proposed scheme may still be expensive in terms of the computational cost. A local refinement strategy as that proposed in [95, 96] would then be necessary.

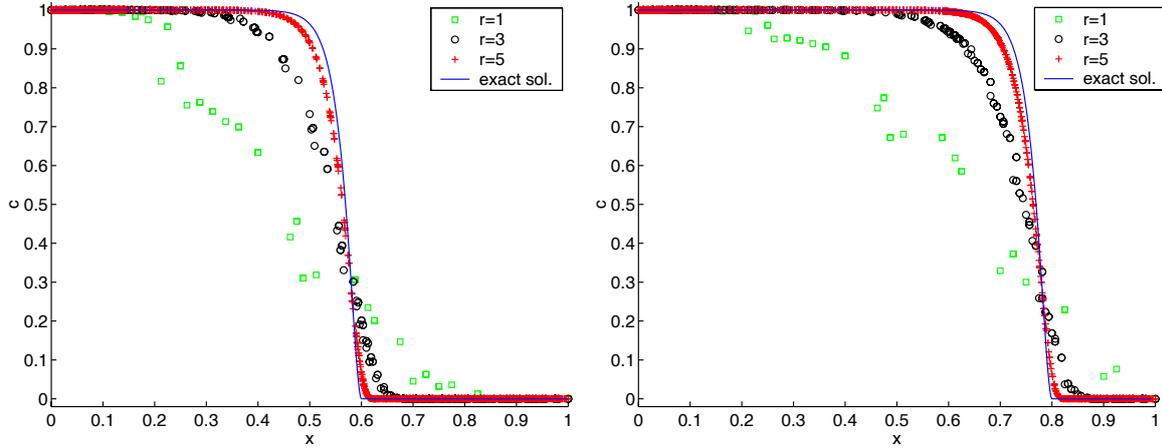
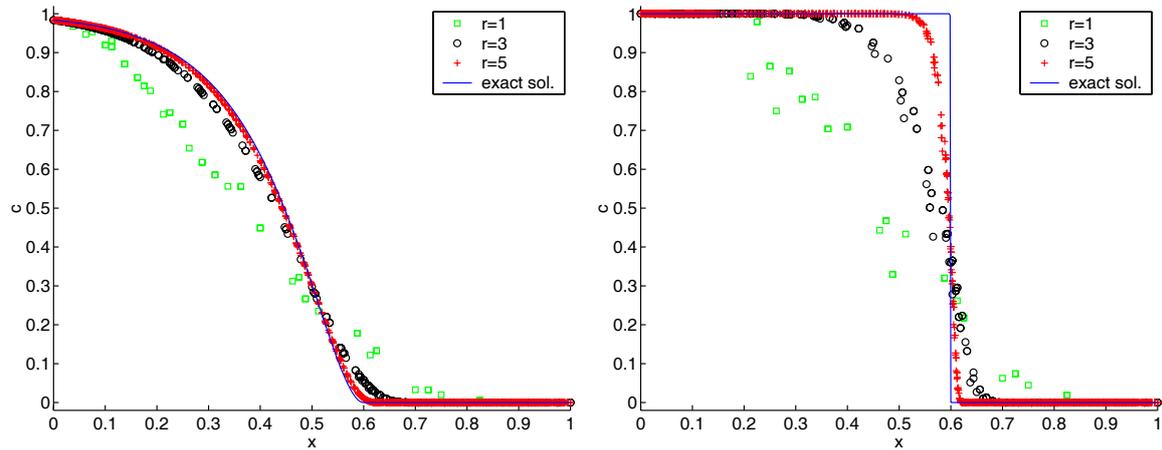
1.8 Appendix A: Technical lemmas

We give here some technical lemmas that were needed in this chapter.

Lemma 1.8.1. *Let us consider the elliptic problem*

$$-\nabla \cdot (\mathbf{S}\nabla p) = q \quad \text{in } \Omega, \quad (1.72a)$$

$$p = 0 \quad \text{on } \partial\Omega, \quad (1.72b)$$


 Figure 1.4: Solution profiles for $y = 0.5$ and $\delta = 0.01$, at $t = 0.5$ (left) and at $t = 0.75$ (right)

 Figure 1.5: Solution profiles for $y = 0.5$ at $t = 0.5$, $\delta = 0.05$ (left) and $\delta = 0.0001$ (right)

where $q \in L^2(\Omega)$. Then the stiffness matrix for the Lagrange multipliers of the hybridization of the lowest-order Raviart–Thomas mixed finite element method on the simplicial mesh \mathcal{T}_h has the form

$$\mathbb{M}_{D,E} = - \sum_{K \in \mathcal{T}_h} (\mathbf{S}_K \nabla \varphi_E, \nabla \varphi_D)_{0,K} \quad D, E \in \mathcal{D}_h^{\text{int}}, \quad (1.73)$$

where

$$\mathbf{S}_K = \left(\frac{1}{|K|} \int_K \mathbf{S}^{-1} \, d\mathbf{x} \right)^{-1} \quad \forall K \in \mathcal{T}_h. \quad (1.74)$$

PROOF:

The hybridization of the lowest-order Raviart–Thomas mixed finite element method for the problem (1.72a)–(1.72b) reads (cf. [33, Section V.1.2]): find $\mathbf{u}_h \in \mathbf{V}_h$, $p_h \in \Phi_h$, and $\lambda_h \in \Lambda_h$

such that

$$\sum_{K \in \mathcal{T}_h} \{(\mathbf{S}^{-1} \mathbf{u}_h, \mathbf{v}_h)_{0,K} - (\nabla \cdot \mathbf{v}_h, p_h)_{0,K} + \langle \mathbf{v}_h \cdot \mathbf{n}, \lambda_h \rangle_{\partial K}\} = 0 \quad \forall \mathbf{v}_h \in \mathbf{V}_h, \quad (1.75a)$$

$$- \sum_{K \in \mathcal{T}_h} (\nabla \cdot \mathbf{u}_h, \phi_h)_{0,K} = -(q, \phi_h)_{0,\Omega} \quad \forall \phi_h \in \Phi_h, \quad (1.75b)$$

$$\sum_{K \in \mathcal{T}_h} \langle \mathbf{u}_h \cdot \mathbf{n}, \mu_h \rangle_{\partial K} = 0 \quad \forall \mu_h \in \Lambda_h. \quad (1.75c)$$

Here, \mathbf{V}_h is the space of elementwise linear vector functions such that $\mathbf{u}_h \in \mathbf{V}_h$ satisfies $\mathbf{u}_h|_K = (a_K + d_K x, b_K + d_K y)$ if $d = 2$ and $\mathbf{u}_h|_K = (a_K + d_K x, b_K + d_K y, c_K + d_K z)$ if $d = 3$ for all $K \in \mathcal{T}_h$, Φ_h is the space of elementwise constant scalar functions, and Λ_h is the space of sidewise constant scalar Lagrange multipliers. For all $D \in \mathcal{D}_h$, we denote $\lambda_h|_{\sigma_D}$ by λ_D and require $\lambda_D = 0$ for all $D \in \mathcal{D}_h^{\text{ext}}$. We now extend the ideas of [38], where the tensor \mathbf{S} is supposed piecewise constant on \mathcal{T}_h .

Let us set $\tilde{\lambda}_h := \sum_{D \in \mathcal{D}_h} \lambda_D \varphi_D$. Using (1.16), we have

$$\sum_{\sigma_D \in \mathcal{E}_K} \lambda_D |\sigma_D| \mathbf{n}_{\sigma_D} = |K| \sum_{\sigma_D \in \mathcal{E}_K} \lambda_D \nabla \varphi_D|_K = |K| \nabla \tilde{\lambda}_h|_K.$$

Then denoting the unit coordinate vectors as \mathbf{e}_i and taking, respectively, $\mathbf{v}_h = \mathbf{e}_i$ in K , $1 \leq i \leq d$, $\mathbf{v}_h = 0$ otherwise as the test functions in (1.75a), we come to

$$\int_K \mathbf{S}^{-1} \mathbf{u}_h \, d\mathbf{x} + |K| \nabla \tilde{\lambda}_h|_K = 0 \quad \forall K \in \mathcal{T}_h.$$

Next we notice that the stiffness matrix does not depend on q and hence we can pose $q = 0$. Considering $\phi_h = 1$ on K and zero otherwise in (1.75b), this yields $d_K = 0$ for all $K \in \mathcal{T}_h$. Hence $\mathbf{u}_h|_K = -\mathbf{S}_K \nabla \tilde{\lambda}_h|_K$ with \mathbf{S}_K given by (1.74). It now suffices to substitute this into (1.75c) to obtain a system for the Lagrange multipliers λ_D , $D \in \mathcal{D}_h^{\text{int}}$, with the matrix given by (1.73). \square

Lemma 1.8.2. *Let us consider the function $B(s)$, $s \in \mathbb{R}$, $B(s) = \beta(s)s - \int_0^s \beta(\tau) \, d\tau$, with β satisfying Assumption (A1). Then $B(s) \geq s^2 c_\beta / 2$ for all $s \in \mathbb{R}$.*

PROOF:

Let us first consider a given $s \geq 0$. We then have for each $h > 0$

$$\frac{B(s+h) - B(s)}{h} = \frac{\beta(s+h) - \beta(s)}{h} s + \beta(s+h) - \frac{1}{h} \int_s^{s+h} \beta(\tau) \, d\tau.$$

This gives, using that $\beta(s+h) - \beta(s) \geq c_\beta h$, which follows from Assumption (A1), and the continuity of β

$$\liminf_{h \rightarrow 0^+} \frac{B(s+h) - B(s)}{h} \geq c_\beta s.$$

Hence, using the fact that $B(0) = 0$ and that $s^2 c_\beta / 2 = 0$ for $s = 0$, we have $B(s) \geq s^2 c_\beta / 2$ for all $s \geq 0$. The proof for $s < 0$ proceeds similarly. \square

Lemma 1.8.3. *Let β satisfy Assumption (A2). Then $[\beta(s)]^2 \leq 2C_\beta^2 + 4L_\beta^2 P^2 + 4L_\beta^2 s^2$ for all $s \in \mathbb{R}$.*

PROOF:

If $s \in [-P, P]$, the assertion of the lemma is trivially satisfied, since by Assumption (A2), $|\beta(s)| \leq C_\beta$. If $s > P$, then using the Lipschitz continuity of β on $[P, +\infty)$, one has

$$\beta(s) = \beta(P) + \beta(s) - \beta(P) \leq \beta(P) + L_\beta(s - P)$$

and similarly for $s < -P$. Thus, using the inequality $(a \pm b)^2 \leq 2(a^2 + b^2)$ and $|\beta(\pm P)| \leq C_\beta$, one has, for $|s| > P$,

$$[\beta(s)]^2 \leq 2C_\beta^2 + 4L_\beta^2 P^2 + 4L_\beta^2 s^2. \quad \square$$

Lemma 1.8.4. *Let $\Omega \subset \mathbb{R}^p$, $p > 1$, be an open bounded set, $\{a_n, n \in \mathbb{N}\}$ a sequence of functions from $L^2(\Omega)$, defined by zero on $\mathbb{R}^p \setminus \Omega$, h_n a sequence of non-negative real values with $\lim_{n \rightarrow \infty} h_n = 0$, and $C > 0$. Let the functions a_n satisfy*

$$\int_{\Omega} \left(a_n(\mathbf{x} + \boldsymbol{\eta}) - a_n(\mathbf{x}) \right)^2 d\mathbf{x} \leq C|\boldsymbol{\eta}| + h_n \quad \forall \boldsymbol{\eta} \in \mathbb{R}^p, \forall n \in \mathbb{N}. \quad (1.76)$$

Then

$$\forall \varepsilon > 0 \quad \exists \zeta > 0 \quad \forall \boldsymbol{\eta} \in \mathbb{R}^p, |\boldsymbol{\eta}| < \zeta \quad \forall n \in \mathbb{N} \quad \int_{\Omega} \left(a_n(\mathbf{x} + \boldsymbol{\eta}) - a_n(\mathbf{x}) \right)^2 d\mathbf{x} \leq \varepsilon. \quad (1.77)$$

PROOF:

Let us consider a fixed $\varepsilon > 0$. Let n_0 be such that $\forall n > n_0$, $|h_n| < \varepsilon/2$. The continuity in mean of the functions a_1, \dots, a_{n_0} implies

$$\int_{\mathbb{R}^p} \left(a_i(\mathbf{x} + \boldsymbol{\eta}) - a_i(\mathbf{x}) \right)^2 d\mathbf{x} \longrightarrow 0 \quad \text{as } |\boldsymbol{\eta}| \rightarrow 0 \quad \forall i \in \{1, \dots, n_0\},$$

or, more precisely,

$$\forall i \in \{1, \dots, n_0\} \quad \forall \varepsilon^* > 0 \quad \exists \zeta_i^* > 0 \quad \forall \boldsymbol{\eta}^* \in \mathbb{R}^p, |\boldsymbol{\eta}^*| < \zeta_i^* \\ \int_{\mathbb{R}^p} \left(a_i(\mathbf{x} + \boldsymbol{\eta}^*) - a_i(\mathbf{x}) \right)^2 d\mathbf{x} \leq \varepsilon^*. \quad (1.78)$$

We set $\varepsilon^* = \varepsilon$ in (1.78) and define $\zeta^* := \min_{i=1, \dots, n_0} \zeta_i^*$. Since $n_0 < \infty$, $\zeta^* > 0$. It is finally enough to choose

$$\zeta = \min \left\{ \zeta^*, \frac{\varepsilon}{2C} \right\}.$$

Indeed, for $n < n_0$, estimate (1.77) is valid due to (1.78). For $n > n_0$, (1.76) and the fact that $|h_n| < \varepsilon/2$ yields the assertion of the lemma. \square

Lemma 1.8.5. *Let $\Omega \subset \mathbb{R}^p$, $p \geq 1$, be an open bounded set and let $d \geq 1$. Let the vector-valued functions $\mathbf{u}^n \in [L^2(\Omega)]^d$ such that $\int_{\Omega} \mathbf{u}^n \cdot \mathbf{u}^n d\mathbf{x} \leq U^2$, $U > 0$, converge weakly in $[L^2(\Omega)]^d$ to some function $\mathbf{u} \in [L^2(\Omega)]^d$. Let the matrix-valued functions $\mathbf{M}^n, \mathbf{M}_{i,j}^n \in L^2(\Omega)$, $1 \leq i, j \leq d$, converge elementwise strongly in $L^2(\Omega)$ to some function \mathbf{M} . Then*

$$\int_{\Omega} \mathbf{M}^n \mathbf{u}^n \cdot \boldsymbol{\psi} d\mathbf{x} \rightarrow \int_{\Omega} \mathbf{M} \mathbf{u} \cdot \boldsymbol{\psi} d\mathbf{x}$$

for all $\boldsymbol{\psi} \in [C(\overline{\Omega})]^d$.

PROOF:

We have

$$\begin{aligned} \int_{\Omega} (\mathbf{M}^n \mathbf{u}^n - \mathbf{M} \mathbf{u}) \cdot \psi \, d\mathbf{x} &= \sum_{i=1}^d \sum_{j=1}^d \int_{\Omega} (\mathbf{M}_{i,j}^n \mathbf{u}_j^n - \mathbf{M}_{i,j} \mathbf{u}_j) \psi_i \, d\mathbf{x} \\ &= \sum_{i=1}^d \sum_{j=1}^d \left(\int_{\Omega} (\mathbf{M}_{i,j}^n - \mathbf{M}_{i,j}) \mathbf{u}_j^n \psi_i \, d\mathbf{x} + \int_{\Omega} (\mathbf{u}_j^n - \mathbf{u}_j) \mathbf{M}_{i,j} \psi_i \, d\mathbf{x} \right). \end{aligned}$$

The first term is bounded by

$$C_{\psi} \sum_{i=1}^d \sum_{j=1}^d \left(\int_{\Omega} (\mathbf{M}_{i,j}^n - \mathbf{M}_{i,j})^2 \, d\mathbf{x} \right)^{\frac{1}{2}} \left(\int_{\Omega} (\mathbf{u}_j^n)^2 \, d\mathbf{x} \right)^{\frac{1}{2}} \leq C_{\psi} U \sum_{i=1}^d \sum_{j=1}^d \left(\int_{\Omega} (\mathbf{M}_{i,j}^n - \mathbf{M}_{i,j})^2 \, d\mathbf{x} \right)^{\frac{1}{2}},$$

using the fact that $|\psi_i| \leq C_{\psi}$, $1 \leq i \leq d$, $C_{\psi} > 0$, and the Cauchy–Schwarz inequality, and thus converges to zero using the strong convergence of $\mathbf{M}_{i,j}^n$ to $\mathbf{M}_{i,j}$, $1 \leq i, j \leq d$. The second term converges to zero by the definition of the weak convergence of \mathbf{u}^n to \mathbf{u} . \square

1.9 Appendix B: A combined finite volume–finite element scheme for contaminant transport simulation on nonmatching grids

We present in this appendix a variant of the scheme from the first part of the chapter for nonmatching grids and apply it to contaminant transport simulation in porous media.

1.9.1 Introduction

We consider in this appendix the equation (1.1) in its precise form describing the reactive miscible displacement of one contaminant in porous media. We suppose that a domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is discretized into a nonoverlapping nonmatching grid possibly containing nonconvex elements, as that given in Figure 1.6 below by the dashed line.

The discretization methods for nonmatching grids represent a very active area of research. They are usually proposed in the context of domain decomposition methods, cf. [102]. The mortar method was developed for elliptic problems discretized by the finite element or spectral methods in [26]. This approach has been later extended to mixed finite element methods [11, 120], finite volume element methods [55], and cell-centered finite volume methods [3, 66]. A nonconsistent but simple, stable, and efficient (see the comparative tests in [66]) cell-centered finite volume scheme for nonoverlapping nonmatching grids has been proposed in [36].

We apply here the ideas of combined finite volume–finite element schemes (cf. [67, 114]) and in particular a variant of the scheme proposed and studied in the first part of this chapter to the discretization of (1.1) on the given grids. We are motivated by the following consideration: the mesh can be nonmatching and can contain nonconvex control volumes for a pure cell-centered finite volume discretization of the equation (1.1) without the diffusion term, cf. [61, Chapter VI]. The mesh is required to match along hyperplanes of \mathbb{R}^d and to consist of convex control volumes only when the diffusion term is present, cf. [61, Chapter III]. We however notice that given a set of points, we can always construct a simplicial mesh (consisting of triangles if $d = 2$ and of tetrahedra if $d = 3$) with vertices given by this set of points. Hence an intuitive idea

is as follows: given a nonmatching mesh with possibly nonconvex elements and with a set of points associated with these elements, construct a simplicial mesh having this set of points as vertices. Then consider a finite element discretization of the diffusion term of (1.1) on the simplicial mesh and a finite volume discretization of the other terms of (1.1) on the original mesh.

We believe that the proposed approach to the discretization of the equation (1.1) on nonmatching grids is in some sense the simplest, yet (at least in our opinion) very efficient. In particular, we do not introduce any supplementary equations or unknowns on the boundary between the regions with nonmatching grids, nor do we use any interpolation of the discrete solutions on this boundary. There is no need for this for the finite volume part and the finite element part uses a conforming mesh. The proposed scheme is similar to that from [36]. The essential difference is that we replace the finite volume diffusion fluxes by the finite element ones. This is very important in the present case, since the diffusion fluxes through the interfaces between the subdomains with nonmatching grids of the scheme proposed in [36] are not consistent, whereas our discrete diffusion fluxes are consistent. Next, the scheme is stable since we avoid spurious oscillations in the convection-dominated case by checking the local Péclet number and by adding exactly the necessary amount of upstream weighting and it possesses a discrete maximum principle under some conditions on the simplicial mesh and on the tensor \mathbf{S} . The scheme can finally be easily implemented in any finite volume code, in order to permit a (nonmatching) local refinement of the mesh and an easy discretization of inhomogeneous and anisotropic tensors, a highly desirable feature in contaminant transport modeling. This was in fact our original motivation.

This appendix is organized as follows. We describe in Section 1.9.2 the problem of reactive transport with equilibrium adsorption in porous media. We propose in Section 1.9.3 a combined finite volume–finite element scheme with the backward Euler finite difference time stepping for the discretization of this problem. We prove in Section 1.9.4 the local conservativity of the scheme and the discrete maximum principle under appropriate conditions on the simplicial mesh and on the tensor \mathbf{S} . Its convergence could be proved using the techniques from the first part of this chapter. Finally, in Section 1.9.5 we demonstrate the performances of the proposed scheme on some model as well as real problems and in Section 1.9.6 we give some concluding remarks.

1.9.2 The contaminant transport problem

Let $(0, T)$ be a time interval, $0 < T < +\infty$. We consider a reactive miscible displacement with equilibrium adsorption of one contaminant in Ω , described by

$$\frac{\partial(\theta c)}{\partial t} + \rho_b \frac{\partial w(c)}{\partial t} - \nabla \cdot (\mathbf{S} \nabla c) + \nabla \cdot (c \mathbf{v}) + \lambda(\theta c + \rho_b w(c)) - q_{\text{out}} c = q_{\text{in}} c_s \quad \text{in } \Omega \times (0, T), \quad (1.79a)$$

$$c(\cdot, 0) = c_0 \quad \text{in } \Omega, \quad (1.79b)$$

$$c = g \quad \text{on } \partial\Omega \times (0, T). \quad (1.79c)$$

The problem (1.79a)–(1.79c) falls into the frame of the problem (1.1)–(1.3) studied in the first part of this chapter, with an additional sink term and an inhomogeneous Dirichlet boundary condition. Neumann or Robin boundary conditions can also be considered. In (1.79a)–(1.79c) $c = c(\mathbf{x}, t)$ is the unknown concentration of the dissolved contaminant ($[\text{ML}^{-d}]$), $\theta = \theta(\mathbf{x}, t)$ is the water content ($[-]$) (we shall hereafter denote by θ_s the saturated water content and by θ_r the residual water content), $\rho_b = \rho_b(\mathbf{x})$ is the bulk density of the porous medium

([ML^{-d}]), and $w : \mathbb{R} \rightarrow \mathbb{R}$ is the equilibrium adsorption function. We suppose that adsorption is sufficiently fast in comparison with the speed of the displacement of the contaminant so that the concentration of the dissolved contaminant c and the concentration ratio of the immobilized contaminant $w(c)$ ([$-$]) are in equilibrium. In particular, we shall consider, for $c \geq 0$, $w(c) = \mu_1 c^{\mu_2}$, where μ_1 and μ_2 are positive constants, in the case of the Freundlich isotherm and $w(c) = \nu_1 \nu_2 c / (1 + \nu_2 c)$, where ν_1 and ν_2 are positive constants, in the case of the Langmuir isotherm. We suppose that the velocity field $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$ ([LT⁻¹]) is given by the Darcy law

$$\mathbf{v} = -\mathbf{K}(p)\nabla(p + z), \quad (1.80)$$

where $\mathbf{K} = \mathbf{K}(\mathbf{x}, p)$ is the hydraulic conductivity tensor ([LT⁻¹]), $z = z(\mathbf{x})$ is the elevation, the upward vertical coordinate ([L]), and $p = p(\mathbf{x}, t)$, the pressure head ([L]), is the solution of the Richards problem, which describes two-phase water–air flow in the subsurface,

$$\frac{\partial \theta(p)}{\partial t} - \nabla \cdot \mathbf{K}(p)\nabla(p + z) = q_{\text{out}} + q_{\text{in}} \quad \text{in } \Omega \times (0, T), \quad (1.81a)$$

$$p(\cdot, 0) = p_0 \quad \text{in } \Omega, \quad (1.81b)$$

$$p = p_D \quad \text{on } \Gamma_D \times (0, T), \quad (1.81c)$$

$$-\mathbf{K}(p)\nabla(p + z) \cdot \mathbf{n} = u_N \quad \text{on } \Gamma_N \times (0, T). \quad (1.81d)$$

Here $\overline{\Gamma_D} \cup \overline{\Gamma_N} = \partial\Omega$, $\Gamma_D \cap \Gamma_N = \emptyset$, $|\Gamma_D| \neq 0$. The dependence of θ and \mathbf{K} on p is given for example by the van Genuchten law, see [118]. We suppose that the diffusion–dispersion tensor $\mathbf{S} = \mathbf{S}(\mathbf{x}, \mathbf{v})$ ([L²T⁻¹]) is given by

$$\begin{aligned} \mathbf{S}_{ii} &= \alpha_T |\mathbf{v}| + (\alpha_L - \alpha_T) \frac{v_i^2}{|\mathbf{v}|} + \sigma \quad i = 1, \dots, d, \\ \mathbf{S}_{ij} = \mathbf{S}_{ji} &= (\alpha_L - \alpha_T) \frac{v_i v_j}{|\mathbf{v}|} \quad i, j = 1, \dots, d, \end{aligned}$$

where v_i are the components of the velocity vector \mathbf{v} and $|\mathbf{v}|$ is its length, $\alpha_L = \alpha_L(\mathbf{x})$ is the longitudinal dispersivity ([L]), $\alpha_T = \alpha_T(\mathbf{x})$ is the transverse dispersivity ([L]), and finally $\sigma = \sigma(\mathbf{x})$ is the molecular diffusion coefficient ([L²T⁻¹]). We consider first-order irreversible reactions such as radioactive decay, hydrolysis, and some forms of biodegradation, where λ is the reaction rate constant ([T⁻¹]). Finally, $q_{\text{in}} = q_{\text{in}}(\mathbf{x}, t)$, $q_{\text{in}} \geq 0$, denotes the sources per unit volume ([T⁻¹]). In the case of a source, we have to specify the concentration of the entering dissolved contaminant c_s . In contrast, the concentration of the leaving dissolved contaminant due to the sinks per unit volume $q_{\text{out}} = q_{\text{out}}(\mathbf{x}, t)$ ([T⁻¹]), $q_{\text{out}} \leq 0$, is given by the unknown concentration c . We refer to [19, 25, 79, 123] for more details.

1.9.3 Combined finite volume–finite element scheme

We define in this section the space and time discretizations and introduce the combined finite volume–finite element scheme.

Space and time discretizations

We suppose a generally nonconstant time step for the time discretization. We split up the time interval $(0, T)$ such that $0 = t_0 < \dots < t_n < \dots < t_N = T$ and define $\Delta t_n := t_n - t_{n-1}$, $n \in \{1, 2, \dots, N\}$. We next describe the space discretization.

As a *primal grid* of Ω , we understand a partition \mathcal{D}_h of Ω into closed polygons such that $\overline{\Omega} = \bigcup_{D \in \mathcal{D}_h} D$ and such that the intersection of interiors of two different polygons is empty. We

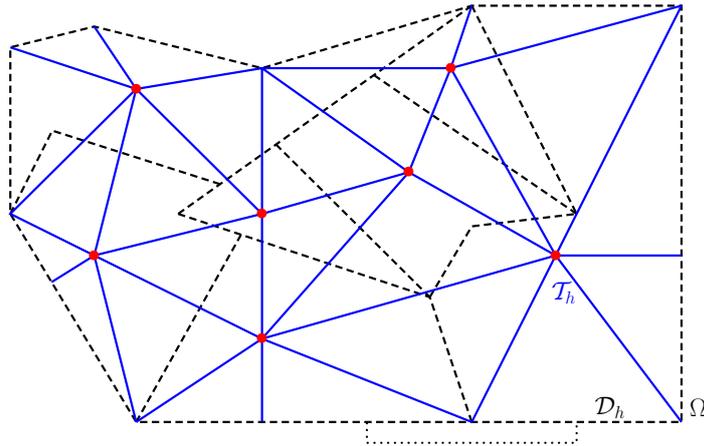


Figure 1.6: Primal nonmatching grid \mathcal{D}_h (dashed) and dual triangular grid \mathcal{T}_h (solid)

in particular admit nonmatching grids, i.e. the case where there exist two different polygons $D, E \in \mathcal{D}_h$ such that their intersection is not an empty set but it is not a common vertex, edge, or side (edge if $d = 2$, face if $d = 3$) of D and E . An example of an admissible primal grid is given in Figure 1.6 by the dashed line. We suppose that there exists a family of points \mathcal{P}_h such that there is one point V_D in the interior of D associated with each $D \in \mathcal{D}_h$.

A *dual grid* of Ω is a partition \mathcal{T}_h of Ω into closed simplices which satisfies the following properties: (i) The set of points \mathcal{P}_h is contained in the set of vertices of \mathcal{T}_h , denoted by \mathcal{V}_h ; (ii) The vertices from $\mathcal{V}_h \setminus \mathcal{P}_h$ lie on the boundary of Ω ; (iii) \mathcal{T}_h is conforming, i.e. the intersection of two different simplices is either an empty set or their common vertex, edge, or face; (iv) $\bar{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$. This definition is not unique: we have a choice in connecting the different points $V_D \in \mathcal{P}_h$ and also a choice in the definition of the vertices on the boundary. The general intention is to find a triangulation such that the transmissibilities $\mathbb{S}_{D,E}^n$ defined below by (1.83) were non-negative, since this implies the discrete maximum principle, cf. Theorem 1.9.4 and Remarks 1.9.5 and 1.9.6 below. An example of a dual grid to a primal nonmatching grid is given in Figure 1.6 by the solid line.

In order to simplify the notation in the next sections, we define still a *fictitious boundary grid* $\mathcal{D}_h^{\text{ext}}$. We associate a fictitious control volume D with each vertex $V \in \mathcal{V}_h$ lying on the boundary $\partial\Omega$. We define D in such way that $D \cap \Omega = \emptyset$, $D \cap \bar{\Omega} \subset \partial\Omega$, and $V \in D \cap \bar{\Omega}$. We finally require that the boundaries of D , $D \in \mathcal{D}_h^{\text{ext}}$, halve the segments of $\partial\Omega$ between the boundary vertices, so that $\bigcup_{D \in \mathcal{D}_h^{\text{ext}}} \{D \cap \bar{\Omega}\} = \partial\Omega$. An example of $D \in \mathcal{D}_h^{\text{ext}}$ is given in Figure 1.6 by the dotted line. We shall use the notation V_D for the vertex associated with $D \in \mathcal{D}_h^{\text{ext}}$, as for the vertices from \mathcal{P}_h and control volumes from \mathcal{D}_h .

We next denote by $\mathcal{N}(D)$ the set of all neighbors of a control volume $D \in \mathcal{D}_h$, i.e. the set of $E \in \mathcal{D}_h \cup \mathcal{D}_h^{\text{ext}}$ such that $D \cap E$ has a positive $(d-1)$ -dimensional measure. In particular, using the above definition of the set $\mathcal{D}_h^{\text{ext}}$, we can easily write the integral over ∂D as $\sum_{E \in \mathcal{N}(D)} \int_{\partial D \cap \partial E} d\gamma(\mathbf{x})$. Similarly, for a vertex $V_D \in \mathcal{P}_h$, we denote by $\mathcal{M}(V_D)$ the set of all vertices $V_E \in \mathcal{V}_h$ such that there exists an edge between V_D and V_E .

The combined scheme

The combined scheme is obtained by the discretization of the diffusion term of (1.79a) by means of the piecewise linear conforming finite element method on \mathcal{T}_h , the discretization of

the other terms of (1.79a) by means of the cell-centered finite volume method on \mathcal{D}_h , and using a finite difference time stepping.

Definition 1.9.1. (Combined scheme) *The fully implicit combined finite volume–finite element scheme for the problem (1.79a)–(1.79c) reads: find the values c_D^n , $D \in \mathcal{D}_h$, $n \in \{0, 1, \dots, N\}$, such that*

$$c_D^0 = \frac{1}{|D|} \int_D c_0(\mathbf{x}) \, d\mathbf{x} \quad D \in \mathcal{D}_h, \quad (1.82a)$$

$$c_D^n = g(V_D, t_n) \quad D \in \mathcal{D}_h^{\text{ext}}, \, n \in \{1, 2, \dots, N\}, \quad (1.82b)$$

$$\begin{aligned} & \frac{\theta_D^n c_D^n - \theta_D^{n-1} c_D^{n-1}}{\Delta t_n} |D| + (\rho_b)_D \frac{w(c_D^n) - w(c_D^{n-1})}{\Delta t_n} |D| - \sum_{V_E \in \mathcal{M}(V_D)} \mathbb{S}_{D,E}^n (c_E^n - c_D^n) \\ & + \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \overline{c_{D,E}^n} + \lambda[\theta_D^n c_D^n + (\rho_b)_D w(c_D^n)] |D| - (q_{\text{out}})_D^n c_D^n |D| = (q_{\text{in}} c_s)_D^n |D| \\ & D \in \mathcal{D}_h, \, n \in \{1, 2, \dots, N\}. \end{aligned} \quad (1.82c)$$

In the above definition we have used

$$\theta_D^n := \frac{1}{|D|} \int_D \theta(\mathbf{x}, t_n) \, d\mathbf{x} \quad D \in \mathcal{D}_h, \, n \in \{0, 1, \dots, N\},$$

$$(\rho_b)_D := \frac{1}{|D|} \int_D \rho_b(\mathbf{x}) \, d\mathbf{x} \quad D \in \mathcal{D}_h,$$

$$(q_{\text{out}})_D^n := \frac{1}{\Delta t_n |D|} \int_{t_{n-1}}^{t_n} \int_D q_{\text{out}}(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad D \in \mathcal{D}_h, \, n \in \{1, 2, \dots, N\},$$

$$(q_{\text{in}} c_s)_D^n := \frac{1}{\Delta t_n |D|} \int_{t_{n-1}}^{t_n} \int_D q_{\text{in}}(\mathbf{x}, t) c_s(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad D \in \mathcal{D}_h, \, n \in \{1, 2, \dots, N\}$$

and we have denoted the flux of \mathbf{v} between $D \in \mathcal{D}_h$ and $E \in \mathcal{N}(D)$ for $n \in \{1, 2, \dots, N\}$ by

$$\mathbf{v}_{D,E}^n := \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \int_{\partial D \cap \partial E} \mathbf{v}(\mathbf{x}, t) \cdot \mathbf{n}_{D,E} \, d\gamma(\mathbf{x}) \, dt,$$

where $\mathbf{n}_{D,E}$ is the unit normal vector of the side $\partial D \cap \partial E$ between D and E , outward to D . For the notational convenience, we define $\mathbf{v}_{D,E}^n$ by 0 if $E \notin \mathcal{N}(D)$. We suppose that the functions g and θ are sufficiently smooth in order to define c_D^n , $D \in \mathcal{D}_h^{\text{ext}}$, and θ_D^n . In analogy with the first part of this chapter, we first define

$$\tilde{\mathbf{S}}^n(\mathbf{x}) := \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \mathbf{S}(\mathbf{x}, t) \, dt \quad \mathbf{x} \in \Omega, \, n \in \{1, 2, \dots, N\}.$$

The transmissibility between V_D and V_E , $D \in \mathcal{D}_h$, $E \in \mathcal{D}_h \cup \mathcal{D}_h^{\text{ext}}$, is then given by

$$\mathbb{S}_{D,E}^n := - \int_{\Omega} \mathbf{S}^n \nabla \varphi_E \cdot \nabla \varphi_D \, d\mathbf{x} \quad n \in \{1, 2, \dots, N\}, \quad (1.83)$$

where we have again two choices of the definition of \mathbf{S}^n . We can either use directly $\mathbf{S}^n = \tilde{\mathbf{S}}^n$, or define a piecewise constant tensor

$$\mathbf{S}^n(\mathbf{y}) = \left(\frac{1}{|K|} \int_K [\tilde{\mathbf{S}}^n(\mathbf{x})]^{-1} \, d\mathbf{x} \right)^{-1} \quad \mathbf{y} \in K, \, K \in \mathcal{T}_h, \, n \in \{1, 2, \dots, N\}.$$

These two choices correspond, respectively, to the arithmetic or harmonic average of the diffusion–dispersion tensor.

Finally, again in analogy with the first part of this chapter, we define the value $\overline{c_{D,E}^n}$ for $D \in \mathcal{D}_h$, $E \in \mathcal{N}(D)$, and $n \in \{1, 2, \dots, N\}$ as follows:

$$\overline{c_{D,E}^n} := \begin{cases} c_D^n + \alpha_{D,E}^n(c_E^n - c_D^n) & \text{if } \mathbf{v}_{D,E}^n \geq 0 \\ c_E^n + \alpha_{D,E}^n(c_D^n - c_E^n) & \text{if } \mathbf{v}_{D,E}^n < 0 \end{cases}.$$

Here $\alpha_{D,E}^n$ is the coefficient of the amount of upstream weighting which is defined by

$$\alpha_{D,E}^n := \frac{\max\left\{\min\left\{\mathbb{S}_{D,E}^n, \frac{1}{2}|\mathbf{v}_{D,E}^n|\right\}, 0\right\}}{|\mathbf{v}_{D,E}^n|}, \quad \mathbf{v}_{D,E}^n \neq 0. \quad (1.84)$$

We set $\alpha_{D,E}^n := 0$ if $\mathbf{v}_{D,E}^n = 0$. We remark that in the scheme studied in the first part of this chapter, there can be nonzero convective and diffusive fluxes between D and E only if D and E neighbors. This is however not the case with the scheme (1.82a)–(1.82c): there can be nonzero convective flux between D and E only if D and E neighbors, but there can be nonzero diffusive flux between D and E even if D and E are not neighbors (because the transmissibility between D and E is given by the grid \mathcal{T}_h). However, as we shall see in Theorem 1.9.4, the local Péclet upstream weighting still guarantees, adding minimal numerical diffusion, the stability of the scheme.

1.9.4 Discrete properties of the scheme

We show in this section two essential properties of the scheme proposed in this appendix. The ideas follow those introduced in [61] for finite volume schemes and extended onto combined finite volume–finite element schemes in the first part of this chapter.

Theorem 1.9.2. (Conservativity of the scheme) *The scheme (1.82a)–(1.82c) is conservative with respect to the primal mesh \mathcal{D}_h .*

PROOF:

Let us take two fixed dual volumes $E \in \mathcal{D}_h$ and $D \in \mathcal{D}_h$. The discrete diffusive flux from D to E is given by $-\mathbb{S}_{D,E}^n(c_E^n - c_D^n)$. The discrete diffusive flux from E to D is given by $-\mathbb{S}_{E,D}^n(c_D^n - c_E^n)$, i.e. we have their equality up to the sign, considering that $\mathbb{S}_{D,E}^n = \mathbb{S}_{E,D}^n$ for all $n \in \{1, 2, \dots, N\}$, which follows from (1.83) using the symmetry of the tensor \mathbf{S} .

For the discrete convective flux from D to E , we have $\mathbf{v}_{D,E}^n[c_D^n + \alpha_{D,E}^n(c_E^n - c_D^n)]$, supposing $\mathbf{v}_{D,E}^n \geq 0$. For the discrete convective flux from E to D , we have $\mathbf{v}_{E,D}^n[c_D^n + \alpha_{E,D}^n(c_E^n - c_D^n)]$, i.e. again the equality up to the sign, considering that $\mathbf{v}_{D,E}^n = -\mathbf{v}_{E,D}^n$ and that $\alpha_{D,E}^n = \alpha_{E,D}^n$, which follows from $\mathbb{S}_{D,E}^n = \mathbb{S}_{E,D}^n$. For $\mathbf{v}_{D,E}^n < 0$, the proof is similar. Hence the combined finite volume–finite element scheme is conservative as the pure finite volume is, cf. [61]. \square

It follows from (1.80) and (1.81a) that $\nabla \cdot \mathbf{v} = q_{\text{out}} + q_{\text{in}} - \partial\theta/\partial t$. This property implies the following lemma:

Lemma 1.9.3. *For all $D \in \mathcal{D}_h$ and $n \in \{1, 2, \dots, N\}$,*

$$\begin{aligned} \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \widehat{c_{D,E}^n} &= \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) + (q_{\text{in}})_D^n c_D^n |D| \\ &\quad + (q_{\text{out}})_D^n c_D^n |D| - \left(\frac{\theta_D^n - \theta_D^{n-1}}{\Delta t_n} \right) c_D^n |D|, \end{aligned}$$

where $(\mathbf{v}_{D,E}^n)^- := \min\{\mathbf{v}_{D,E}^n, 0\}$ and where the definition of $(q_{\text{in}})_D^n$ is as that of $(q_{\text{out}})_D^n$.

PROOF:

Considering that $\mathbf{v}_{D,E}^n = (\mathbf{v}_{D,E}^n)^+ + (\mathbf{v}_{D,E}^n)^-$, where $(\mathbf{v}_{D,E}^n)^+ := \max\{\mathbf{v}_{D,E}^n, 0\}$, and the definition of the upstream weighting, we have

$$\begin{aligned}
 \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n \widehat{c_{D,E}^n} &= \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^+ c_D^n + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- c_E^n \\
 &= \sum_{E \in \mathcal{N}(D)} \mathbf{v}_{D,E}^n c_D^n + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) \\
 &= c_D^n \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \int_D \nabla \cdot \mathbf{v}(\mathbf{x}, t) \, d\mathbf{x} \, dt + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) \\
 &= c_D^n (q_{\text{in}})_D^n |D| + c_D^n (q_{\text{out}})_D^n |D| + \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) \\
 &\quad - \left(\frac{\theta_D^n - \theta_D^{n-1}}{\Delta t_n} \right) c_D^n |D|. \quad \square
 \end{aligned}$$

We now give an important theorem, guaranteeing that under certain conditions on \mathcal{T}_h and \mathbf{S} , we obtain physically correct results.

Theorem 1.9.4. (Discrete maximum principle) *Let $\mathbb{S}_{D,E}^n \geq 0$ for all $D \in \mathcal{D}_h$, $V_E \in \mathcal{M}(V_D)$, and all $n \in \{1, 2, \dots, N\}$. Let the initial, sources, and Dirichlet boundary concentrations satisfy $0 \leq c_0 \leq M$, $0 \leq c_s \leq M$, and $0 \leq g \leq M$, respectively. Let finally the adsorption function w be nondecreasing and such that $w(0) = 0$ and let $\lambda \geq 0$. Then the solution of the problem (1.82a)–(1.82c) satisfies*

$$0 \leq c_D^n \leq M$$

for all $D \in \mathcal{D}_h$, $n \in \{1, 2, \dots, N\}$.

PROOF:

Setting $\mathbb{T}_{D,E}^n := \mathbb{S}_{D,E}^n - |\mathbf{v}_{D,E}^n| \alpha_{D,E}^n$, $D \in \mathcal{D}_h$, $E \in \mathcal{L}(D)$, where $E \in \mathcal{L}(D)$ if $E \in \mathcal{N}(D)$ or if $V_E \in \mathcal{M}(V_D)$, and using Lemma 1.9.3, we can rewrite the scheme (1.82a)–(1.82c) as

$$\begin{aligned}
 &\frac{\theta_D^{n-1} c_D^n - \theta_D^{n-1} c_D^{n-1}}{\Delta t_n} |D| + (\rho_b)_D \frac{w(c_D^n) - w(c_D^{n-1})}{\Delta t_n} |D| - \sum_{E \in \mathcal{L}(D)} \mathbb{T}_{D,E}^n (c_E^n - c_D^n) \\
 &+ \sum_{E \in \mathcal{N}(D)} (\mathbf{v}_{D,E}^n)^- (c_E^n - c_D^n) + \lambda [\theta_D^n c_D^n + (\rho_b)_D w(c_D^n)] |D| = (q_{\text{in}} c_s)_D^n |D| - c_D^n (q_{\text{in}})_D^n |D| \\
 &D \in \mathcal{D}_h, n \in \{1, 2, \dots, N\}.
 \end{aligned}$$

In view of the definition of $\mathbb{T}_{D,E}^n$ and of (1.84), one has $\mathbb{T}_{D,E}^n \geq 0$ for all $D \in \mathcal{D}_h$, $E \in \mathcal{L}(D)$, and $n \in \{1, 2, \dots, N\}$. We now make use of an induction argument. We remark that $0 \leq c_D^n \leq M$ is satisfied for $n = 0$, using the assumption on c_0 . Let us suppose that $0 \leq c_D^{n-1} \leq M$ for all $D \in \mathcal{D}_h$ for a fixed $(n-1) \in \{0, 1, \dots, N-1\}$. Since the set $\mathcal{D}_h \cup \mathcal{D}_h^{\text{ext}}$ is finite, there exist $D_0, D_1 \in \mathcal{D}_h \cup \mathcal{D}_h^{\text{ext}}$ such that $c_{D_0}^n \leq c_D^n \leq c_{D_1}^n$ for all $D \in \mathcal{D}_h \cup \mathcal{D}_h^{\text{ext}}$. Using a contradiction argument we prove below that $c_{D_0}^n \geq 0$ and $c_{D_1}^n \leq M$. Suppose that $c_{D_0}^n < 0$. We remark that $D_0 \in \mathcal{D}_h$, using the assumption on the Dirichlet boundary condition g . Then, since $\mathbb{T}_{D_0,E}^n \geq 0$ and $-(\mathbf{v}_{D_0,E}^n)^- \geq 0$, we have

$$\sum_{E \in \mathcal{L}(D_0)} \mathbb{T}_{D_0,E}^n (c_E^n - c_{D_0}^n) + \sum_{E \in \mathcal{N}(D_0)} -(\mathbf{v}_{D_0,E}^n)^- (c_E^n - c_{D_0}^n) \geq 0.$$

This yields

$$\begin{aligned} & \frac{\theta_{D_0}^{n-1} c_{D_0}^n - \theta_{D_0}^{n-1} c_{D_0}^{n-1}}{\Delta t_n} |D_0| + (\rho_b)_{D_0} \frac{w(c_{D_0}^n) - w(c_{D_0}^{n-1})}{\Delta t_n} |D_0| \\ & + \lambda [\theta_{D_0}^n c_{D_0}^n + (\rho_b)_{D_0} w(c_{D_0}^n)] |D_0| - (q_{\text{in}} c_s)_{D_0}^n |D_0| + c_{D_0}^n (q_{\text{in}})_{D_0}^n |D_0| \geq 0. \end{aligned}$$

Now $c_{D_0}^n < 0$ implies $c_{D_0}^n (q_{\text{in}})_{D_0}^n \leq 0$ and one also has $-(q_{\text{in}} c_s)_{D_0}^n \leq 0$. Using the fact that $\theta \geq \theta_s > 0$ and that w is nondecreasing and satisfies $w(0) = 0$, $\theta_{D_0}^n c_{D_0}^n + (\rho_b)_{D_0} w(c_{D_0}^n) \leq 0$ follows. Finally, $w(c_{D_0}^n) - w(c_{D_0}^{n-1}) \leq 0$, using that w is nondecreasing. Thus $c_{D_0}^n \geq c_{D_0}^{n-1}$, which is a contradiction.

Let us now suppose that $c_{D_1}^n > M$. In view of the Dirichlet boundary condition, D_1 again necessarily lies in \mathcal{D}_h . Similarly as in the previous case, one comes to

$$\begin{aligned} & \frac{\theta_{D_1}^{n-1} c_{D_1}^n - \theta_{D_1}^{n-1} c_{D_1}^{n-1}}{\Delta t_n} |D_1| + (\rho_b)_{D_1} \frac{w(c_{D_1}^n) - w(c_{D_1}^{n-1})}{\Delta t_n} |D_1| \\ & + \lambda [\theta_{D_1}^n c_{D_1}^n + (\rho_b)_{D_1} w(c_{D_1}^n)] |D_1| - (q_{\text{in}} c_s)_{D_1}^n |D_1| + c_{D_1}^n (q_{\text{in}})_{D_1}^n |D_1| \leq 0. \end{aligned}$$

We can estimate

$$-(q_{\text{in}} c_s)_{D_1}^n \geq -M (q_{\text{in}})_{D_1}^n \geq -c_{D_1}^n (q_{\text{in}})_{D_1}^n.$$

Simply $\theta_{D_1}^n c_{D_1}^n + (\rho_b)_{D_1} w(c_{D_1}^n) \geq 0$ and $w(c_{D_1}^n) - w(c_{D_1}^{n-1}) \geq 0$. This implies $c_{D_1}^n \leq c_{D_1}^{n-1}$, which is again a contradiction. \square

Remark 1.9.5. (Discrete maximum principle) *We see that the discrete maximum principle holds as soon as the transmissibilities $\mathbb{S}_{D,E}^n$ defined by (1.83) are non-negative. This is e.g. the case, in two space dimensions, when \mathbf{S} reduces to a constant scalar function and when \mathcal{T}_h is Delaunay, that is the circumcircle of each triangle does not contain any vertex in its interior, and under the additional condition that no circumcenters of boundary triangles lie outside the domain, cf. [80, 100]. Remark that given a set of points, we can always construct a Delaunay triangulation. In three space dimensions, however, a Delaunay tetrahedral mesh in general does not guarantee the non-negativity of the finite element transmissibilities, cf. [86, 100]. We refer to Remark 1.9.7 for the modification of the proposed scheme, which guarantees the discrete maximum principle in both two and three space dimensions under the condition that \mathbf{S} is a constant scalar function.*

Remark 1.9.6. (Dual Delaunay triangulation for a locally refined square grid) *Let us consider a locally refined square grid, where a square is refined into 9 subsquares and where the difference of levels of refinement of two neighboring squares is at most one, such as that given in Figure 1.7 by the dashed line. Then an example of a dual Delaunay triangulation is given in Figure 1.7 by the solid line.*

Remark 1.9.7. (A two-grid finite volume scheme verifying the discrete maximum principle for \mathbf{S} constant and scalar) *When \mathbf{S} is a constant scalar function, we can replace the discretization of the diffusion term by the finite element method on a dual simplicial grid by a finite volume discretization on a Voronoï grid given by the points from \mathcal{V}_h . Recall that in two space dimensions, this would lead to the same scheme for the Voronoï mesh dual to a Delaunay triangulation, cf. [61, Section III.12.2]. The interest in it is that in three space dimensions, the finite volume discretization of a Laplacian on a Voronoï grid still leads to positive transmissibilities; compare this with Remark 1.9.5.*

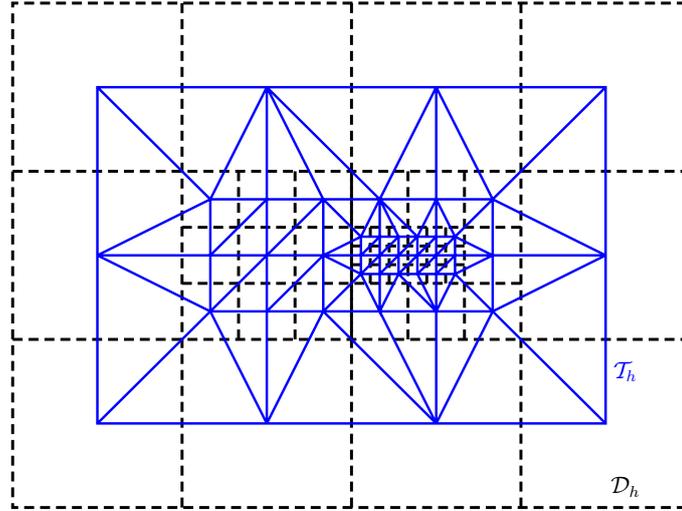


Figure 1.7: Primal locally refined square grid \mathcal{D}_h (dashed) and dual triangular grid \mathcal{T}_h (solid)

1.9.5 Numerical simulations

We present here the results of two numerical experiments in space dimension two.

Model problem with a known analytical solution

The purpose of this problem is to test the proposed scheme on nonmatching grids. Let us consider a linear model problem of the type (1.79a)–(1.79c) with constant coefficients given by

$$\begin{aligned} \theta &= 1, \quad w = 0, \\ \mathbf{S} &= \delta \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{v} = (v_1, v_2), \\ \lambda &= 0, \quad q_{\text{out}} = 0, \quad q_{\text{in}} = 0. \end{aligned}$$

The initial and boundary conditions are given by the exact solution

$$c(x, y, t) = \frac{1}{200\nu t + 1} e^{-50 \frac{(x-x_0-v_1 t)^2 + (y-y_0-v_2 t)^2}{200\nu t + 1}}$$

representing a Gaussian peak starting at the point (x_0, y_0) , being transported by the convective field \mathbf{v} , and diffusing. Let us in particular consider

$$\begin{aligned} \Omega &= (0, 3) \times (0, 3), \quad T = 2, \\ v_1 &= 0.8, \quad v_2 = 0.4, \quad x_0 = 0.5, \quad y_0 = 1.35. \end{aligned}$$

We consider the discretization of the domain Ω into N^2 squares with $N = 10, 20, 40$, and 80 . We shall call these grids in the following as unrefined grids. We next consider local refinements and uniform refinements of these grids, where one square is refined into 9 subsquares. We never refine twice along a given edge. An example of a locally refined grid and the appropriate dual triangular grid is given in Figure 1.7. In view of Remark 1.9.6, the scheme (1.82a)–(1.82c) for the considered problem satisfies the discrete maximum principle. We divide the time interval $(0, T)$ into $1.6N$ time steps and consider two values of the parameter δ : for $\delta = 0.1$, the problem is diffusion-dominated, and for $\delta = 0.001$, the problem is convection-dominated.

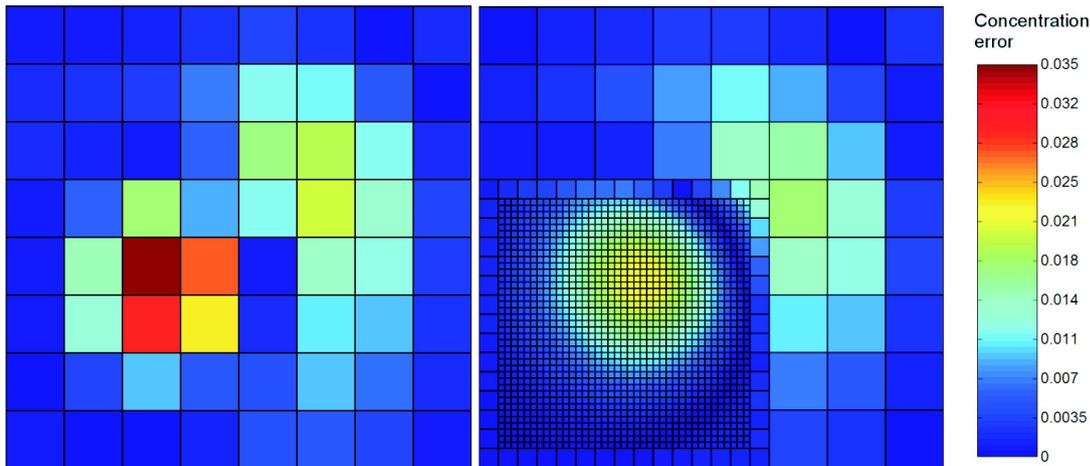


Figure 1.8: Errors of the approximate solutions on unrefined 20×20 (left) and locally refined (right) grids for $\delta = 0.1$ at $t = 3/8$, cut of the domain

N	T. st.	Unkn.	$\ c_h - c\ _{0,h,\Omega}$	Unkn.	$\ c_h - c\ _{0,h,\Omega}$	Unkn.	$\ c_h - c\ _{0,h,\Omega}$
10	16	100	0.01277	316	0.00547	900	0.00491
20	32	400	0.00354	1264	0.00301	3600	0.00270
40	64	1600	0.00173	5056	0.00158	14400	0.00143
80	128	6400	0.00086	20224	0.00080	57600	0.00074

Table 1.2: N, number of time steps, number of unknowns, and discrete $L^2(\Omega)$ errors for $\delta = 0.1$ at $t = 2$ for unrefined, locally refined, and uniformly refined square grids, respectively

N	T. st.	Unkn.	$\ c_h - c\ _{0,h,\Omega}$	Unkn.	$\ c_h - c\ _{0,h,\Omega}$	Unkn.	$\ c_h - c\ _{0,h,\Omega}$
10	16	100	0.1419	316	0.1347	900	0.1347
20	32	400	0.1364	1264	0.1222	3600	0.1222
40	64	1600	0.1249	5056	0.1032	14400	0.1031
80	128	6400	0.1064	20224	0.0777	57600	0.0777

Table 1.3: N, number of time steps, number of unknowns, and discrete $L^2(\Omega)$ errors for $\delta = 0.001$ at $t = 2$ for unrefined, locally refined, and uniformly refined square grids, respectively

We first perform a simple test. We consider $\delta = 0.1$, the time step of length $1/16$, a 20×20 grid of Ω , and its local refinement in a part of the subdomain where c is nonzero. The pointwise errors in centers of the cells at $t = 3/8$ are given in Figure 1.8. We can see that we have decreased the error in the refined part, whereas in the unrefined part, the error almost does not change. In particular, we do not produce any error around the interface between the refined and unrefined subdomains.

We next consider the whole time interval $(0, 2)$. During this interval, the peak of the exact solution moves from the point $(0.5, 1.35)$ to the point $(2.1, 2.15)$. We consider the unrefined, locally refined, and uniformly refined grids; the locally refined subdomain can be seen in

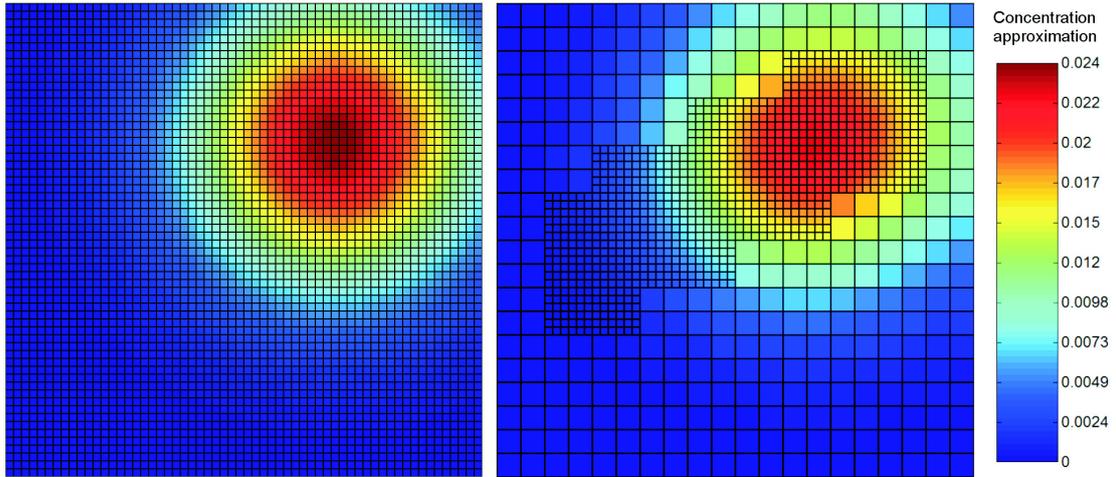


Figure 1.9: Exact solution (left) and approximate solution on a locally refined grid (right) for $\delta = 0.1$ at $t = 2$

Figure 1.9. We give the discrete $L^2(\Omega)$ errors at time $t = 2$ for $\delta = 0.1$ in Table 1.2 and for $\delta = 0.001$ in Table 1.3. The discrete norm $\|c_h - c\|_{0,h,\Omega}$ is the $L^2(\Omega)$ norm of the difference of the piecewise constant solution on the square grid and the exact solution, evaluated with a quadrature formula. We can see an expected linear convergence in the diffusion-dominated case. The error is significantly decreased by refining the grid locally around the region where c is nonzero. An example of the solution on a locally refined grid at $t = 2$ for $\delta = 0.1$ is given in Figure 1.9, together with the exact solution on the same grid refined uniformly. In the convection-dominated case, however, the error of the approximate solution is significant. The one-level local refinement is not sufficient in this case.

Real problem

We simulate in this section a real flow–transport problem provided by the HydroExpert company, Paris. We use for this purpose the software program TALISMAN of this company, cf. [104]. The combined scheme on a grid represented in Figure 1.7 is implemented in this code.

The domain $\Omega \in \mathbb{R}^3$ is an aquifer with bottom coordinate $z_b = z_b(x, y)$, top coordinate $z_t = z_t(x, y)$, and aperture $e = z_t - z_b = e(x, y)$, see Figure 1.10. We consider the Dupuit approximation of the Richards equation in Ω , consisting in integrating the Richards equation over the aquifer aperture e under the assumption that the flow is only horizontal, cf. [23]. We consider in addition the effect of water compressibility and a nonlinear discharge function. Let us denote by Ω' the horizontal plane of Ω , cf. Figure 1.10, and by Γ'_D and Γ'_N the Dirichlet and Neumann boundaries of Ω' , respectively. The flow problem then reads

$$\frac{\partial \tilde{\theta}(h)}{\partial t} - \nabla \cdot \tilde{\mathbf{K}}(h) \nabla h + Q_d(h) = \tilde{q}_{\text{out}} + \tilde{q}_{\text{in}} \quad \text{in } \Omega' \times (0, T), \quad (1.85a)$$

$$h(\cdot, 0) = h_0 \quad \text{in } \Omega', \quad (1.85b)$$

$$h = h_D \quad \text{on } \Gamma'_D \times (0, T), \quad (1.85c)$$

$$-\tilde{\mathbf{K}}(h) \nabla h \cdot \mathbf{n} = 0 \quad \text{on } \Gamma'_N \times (0, T), \quad (1.85d)$$

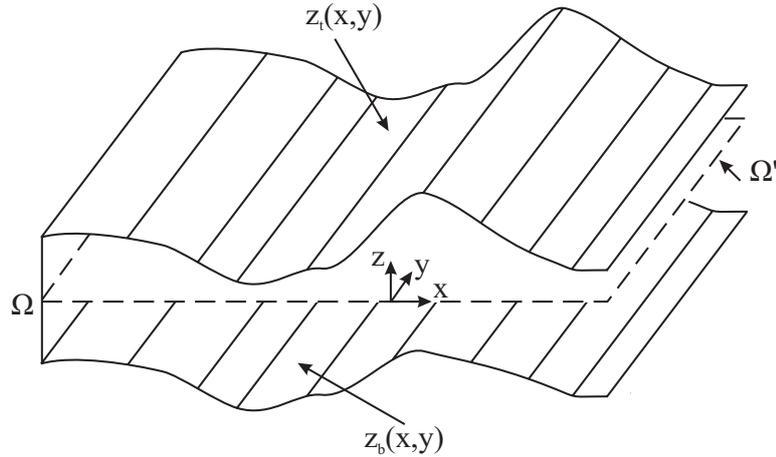


Figure 1.10: Considered domain $\Omega \in \mathbb{R}^3$ with its horizontal plane $\Omega' \in \mathbb{R}^2$

where

$$\frac{\partial \tilde{\theta}(h)}{\partial h} = \begin{cases} E_l & \text{if } h \leq z_t \\ E_s e & \text{if } h \geq z_t \end{cases}$$

and

$$\tilde{\mathbf{K}}(h) = \begin{cases} \mathbf{K}_s(h - z_b) & \text{if } h \leq z_t \\ \mathbf{K}_s e & \text{if } h \geq z_t \end{cases}$$

and finally

$$Q_d(h) = \begin{cases} 0 & \text{if } h \leq z_t \\ \mathbf{K}_d(h - z_t) & \text{if } h \geq z_t \end{cases} .$$

The problem (1.85a)–(1.85d) is two-dimensional, all the variables and in particular the unknown piezometric head h ([L]), $h = p + z$, are only functions of the horizontal coordinates x, y , and the gradient and divergence operators are also only two-dimensional. The storativity E_l ([–]) is related to the water content θ by $E_l = \theta_s - \theta_r$. The specific storativity E_s ([L⁻¹]) is given by the water compressibility and is usually very small in comparison with E_l/e . The tensor \mathbf{K}_s ([LT⁻¹]) expresses the hydraulic conductivity in the saturated state. The discharge Q_d ([LT⁻¹]) depends on the hydraulic conductance \mathbf{K}_d ([T⁻¹]). In analogy with the Darcy law, we define

$$\tilde{\mathbf{v}} := -\tilde{\mathbf{K}}(h)\nabla h. \quad (1.86)$$

Notice that $\tilde{\mathbf{v}}$ is a two-dimensional vector in Ω' with the units of [L²T⁻¹]. The flux of $\tilde{\mathbf{v}}$ through a segment $\mathbf{b} \in \Omega'$, $\int_{\mathbf{b}} \mathbf{v}(\mathbf{x}, t) \cdot \mathbf{n}_{\mathbf{b}} d\gamma(\mathbf{x})$ ([L³T⁻¹]), approximates the flux of groundwater over a vertical face in Ω , whose intersection with the horizontal plane Ω' is the segment \mathbf{b} .

We make similar approximations in the contaminant transport problem (1.79a)–(1.79c). We namely suppose that the concentration c does not vary with z and that the diffusion and convection are only two-dimensional in the horizontal plane Ω' of Ω . By formally integrating the three-dimensional convection–reaction–diffusion equation in Ω over e and adding the discharge, we replace the functions θ , q_{in} , and q_{out} defined in Ω by the functions $\tilde{\theta}$, \tilde{q}_{in} , and $\tilde{q}_{\text{out}} - Q_d(h)$ defined in Ω' . We finally use $\tilde{\mathbf{v}}$ instead of \mathbf{v} in the convection term and in the definition of the diffusion–dispersion tensor \mathbf{S} and consider the gradient and divergence operators only in Ω' . Notice that we have to replace the molecular diffusion coefficient σ by $\tilde{\sigma} = \sigma e$. Hence the final transport problem is, as the flow problem, two-dimensional in the plane Ω' , with the three-dimensional units (namely, the concentration is measured in [ML⁻³]).

With the assumptions of the previous paragraph, the transport scheme is constructed as follows. We consider a (nonmatching locally refined) square grid \mathcal{D}_h of Ω' and its dual triangular grid \mathcal{T}_h , as that in Figure 1.7. We associate with each $D \in \mathcal{D}_h$ an aperture e_D , given for instance as the mean of e over D . We seek the values c_D^n , $D \in \mathcal{D}_h$, $n \in \{1, 2, \dots, N\}$, such that

$$\begin{aligned} & \frac{\tilde{\theta}_D^n c_D^n - \tilde{\theta}_D^{n-1} c_D^{n-1}}{\Delta t_n} |D| + (\rho_b)_D \frac{w(c_D^n) - w(c_D^{n-1})}{\Delta t_n} |D| e_D - \sum_{V_E \in \mathcal{M}(V_D)} \tilde{S}_{D,E}^n (c_E^n - c_D^n) \\ & + \sum_{E \in \mathcal{N}(D)} \tilde{\mathbf{v}}_{D,E}^n \overline{c_{D,E}^n} + \lambda [\tilde{\theta}_D^n c_D^n + (\rho_b)_D w(c_D^n) e_D] |D| - [\tilde{q}_{\text{out}} - Q_d(h)]_D^n c_D^n |D| = (\tilde{q}_{\text{in}} c_s)_D^n |D| \\ & D \in \mathcal{D}_h, n \in \{1, 2, \dots, N\} \end{aligned}$$

with appropriately prescribed initial and boundary conditions. Here $\tilde{\theta}_D^n$, $D \in \mathcal{D}_h$, $n \in \{0, 1, \dots, N\}$, are the approximations of $\tilde{\theta}$ from (1.85a), given by a flow numerical scheme. In a similar manner, $\tilde{\mathbf{v}}_{D,E}^n$ are the approximations of the flux of $\tilde{\mathbf{v}}$ given by (1.86) through the interface between the control volumes $D \in \mathcal{D}_h$, $E \in \mathcal{N}(D)$ at time t_n . Note that from (1.85a) and using a locally conservative flow numerical scheme (such as a finite volume one), $\tilde{\mathbf{v}}_{D,E}^n = -\tilde{\mathbf{v}}_{E,D}^n$. We define $(\rho_b)_D$ e.g. as the mean of the bulk density ρ_b over the cube $D \times e_D$. The transmissibilities $\tilde{S}_{D,E}^n$ are defined by (1.83), while employing $\tilde{\mathbf{v}}$ and $\tilde{\sigma}$ in the definition of the diffusion–dispersion tensor \mathbf{S} instead of \mathbf{v} and σ . Note finally that again for non-negative transmissibilities, the scheme verifies the discrete maximum principle, which is in particular the consequence of $\nabla \cdot \tilde{\mathbf{v}} = \tilde{q}_{\text{out}} + \tilde{q}_{\text{in}} - Q_d(h) - \partial \tilde{\theta} / \partial t$ following from (1.85a) and (1.86).

The parameters of the given aquifer are visualized in Figure 1.11. Its horizontal plane Ω' fits into a rectangle 1500 × 2400 meters. Its aperture e is smaller than 9 meters and the above sea level of its top ranges between 137 and 146 meters. There is a small valley in the western part of the region, going in the north-southern direction. The given aquifer consists predominantly of sands with saturated hydraulic conductivity \mathbf{K}_s in orders of 10^{-3} m s^{-1} , but there is an important clay barrier with \mathbf{K}_s as low as 10^{-6} m s^{-1} along the eastern boundary and in the southeastern part of the aquifer. There is no discharge in the entire aquifer except of the valley, where the hydraulic conductance \mathbf{K}_d equals to 0.01 s^{-1} . We suppose that the storativity E_l and the specific storativity E_s are constant throughout the aquifer and equal respectively to 0.1 and 10^{-4} m^{-1} and that the saturated water content θ_s (porosity) equals to 0.3. The aquifer is receiving a constant effective recharge of 11 cm per year and there are 88 point sources of $10^{-9} \text{ m}^3 \text{ s}^{-1}$ distributed in the domain. Dirichlet boundary conditions are prescribed (fixed piezometric head imposed on the boundary of Ω'). The initial piezometric head is given also in Figure 1.11. The horizontal plane Ω' of the aquifer is divided into 8759 identical squares of 15 × 15 meters and the simulation period of one year is divided into 10 equidistant time steps.

The task is, given the hydrodynamical parameters specified in the previous paragraph, to simulate the propagation of a contaminant entering the aquifer at a concentration of 10 kg m^{-3} through a new source of a constant yield of $10^{-3} \text{ m}^3 \text{ s}^{-1}$. This source is located in the southeastern part of the domain near the clay barrier. We suppose that the bulk density ρ_b of the porous medium is constant throughout the aquifer and equals to 1600 kg m^{-3} . Also the longitudinal and transverse dispersivities α_L and α_T , as well as the molecular diffusion coefficient σ , are supposed constant and equal to, respectively, 10 m, 1 m, and $10^{-9} \text{ m}^2 \text{ s}^{-1}$. We consider the Langmuir adsorption isotherm with in particular $\nu_1 = 10^{-8}$ and $\nu_2 = 10^5 \text{ m}^3 \text{ kg}^{-1}$. Finally, first-order irreversible hydrolysis is supposed to take place so that the reaction rate

constant λ equals to 10^{-7} s^{-1} . The polluted region will be sufficiently far away from the boundary, so that Robin boundary conditions for the transport problem are prescribed (zero total concentration flux through the boundary of Ω').

First the flow problem was solved. We give in Figure 1.12 the flow field near the pollution source at the end of the simulated period. The absolute majority of the groundwater flows around the clay barrier because of its low conductivity. Then, using the water content and Darcy velocity values on each simulated period, the transport problem was dealt with. This problem is not convection-dominated; its complexity lies rather in the high ratio of the longitudinal and transverse dispersivities and in the complex flow field. Since the molecular diffusion coefficient is very small, the concentration should mainly follow the flow field and can in fact only enter the clay barrier due to the dispersion. The simulations were first performed on an unrefined grid and then, to justify the results, the mesh was refined. It again turned out that the refinement of the most exposed parts is sufficient. We give in particular in Figure 1.13 an example of a still very coarse grid that nevertheless already yields an accurate result. Although the combined scheme in the given case theoretically does not guarantee the discrete maximum principle, there were virtually no negative concentrations.

1.9.6 Concluding remarks

The code TALISMAN was originally a finite volume code working with regular cartesian grids. There was an interest in enabling a local refinement in this code, while subdividing the computational cells independently (as in Figure 1.7). The first attempt was to maintain the pure cell-centered finite volume scheme on square cells, while neglecting the orthogonality condition (i.e. to use the same scheme as that proposed in [36]). This violating of consistency while refining the grid however showed to produce an error rather than to decrease it. Another problem was how to discretize the inhomogeneous and anisotropic diffusion–dispersion tensor, especially on locally refined grids.

It was the implementation of the combined finite volume–finite element scheme in this code that overcame all the above difficulties. By constructing the dual triangular grid as in Figure 1.7, the combined scheme completely falls into the theoretical frame of Sections 1.9.3 and 1.9.4. The numerical experiments in Section 1.9.5 confirm that with this scheme, the local refinement does not produce any errors but substantially improves the results, and this without any considerable increase of the number of unknowns. In addition, inhomogeneous and anisotropic diffusion–dispersion tensors are easily incorporated. These results may serve as an example of the efficiency of the ideas proposed in this appendix.

A local refinement of the computational grid fixed throughout the calculation cannot of course lead to satisfactory results namely in the convection-dominated case, as it was illustrated in Section 1.9.5. We intend to use in the future an adaptive local mesh refinement. The idea is to refine the mesh automatically in the regions where the precision is not sufficient and to derefine it again as the precision gets sufficient. The derivation of a posteriori error estimates and development of a local refinement indicator such as that proposed in [95, 96] is a challenging task for a future work.

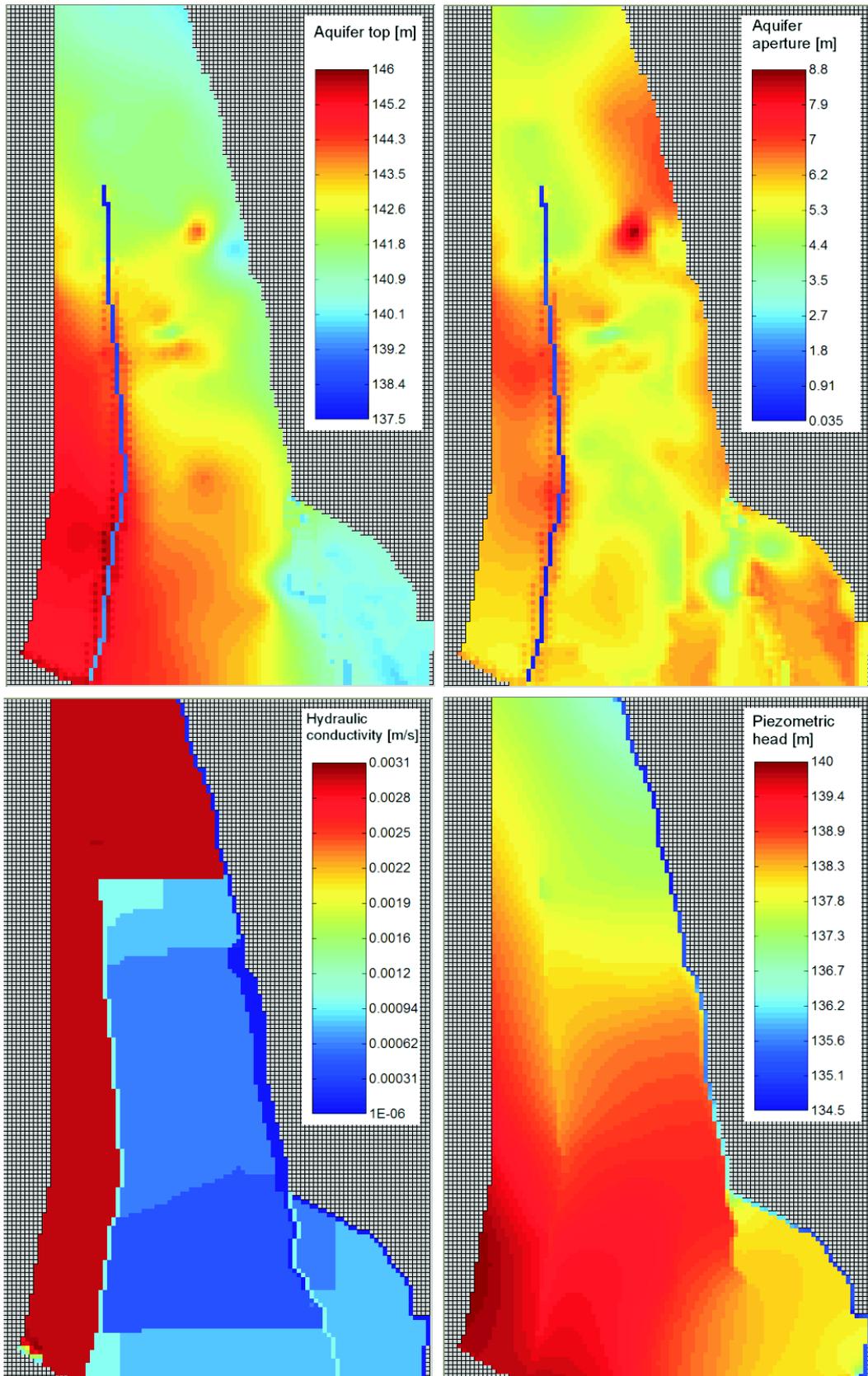


Figure 1.11: Simulated aquifer properties

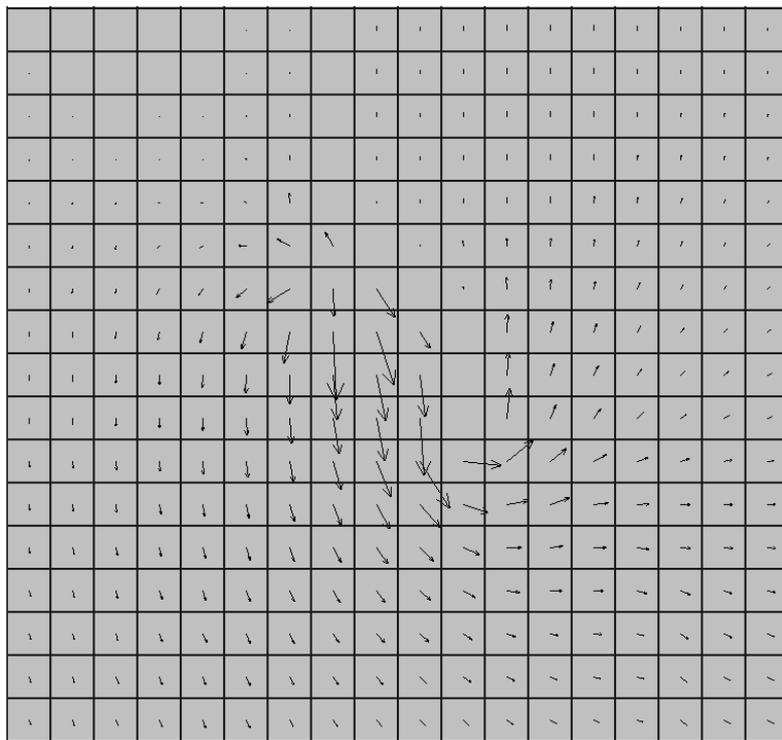


Figure 1.12: Flow field at the end of the simulated period

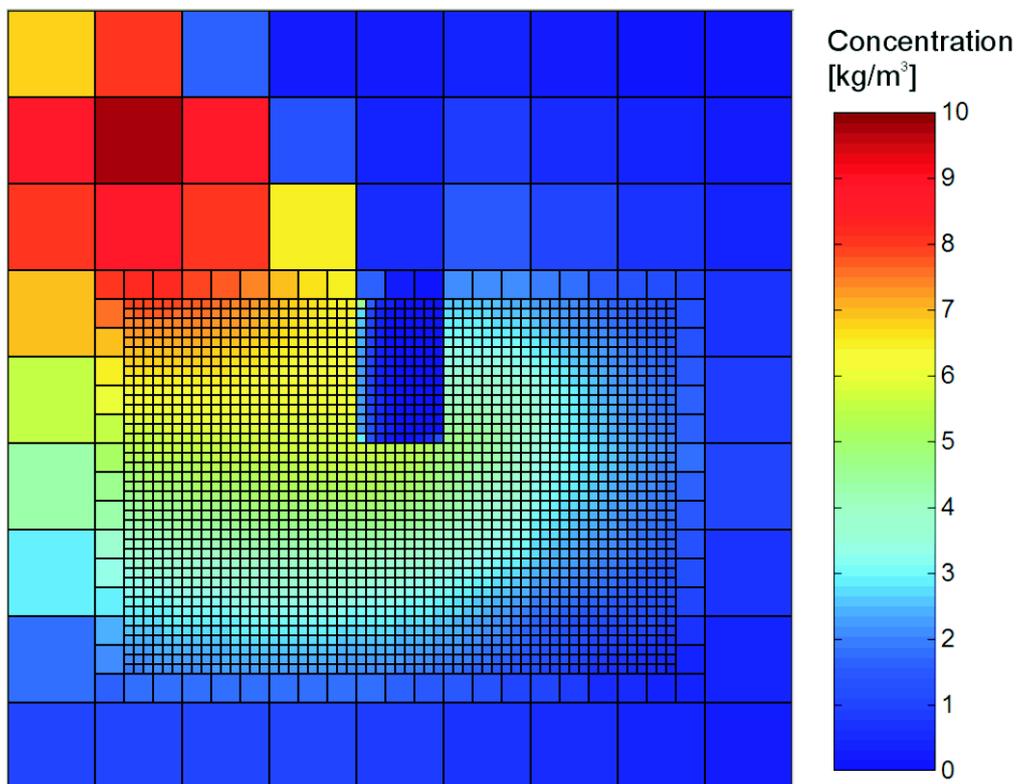


Figure 1.13: Concentration at the end of the simulated period, locally refined mesh

Chapitre 2

Discrete Poincaré–Friedrichs inequalities

We present in this chapter a direct proof of the discrete Poincaré–Friedrichs inequalities for a class of nonconforming approximations of the Sobolev space $H^1(\Omega)$, indicate optimal values of the constants in these inequalities, and extend the discrete Friedrichs inequality onto domains only bounded in one direction. We consider a polygonal domain Ω in two or three space dimensions and its shape-regular simplicial triangulation. The nonconforming approximations of $H^1(\Omega)$ consist of functions from H^1 on each element such that the mean values of their traces on interelement boundaries coincide. The key idea is to extend the proof of the discrete Poincaré–Friedrichs inequalities for piecewise constant functions used in the finite volume method. The results have applications in the analysis of nonconforming numerical methods, such as nonconforming finite element or discontinuous Galerkin methods.

2.1 Introduction

The Friedrichs (also called Poincaré) inequality

$$\int_{\Omega} g^2(\mathbf{x}) \, d\mathbf{x} \leq c_F \int_{\Omega} |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \quad \forall g \in H_0^1(\Omega) \quad (2.1)$$

and the Poincaré (also called mean Poincaré) inequality

$$\int_{\Omega} g^2(\mathbf{x}) \, d\mathbf{x} \leq c_P \int_{\Omega} |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} + \tilde{c}_P \left(\int_{\Omega} g(\mathbf{x}) \, d\mathbf{x} \right)^2 \quad \forall g \in H^1(\Omega) \quad (2.2)$$

(cf. [91]) play an important role in the theory of partial differential equations. We consider here a bounded polygonal domain (open and connected set) $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, $H^1(\Omega)$ is the Sobolev space of $L^2(\Omega)$ functions with square-integrable generalized derivatives, and $H_0^1(\Omega)$ is the subspace of $H^1(\Omega)$ of functions with zero trace on the boundary $\partial\Omega$ of Ω . We refer for instance to [4] for details on the spaces $H^1(\Omega)$, $H_0^1(\Omega)$.

Let $\{\mathcal{T}_h\}_h$ be a family of simplicial triangulations of Ω (consisting of triangles in space dimension two and of tetrahedra in space dimension three). Let the spaces $W(\mathcal{T}_h)$ be formed by functions locally in $H^1(K)$ on each $K \in \mathcal{T}_h$ such that the mean values of their traces on interior sides (edges if $d = 2$, faces if $d = 3$) coincide. Finally, let $W_0(\mathcal{T}_h) \subset W(\mathcal{T}_h)$ be such that the mean values of the traces on exterior sides of functions from $W_0(\mathcal{T}_h)$ are equal to zero (precise definitions are given in the next section). These spaces are nonconforming approximations of the continuous ones, i.e. $W_0(\mathcal{T}_h) \not\subset H_0^1(\Omega)$ and $W(\mathcal{T}_h) \not\subset H^1(\Omega)$. We investigate in this chapter analogies of (2.1) and (2.2) in the forms

$$\int_{\Omega} g^2(\mathbf{x}) \, d\mathbf{x} \leq C_F \sum_{K \in \mathcal{T}_h} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \quad \forall g \in W_0(\mathcal{T}_h), \forall h > 0, \quad (2.3)$$

$$\int_{\Omega} g^2(\mathbf{x}) \, d\mathbf{x} \leq C_P \sum_{K \in \mathcal{T}_h} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} + \tilde{C}_P \left(\int_{\Omega} g(\mathbf{x}) \, d\mathbf{x} \right)^2 \quad \forall g \in W(\mathcal{T}_h), \forall h > 0. \quad (2.4)$$

The validity of (2.3) for $W_0(\mathcal{T}_h)$ consisting of piecewise linear functions (used e.g. in the Crouzeix–Raviart finite element method) has been established in [116, Proposition 4.13] provided that Ω is convex and in [51] for a generally nonconvex Ω but with triangulations that are not locally refined. These results have been later extended in [84] onto \mathcal{T}_h only satisfying the shape regularity (minimal angle) assumption and onto spaces that include $W_0(\mathcal{T}_h)$. Another proof of this last result is presented in [28]. This paper also shows how to extend the discrete Friedrichs and Poincaré inequalities to general polygonal (nonmatching) partitions of Ω and to functions that do not satisfy the equality of the means of traces on interior sides, provided that (2.3), (2.4) are satisfied.

It was shown in [84] and in [28] that the constants C_F , C_P only depend on the domain Ω and on the shape regularity of the meshes. We establish in this chapter the exact dependence of C_F , C_P on these parameters. We show that in space dimension two C_F only depends on the area of Ω and that in space dimension two or three C_F only depends on the square of the infimum of the diameters of Ω in one direction. For convex domains, C_P only depends on the square of the diameter of Ω and on the ratio between the area of the circumscribed ball and the area of Ω . For nonconvex domains, our results involve a more complicated dependence of C_P on Ω . The above-mentioned dependencies are optimal in the sense that they coincide with the dependencies of c_F , c_P on Ω in the continuous case. The dependence of C_F on Ω also allows for

the extension of the discrete Friedrichs inequality to domains only bounded in one direction. We finally show that C_F depends, in space dimension two and provided that it is expressed using the area of Ω , on the square of a parameter describing the shape regularity of the meshes given in the next section. This dependence still holds true for C_F in space dimension two or three and expressed using the square of the infimum of the diameters of Ω in one direction and also for C_P , provided that the mesh is not locally refined. We present an example showing that this dependence is optimal. For locally refined meshes, our results involve a more complicated dependence on the shape regularity parameter. The established constants are necessary in the analysis of nonconforming numerical methods, such as nonconforming finite element or discontinuous Galerkin methods.

Our proof of the discrete Friedrichs and Poincaré inequalities on the spaces $W_0(\mathcal{T}_h)$, $W(\mathcal{T}_h)$ respectively is more direct than those presented in [84] and in [28]; in particular, all the necessary intermediate results are proved here. In [84] the author uses a Clément-type interpolation operator (cf. [43]) mapping the space $W_0(\mathcal{T}_h)$ to $H_0^1(\Omega)$. In [28] the key idea is to construct nonconforming P_1 interpolants of functions from $W(\mathcal{T}_h)$ and to connect the nonconforming P_1 finite elements and conforming P_2 finite elements (in space dimension two) or conforming P_3 finite elements (in space dimension three). In both cases one finally makes use of the continuous inequalities (2.1), (2.2). Our main idea is to construct a piecewise constant interpolant and to extend the discrete Poincaré–Friedrichs inequalities for piecewise constant functions known from finite volume methods, see [59, 61]. In particular, we do not make use of the continuous inequalities; since $H_0^1(\Omega) \subset W_0(\mathcal{T}_h)$ and $H^1(\Omega) \subset W(\mathcal{T}_h)$, we rather prove them.

The rest of the chapter is organized as follows. In Section 2.2 we describe the assumptions on \mathcal{T}_h , define a dual mesh \mathcal{D}_h where the dual elements are associated with the sides of \mathcal{T}_h , define the function spaces used in the sequel, and introduce the interpolation operator. In Section 2.3 we give the discrete Friedrichs inequality for piecewise constant functions on \mathcal{D}_h . In Section 2.4 we prove some interpolation estimates on functions from $H^1(K)$, where K is a simplex in two or three space dimensions. In Section 2.5 we prove the discrete Friedrichs inequality for functions from $W_0(\mathcal{T}_h)$, using their interpolation by piecewise constant functions on \mathcal{D}_h . In Section 2.6 we show how this proof simplifies for Crouzeix–Raviart finite elements in two space dimensions. Finally, Section 2.7 is devoted to the proof of the discrete Poincaré inequality for piecewise constant functions on \mathcal{D}_h and Section 2.8 to the extension of this result to functions from $W(\mathcal{T}_h)$.

2.2 Notation and assumptions

Throughout this chapter, we shall mean by “segment” a segment of a straight line. Let us consider a domain $K \subset \mathbb{R}^d$, $d = 2, 3$. We denote by $\|\cdot\|_{0,K}$ the norm on $L^2(K)$, $\|g\|_{0,K}^2 = \int_K g^2(\mathbf{x}) \, d\mathbf{x}$, by $|K|$ is the d -dimensional Lebesgue measure of K , by $|\sigma|$ the $(d-1)$ -dimensional Lebesgue measure of σ , a part of a hyperplane in \mathbb{R}^d , and by $|\mathbf{s}|$ the length of a segment \mathbf{s} . Let \mathbf{b} be a vector. We shall mean by the diameter of K in the direction of \mathbf{b} , denoted by $\text{diam}_{\mathbf{b}}(K)$, the supremum of the lengths of segments \mathbf{s} with the direction vector \mathbf{b} such that $\mathbf{s} \subset K$. The diameter of K is the supremum of the lengths of all the segments \mathbf{s} such that $\mathbf{s} \subset K$.

Triangulation

We suppose that \mathcal{T}_h for all $h > 0$ consists of closed simplices such that $\bar{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$ and such that if $K, L \in \mathcal{T}_h$, $K \neq L$, then $K \cap L$ is either an empty set or a common face, edge, or

vertex of K and L . The parameter h is defined by $h := \max_{K \in \mathcal{T}_h} \text{diam}(K)$. We denote by \mathcal{E}_h the set of all sides, by $\mathcal{E}_h^{\text{int}}$ the set of all interior sides, by $\mathcal{E}_h^{\text{ext}}$ the set of all exterior sides, and by \mathcal{E}_K the set of all the sides of an element $K \in \mathcal{T}_h$. We make the following shape regularity assumption on $\{\mathcal{T}_h\}_h$:

Assumption (A) (Shape regularity assumption)

There exists a constant $\kappa_{\mathcal{T}} > 0$ such that

$$\min_{K \in \mathcal{T}_h} \frac{|K|}{\text{diam}(K)^d} \geq \kappa_{\mathcal{T}} \quad \forall h > 0.$$

Assumption (A) is equivalent to the existence of a constant $\theta_{\mathcal{T}} > 0$ such that

$$\max_{K \in \mathcal{T}_h} \frac{\text{diam}(K)}{\rho_K} \leq \theta_{\mathcal{T}} \quad \forall h > 0, \quad (2.5)$$

where ρ_K is the diameter of the largest ball inscribed in the simplex K . Finally, Assumption (A) is equivalent to the existence of a constant $\phi_{\mathcal{T}} > 0$ such that

$$\min_{K \in \mathcal{T}_h} \phi_K \geq \phi_{\mathcal{T}} \quad \forall h > 0. \quad (2.6)$$

Here ϕ_K is the smallest angle of the simplex K (plain angle in radians if $d = 2$ and spheric angle in steradians if $d = 3$).

In the sequel we shall consider apart triangulations that may not be locally refined, i.e. the case where the following assumption holds:

Assumption (B) (Inverse assumption)

There exists a constant $\zeta_{\mathcal{T}} > 0$ such that

$$\max_{K \in \mathcal{T}_h} \frac{h}{\text{diam}(K)} \leq \zeta_{\mathcal{T}} \quad \forall h > 0.$$

Assumptions (A) and (B) imply

$$\min_{K \in \mathcal{T}_h} \frac{|K|}{h^d} \geq \tilde{\kappa}_{\mathcal{T}} \quad \forall h > 0, \quad (2.7)$$

where $\tilde{\kappa}_{\mathcal{T}} := \kappa_{\mathcal{T}}/\zeta_{\mathcal{T}}^d$.

Dual mesh

In the sequel we will use a dual mesh \mathcal{D}_h to \mathcal{T}_h such that $\overline{\Omega} = \bigcup_{D \in \mathcal{D}_h} D$. There is one dual element D associated with each side $\sigma_D \in \mathcal{E}_h$. We construct it by connecting the barycentres of every $K \in \mathcal{T}_h$ that contains σ_D through the vertices of σ_D . For $\sigma_D \in \mathcal{E}_h^{\text{ext}}$, the contour of D is completed by the side σ_D itself. We refer to Fig. 2.1 for the two-dimensional case. We denote by $\mathcal{D}_h^{\text{int}}$ the set of all interior and by $\mathcal{D}_h^{\text{ext}}$ the set of all boundary dual elements. As for the primal mesh, we set \mathcal{F}_h , $\mathcal{F}_h^{\text{int}}$, $\mathcal{F}_h^{\text{ext}}$, and \mathcal{F}_D for the dual mesh sides. We denote by Q_D the barycentre of a side σ_D and for two adjacent elements $D, E \in \mathcal{D}_h$, we set $\sigma_{D,E} := \partial D \cap \partial E$, $d_{D,E} := |Q_E - Q_D|$, and $K_{D,E}$ the element of \mathcal{T}_h such that $\sigma_{D,E} \subset K_{D,E}$. We remark that

$$|K \cap D| = \frac{|K|}{d+1} \quad (2.8)$$

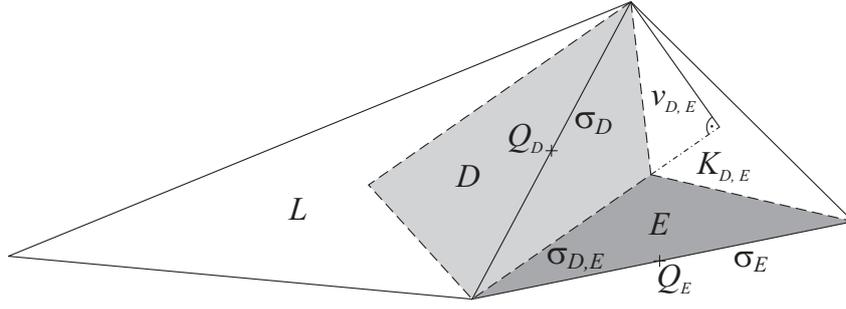


Figure 2.1: Triangles $K, L \in \mathcal{T}_h$ and dual elements $D, E \in \mathcal{D}_h$ with edges $\sigma_D, \sigma_E \in \mathcal{E}_h$

for each $K \in \mathcal{T}_h$ and $D \in \mathcal{D}_h$ such that $\sigma_D \in \mathcal{E}_K$. Let us now consider $\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}$, $\sigma_{D,E} = \partial D \cap \partial E$ in the two-dimensional case. Let $K_{D,E} \cap D$ be in the clockwise direction from $K_{D,E} \cap E$. We then define $v_{D,E}$ as the height of the triangle $|K_{D,E} \cap D|$ with respect to its base $\sigma_{D,E}$ and have (see Fig. 2.1)

$$|K_{D,E} \cap D| = \frac{|\sigma_{D,E}| v_{D,E}}{2}. \quad (2.9)$$

Function spaces

We define the space $W(\mathcal{T}_h)$ by

$$\begin{aligned} W(\mathcal{T}_h) := & \left\{ g \in L^2(\Omega); g|_K \in H^1(K) \quad \forall K \in \mathcal{T}_h, \right. \\ & \int_{\sigma_{K,L}} g|_K(\mathbf{x}) \, d\gamma(\mathbf{x}) = \int_{\sigma_{K,L}} g|_L(\mathbf{x}) \, d\gamma(\mathbf{x}) \\ & \left. \forall \sigma_{K,L} \in \mathcal{E}_h^{\text{int}}, \sigma_{K,L} = \partial K \cap \partial L \right\}. \end{aligned} \quad (2.10)$$

We keep the same notation for the function g and its trace and denote $d\gamma(\mathbf{x})$ the integration symbol for the Lebesgue measure on a hyperplane of Ω . The space $W_0(\mathcal{T}_h)$ is defined by

$$W_0(\mathcal{T}_h) := \left\{ g \in W(\mathcal{T}_h); \int_{\sigma} g(\mathbf{x}) \, d\gamma(\mathbf{x}) = 0 \quad \forall \sigma \in \mathcal{E}_h^{\text{ext}} \right\}. \quad (2.11)$$

We finally define

$$|g|_{1,\mathcal{T}} := \left(\sum_{K \in \mathcal{T}_h} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \right)^{\frac{1}{2}},$$

which is a seminorm on $W(\mathcal{T}_h)$ and a norm on $W_0(\mathcal{T}_h)$. The spaces $X(\mathcal{T}_h) \subset W(\mathcal{T}_h)$ and $X_0(\mathcal{T}_h) \subset W_0(\mathcal{T}_h)$ are defined by piecewise linear functions on \mathcal{T}_h . Note that the functions from $X(\mathcal{T}_h)$ are continuous in barycentres of interior sides and that the functions from $X_0(\mathcal{T}_h)$ are moreover equal to zero in barycentres of exterior sides.

The space $Y(\mathcal{D}_h)$ is the space of piecewise constant functions on \mathcal{D}_h ,

$$Y(\mathcal{D}_h) := \{c \in L^2(\Omega); c|_D \text{ is constant } \forall D \in \mathcal{D}_h\},$$

and $Y_0(\mathcal{D}_h)$ is its subspace of functions equal to zero on all $D \in \mathcal{D}_h^{\text{ext}}$,

$$Y_0(\mathcal{D}_h) := \{c \in Y(\mathcal{D}_h); c|_D = 0 \quad \forall D \in \mathcal{D}_h^{\text{ext}}\}.$$

For $c \in Y(\mathcal{D}_h)$ given by the values c_D on $D \in \mathcal{D}_h$, we define

$$\begin{aligned} |c|_{1,\mathcal{T},*} &:= \left(\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|}{v_{D,E}} (c_E - c_D)^2 \right)^{\frac{1}{2}}, \\ |c|_{1,\mathcal{T},\dagger} &:= \left(\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|}{\text{diam}(K_{D,E})} (c_E - c_D)^2 \right)^{\frac{1}{2}}, \\ |c|_{1,\mathcal{T},\ddagger} &:= \left(\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|}{d_{D,E}} (c_E - c_D)^2 \right)^{\frac{1}{2}}; \end{aligned}$$

$|\cdot|_{1,\mathcal{T},*}$, $|\cdot|_{1,\mathcal{T},\dagger}$, and $|\cdot|_{1,\mathcal{T},\ddagger}$ are seminorms on $Y(\mathcal{D}_h)$ and norms on $Y_0(\mathcal{D}_h)$.

Interpolation operator

The interpolation operator I associates to a function $g \in W(\mathcal{T}_h)$ a function $I(g) \in Y(\mathcal{D}_h)$ such that

$$I(g)|_D = g_D := \frac{1}{|\sigma_D|} \int_{\sigma_D} g|_K(\mathbf{x}) \, d\gamma(\mathbf{x}) \quad \forall D \in \mathcal{D}_h,$$

where $K \in \mathcal{T}_h$ is such that $\sigma_D \in \mathcal{E}_K$. Note that by (2.10), if $\sigma_D \in \mathcal{E}_K$ and $\sigma_D \in \mathcal{E}_L$, $K \neq L$, the choice between K and L does not matter. We recall that $\sigma_D \in \mathcal{E}_h$ is the side associated with the dual element $D \in \mathcal{D}_h$. Note that for $g \in W_0(\mathcal{T}_h)$, $I(g) \in Y_0(\mathcal{D}_h)$.

2.3 Discrete Friedrichs inequality for piecewise constant functions

In finite volume methods (cf. [61]) one can prove the discrete Friedrichs inequality for piecewise constant functions for meshes that satisfy the following orthogonality property: there exists a point associated with each element of the mesh such that the straight line connecting these points for two neighboring elements is orthogonal to the common side of these two elements. The proofs in [59, 61] rely on this property of the meshes. We present in this section analogies of Lemma 9.5 and consequent Remark 9.13 and of Lemma 9.1 of [61] for the mesh \mathcal{D}_h , where the orthogonality property is not necessarily satisfied.

Theorem 2.3.1. (Discrete Friedrichs inequality for piecewise constant functions in two space dimensions) *Let $d = 2$. Then for all $c \in Y_0(\mathcal{D}_h)$,*

$$\|c\|_{0,\Omega}^2 \leq \frac{|\Omega|}{2} |c|_{1,\mathcal{T},*}^2.$$

PROOF:

Let $\mathbf{b}_1 = (1, 0)$ and $\mathbf{b}_2 = (0, 1)$ be two fixed unit vectors in the axis directions. For all $\mathbf{x} \in \Omega$, let \mathcal{B}_x^1 and \mathcal{B}_x^2 be the straight lines going through \mathbf{x} and defined by the vectors \mathbf{b}_1 , \mathbf{b}_2 respectively. Let the functions $\chi_\sigma^{(i)}(\mathbf{x})$, $i = 1, 2$, for each $\sigma \in \mathcal{F}_h^{\text{int}}$ be defined by

$$\chi_\sigma^{(i)}(\mathbf{x}) := \begin{cases} 1 & \text{if } \sigma \cap \mathcal{B}_x^i \neq \emptyset \\ 0 & \text{if } \sigma \cap \mathcal{B}_x^i = \emptyset \end{cases}.$$

Let finally $D \in \mathcal{D}_h^{\text{int}}$ be fixed. Then for a.e. $\mathbf{x} \in D$, \mathcal{B}_x^i , $i = 1, 2$, do not contain any vertex of the dual mesh and $\mathcal{B}_x^i \cap \sigma$, $i = 1, 2$, contain at most one point of all $\sigma \in \mathcal{F}_h$. This implies

that for a.e. $\mathbf{x} \in D$, $\mathcal{B}_{\mathbf{x}}^i$, $i = 1, 2$, always have to intersect the interior of some $E \in \mathcal{D}_h^{\text{ext}}$ before “leaving” or after “entering” Ω (we recall that Ω may be nonconvex). Using this, the fact that $c_E = 0$ for all $E \in \mathcal{D}_h^{\text{ext}}$, and the triangle inequality, we have

$$2|c_D| \leq \sum_{\sigma_{F,G} \in \mathcal{F}_h^{\text{int}}} |c_G - c_F| \chi_{\sigma_{F,G}}^{(i)}(\mathbf{x}) \quad \text{for a.e. } \mathbf{x} \in D, \quad i = 1, 2.$$

This gives

$$|c_D|^2 \leq \frac{1}{4} \sum_{\sigma_{F,G} \in \mathcal{F}_h^{\text{int}}} |c_G - c_F| \chi_{\sigma_{F,G}}^{(1)}(\mathbf{x}) \sum_{\sigma_{F,G} \in \mathcal{F}_h^{\text{int}}} |c_G - c_F| \chi_{\sigma_{F,G}}^{(2)}(\mathbf{x}) \quad \text{for a.e. } \mathbf{x} \in D,$$

which is obviously valid also for $D \in \mathcal{D}_h^{\text{ext}}$, considering that $c_D = 0$ on $D \in \mathcal{D}_h^{\text{ext}}$. Integrating the above inequality over D and summing over $D \in \mathcal{D}_h$ yields

$$\sum_{D \in \mathcal{D}_h} c_D^2 |D| \leq \frac{1}{4} \int_{\Omega} \left(\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E - c_D| \chi_{\sigma_{D,E}}^{(1)}(\mathbf{x}) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E - c_D| \chi_{\sigma_{D,E}}^{(2)}(\mathbf{x}) \right) d\mathbf{x}.$$

Let $\alpha = \inf\{x_1; (x_1, x_2) \in \Omega\}$ and $\beta = \sup\{x_1; (x_1, x_2) \in \Omega\}$. For each $x_1 \in (\alpha, \beta)$, we denote by $J(x_1)$ the set of x_2 such that $\mathbf{x} = (x_1, x_2) \in \Omega$. We now notice that $\chi_{\sigma}^{(1)}(\mathbf{x})$ only depends on x_2 and that $\chi_{\sigma}^{(2)}(\mathbf{x})$ only depends on x_1 . Thus

$$\begin{aligned} & \int_{\alpha}^{\beta} \int_{J(x_1)} \left(\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E - c_D| \chi_{\sigma_{D,E}}^{(1)}(x_2) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E - c_D| \chi_{\sigma_{D,E}}^{(2)}(x_1) \right) dx_2 dx_1 \\ &= \int_{\alpha}^{\beta} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E - c_D| \chi_{\sigma_{D,E}}^{(2)}(x_1) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E - c_D| \int_{J(x_1)} \chi_{\sigma_{D,E}}^{(1)}(x_2) dx_2 dx_1 \\ &\leq \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E - c_D| |\sigma_{D,E}| \int_{\alpha}^{\beta} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |c_E - c_D| \chi_{\sigma_{D,E}}^{(2)}(x_1) dx_1, \end{aligned}$$

where we have used $\int_{J(x_1)} \chi_{\sigma_{D,E}}^{(1)}(x_2) dx_2 \leq |\sigma_{D,E}|$. Using analogously $\int_{\alpha}^{\beta} \chi_{\sigma_{D,E}}^{(2)}(x_1) dx_1 \leq |\sigma_{D,E}|$, we come to

$$\sum_{D \in \mathcal{D}_h} c_D^2 |D| \leq \frac{1}{4} \left(\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |\sigma_{D,E}| |c_E - c_D| \right)^2.$$

Finally, using the Cauchy–Schwarz inequality, we have

$$\sum_{D \in \mathcal{D}_h} c_D^2 |D| \leq \frac{1}{4} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |\sigma_{D,E}| v_{D,E} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|}{v_{D,E}} (c_E - c_D)^2.$$

The equality $\sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} |\sigma_{D,E}| v_{D,E} = 2|\Omega|$, which follows from (2.9), concludes the proof. \square

Remark 2.3.2. (Discrete Friedrichs inequality for piecewise constant functions on equilateral simplices) Let $\mathbf{b} \in \mathbb{R}^d$ be a fixed vector and let \mathcal{T}_h consist of equilateral simplices. Then for all $c \in Y_0(\mathcal{D}_h)$,

$$\|c\|_{0,\Omega}^2 \leq [\text{diam}_{\mathbf{b}}(\Omega) + 2h]^2 |c|_{1,\mathcal{T},\ddagger}^2.$$

This follows from [61, Lemma 9.1] (cf. alternatively [59, Lemma 1]), since the dual mesh \mathcal{D}_h satisfies in this case the orthogonality property.

Lemma 2.3.3. *Let Assumption (B) be satisfied and let $\mathbf{b} \subset \Omega$ be a segment that does not contain any vertex of the dual mesh \mathcal{D}_h . Then*

$$A := \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \sigma_{D,E} \cap \mathbf{b} \neq \emptyset} \text{diam}(K_{D,E}) \leq C_{d,\mathcal{T}} \text{diam}_{\mathbf{b}}(\Omega),$$

where

$$C_{d,\mathcal{T}} = \frac{2^d(d-1)}{\tilde{\kappa}_{\mathcal{T}}} (1 + 2\theta_{\mathcal{T}}). \quad (2.12)$$

PROOF:

The number of nonzero terms of A is equal to the number of interior dual sides intersected by \mathbf{b} . In view of the fact that \mathbf{b} does not contain any vertex of the dual mesh, this number is bounded by $2(d-1)$ -times the number of simplices $K \in \mathcal{T}_h$ whose interior is intersected by \mathbf{b} . All intersected simplices have to be entirely in the rectangle/rectangular parallelepiped constructed around \mathbf{b} , with the distance between \mathbf{b} and its boundary equal to h . Considering the consequence (2.7) of Assumptions (A) and (B), we can estimate the number of intersected elements by

$$\frac{(2h)^{d-1}(|\mathbf{b}| + 2h)}{\tilde{\kappa}_{\mathcal{T}} h^d}.$$

Using in addition $\text{diam}(K_{D,E}) \leq h$ and $|\mathbf{b}| \leq \text{diam}_{\mathbf{b}}(\Omega)$, we have

$$A \leq \frac{2^d(d-1)}{\tilde{\kappa}_{\mathcal{T}}} (\text{diam}_{\mathbf{b}}(\Omega) + 2h).$$

Noticing that

$$h \leq \theta_{\mathcal{T}} \text{diam}_{\mathbf{b}}(\Omega) \quad (2.13)$$

by the consequence (2.5) of Assumption (A) concludes the proof. \square

Lemma 2.3.4. *Let $\mathbf{b} \subset \Omega$ be a segment that does not contain any vertex of the dual mesh \mathcal{D}_h . Then*

$$A := \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \sigma_{D,E} \cap \mathbf{b} \neq \emptyset} \text{diam}(K_{D,E}) \leq C_{d,\mathcal{T}} \text{diam}_{\mathbf{b}}(\Omega), \quad (2.14)$$

where

$$C_{d,\mathcal{T}} = 4N(d-1)\theta_{\mathcal{T}}^{2N}, \quad N = \frac{2^{d-1}\pi}{\phi_{\mathcal{T}}}. \quad (2.15)$$

PROOF:

The number of nonzero terms of A is equal to the number of interior dual sides intersected by \mathbf{b} . In view of the fact that \mathbf{b} does not contain any vertex of the dual mesh, this number is bounded by $2(d-1)$ -times the number of simplices $K \in \mathcal{T}_h$ whose interior is intersected by \mathbf{b} . We consider two different cases.

If the segment \mathbf{b} intersects at most N simplices, where N is given by (2.15), we estimate

$$\text{diam}(K) \leq \theta_{\mathcal{T}} \rho_K \leq \theta_{\mathcal{T}} \text{diam}_{\mathbf{b}}(\Omega) \quad \forall K \in \mathcal{T}_h,$$

which follows from the consequence (2.5) of Assumption (A), to see that

$$A \leq 2(d-1)N\theta_{\mathcal{T}} \text{diam}_{\mathbf{b}}(\Omega).$$

This in view of $N \geq 1$ and $\theta_{\mathcal{T}} > 1$ implies (2.14) with $C_{d,\mathcal{T}}$ given by (2.15).

We next consider the case where the segment \mathbf{b} intersects at least $N + 1$ simplices. We divide it into a system of nonoverlapping segments $\{\mathbf{b}_k\}_{k=1}^M$ such that $\mathbf{b} = \cup_{k=1}^M \mathbf{b}_k$. We further require that each \mathbf{b}_k intersects at least N and at most $2N$ simplices and that no simplex has an intersection with a positive 1-dimensional Lebesgue measure with two different segments. We then have

$$A \leq 2(d-1) \sum_{k=1}^M \sum_{K \in \mathcal{T}_h; K^\circ \cap \mathbf{b}_k \neq \emptyset} \text{diam}(K).$$

Next it follows from the consequence (2.5) of Assumption (A) that $\rho_K \leq \theta_T \rho_L$ if $K, L \in \mathcal{T}_h$ are neighboring elements. Recall that ρ_K is the diameter of the largest ball inscribed in the simplex K . Thus we come to

$$\frac{\max_{K \in \mathcal{T}_h; K^\circ \cap \mathbf{b}_k \neq \emptyset} \rho_K}{\min_{K \in \mathcal{T}_h; K^\circ \cap \mathbf{b}_k \neq \emptyset} \rho_L} \leq \theta_T^{2N-1} \quad \forall k = 1, \dots, M.$$

We further claim that

$$\min_{K \in \mathcal{T}_h; K^\circ \cap \mathbf{b}_k \neq \emptyset} \rho_K \leq |\mathbf{b}_k| \quad \forall k = 1, \dots, M,$$

i.e. if we take N simplices intersected by a straight line, where N is given by (2.15), then the length of the intersection is at least equal to the smallest diameter of the inscribed balls of the simplices. We show this by contradiction. Let us suppose that N simplices are intersected by a segment \mathbf{l} and that the length of the intersection is smaller or equal to the smallest diameter of the inscribed balls of the simplices. By contradiction, the centers of all the inscribed balls lie outside of the segment \mathbf{l} . Now with each simplex intersected by \mathbf{l} , we add an angle greater or equal to $\phi_{\mathcal{T}}$ by the consequence (2.6) of Assumption (A). Since we have N simplices, their angles fill the whole circle (2π , $d = 2$) or sphere (4π , $d = 3$), which already yields the contradiction.

Using the last two estimates, the fact that each \mathbf{b}_k intersects at most $2N$ simplices, and once more the consequence (2.5) of Assumption (A), we have

$$A \leq 2(d-1) \sum_{k=1}^M 2N \theta_T^{2N} |\mathbf{b}_k| \leq 4N(d-1) \theta_T^{2N} |\mathbf{b}| \leq 4N(d-1) \theta_T^{2N} \text{diam}_{\mathbf{b}}(\Omega).$$

This proves (2.14) with $C_{d,\mathcal{T}}$ given by (2.15) for the second case and consequently the whole lemma. \square

Theorem 2.3.5. (Discrete Friedrichs inequality for piecewise constant functions)

Let $\mathbf{b} \in \mathbb{R}^d$ be a fixed vector. Then for all $c \in Y_0(\mathcal{D}_h)$,

$$\|c\|_{0,\Omega}^2 \leq C_{d,\mathcal{T}} [\text{diam}_{\mathbf{b}}(\Omega)]^2 |c|_{1,\mathcal{T},\dagger}^2,$$

where $C_{d,\mathcal{T}}$ is given by (2.12) when Assumption (B) is satisfied and by (2.15) in the general case.

PROOF:

For all $\mathbf{x} \in \Omega$, we denote by $\mathcal{B}_{\mathbf{x}}$ the straight semi-line defined by the origin \mathbf{x} and the vector \mathbf{b} . Let $\mathbf{y}(\mathbf{x}) \in \partial\Omega \cap \mathcal{B}_{\mathbf{x}}$ be the point where $\mathcal{B}_{\mathbf{x}}$ intersects $\partial\Omega$ for the first time. Then $[\mathbf{x}, \mathbf{y}(\mathbf{x})] \subset \overline{\Omega}$. We finally define a function $\chi_\sigma(\mathbf{x})$ for each $\sigma \in \mathcal{F}_h^{\text{int}}$ by

$$\chi_\sigma(\mathbf{x}) := \begin{cases} 1 & \text{if } \sigma \cap [\mathbf{x}, \mathbf{y}(\mathbf{x})] \neq \emptyset \\ 0 & \text{if } \sigma \cap [\mathbf{x}, \mathbf{y}(\mathbf{x})] = \emptyset \end{cases}. \quad (2.16)$$

Let $D \in \mathcal{D}_h^{\text{int}}$ be fixed. Then for a.e. $\mathbf{x} \in D$, $\mathcal{B}_{\mathbf{x}}$ does not contain any vertex of the dual mesh and $\mathcal{B}_{\mathbf{x}} \cap \sigma$ contains at most one point of all $\sigma \in \mathcal{F}_h$. This implies that for a.e. $\mathbf{x} \in D$, $\mathcal{B}_{\mathbf{x}}$ always has to intersect the interior of some $E \in \mathcal{D}_h^{\text{ext}}$ before “leaving” Ω . Using this, the fact that $c_E = 0$ for all $E \in \mathcal{D}_h^{\text{ext}}$, and the triangle inequality, we have

$$|c_D| \leq \sum_{\sigma_{F,G} \in \mathcal{F}_h^{\text{int}}} |c_G - c_F| \chi_{\sigma_{F,G}}(\mathbf{x}) \quad \text{for a.e. } \mathbf{x} \in D.$$

The Cauchy–Schwarz inequality yields

$$|c_D|^2 \leq \sum_{\sigma_{F,G} \in \mathcal{F}_h^{\text{int}}} \chi_{\sigma_{F,G}}(\mathbf{x}) \text{diam}(K_{F,G}) \sum_{\sigma_{F,G} \in \mathcal{F}_h^{\text{int}}} \frac{(c_G - c_F)^2}{\text{diam}(K_{F,G})} \chi_{\sigma_{F,G}}(\mathbf{x}) \quad \text{for a.e. } \mathbf{x} \in D, \quad (2.17)$$

which is obviously valid also for $D \in \mathcal{D}_h^{\text{ext}}$, considering that $c_D = 0$ on $D \in \mathcal{D}_h^{\text{ext}}$. Integrating the above inequality over D , summing over $D \in \mathcal{D}_h$, and using Lemma 2.3.3 when Assumption (B) is satisfied and Lemma 2.3.4 in the general case yields

$$\sum_{D \in \mathcal{D}_h} |c_D|^2 |D| \leq C_{d,T} \text{diam}_{\mathbf{b}}(\Omega) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{(c_E - c_D)^2}{\text{diam}(K_{D,E})} \int_{\Omega} \chi_{\sigma_{D,E}}(\mathbf{x}) \, d\mathbf{x}.$$

Now the value $\int_{\Omega} \chi_{\sigma_{D,E}}(\mathbf{x}) \, d\mathbf{x}$ is the measure of the set of points of Ω located inside a cylinder whose basis is $\sigma_{D,E}$ and generator vector is $-\mathbf{b}$. Thus

$$\int_{\Omega} \chi_{\sigma_{D,E}}(\mathbf{x}) \, d\mathbf{x} \leq |\sigma_{D,E}| \text{diam}_{\mathbf{b}}(\Omega),$$

which leads to the assertion of the lemma. \square

2.4 Interpolation estimates on functions from $H^1(K)$

Lemma 2.4.1. *Let K be a simplex, σ its side, and $g \in H^1(K)$. We set*

$$g_K := \frac{1}{|K|} \int_K g(\mathbf{x}) \, d\mathbf{x}, \quad (2.18)$$

$$g_{\sigma} := \frac{1}{|\sigma|} \int_{\sigma} g(\mathbf{x}) \, d\gamma(\mathbf{x}). \quad (2.19)$$

Then

$$(g_K - g_{\sigma})^2 \leq c_d \frac{\text{diam}(K)^2}{|K|} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x}, \quad (2.20)$$

$$\int_K [g(\mathbf{x}) - g_{\sigma}]^2 \, d\mathbf{x} \leq c_d \text{diam}(K)^2 \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x}, \quad (2.21)$$

where

$$c_d = 6 \text{ for } d = 2, \quad c_d = 9 \text{ for } d = 3. \quad (2.22)$$

PROOF:

The inequality (2.20) is proved as a part of [61, Lemma 9.4] or [59, Lemma 2] for $d = 2$. In these references a general convex polygonal element K is considered; the fact that $c_d = 6$

follows by considering a triangular element. The inequality (2.21) also follows from these proofs, using the Cauchy–Schwarz inequality. We now give the proof for the three-dimensional case, following the ideas of the proof for $d = 2$.

Let us consider a tetrahedron K and its face σ . Let us denote the space coordinates by x_1, x_2, x_3 . We assume, without loss of generality, that $\sigma \subset \{0\} \times \mathbb{R} \times \mathbb{R}^+$, that one vertex of σ lies in the origin, that the longest edge of σ lies on x_2^+ , and that $K \subset \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}$. Let $\mathbf{a} = (\alpha, \beta, \gamma)$ be the vertex that does not lie on σ . For all $x_1 \in [0, \alpha]$, we set $J(x_1) = \{x_2 \in \mathbb{R} \text{ such that } (x_1, x_2, x_3) \in K \text{ for some } x_3 \in \mathbb{R}\}$. For all $x_2 \in J(x_1)$ with $x_1 \in [0, \alpha]$ given, we set $J(x_1, x_2) = \{x_3 \in \mathbb{R} \text{ such that } (x_1, x_2, x_3) \in K\}$. For a.e. $\mathbf{x} = (x_1, x_2, x_3) \in K$ and a.e. $\mathbf{y} = (0, y_2, y_3) \in \sigma$, we set $\mathbf{z}(\mathbf{x}, \mathbf{y}) = t\mathbf{a} + (1-t)\mathbf{y}$ with $t = \frac{x_1}{\alpha}$. Since K is convex, $\mathbf{z}(\mathbf{x}, \mathbf{y}) \in K$ and we have $\mathbf{z}(\mathbf{x}, \mathbf{y}) = (x_1, z_2(x_1, y_2), z_3(x_1, y_3))$ with $z_2(x_1, y_2) = \frac{x_1}{\alpha}\beta + (1 - \frac{x_1}{\alpha})y_2$ and $z_3(x_1, y_3) = \frac{x_1}{\alpha}\gamma + (1 - \frac{x_1}{\alpha})y_3$.

Using the Cauchy–Schwarz inequality, we have

$$\begin{aligned} \int_K [g(\mathbf{x}) - g_\sigma]^2 d\mathbf{x} &= \int_K \left[\frac{1}{|\sigma|} \int_\sigma g(\mathbf{x}) d\gamma(\mathbf{y}) - \frac{1}{|\sigma|} \int_\sigma g(\mathbf{y}) d\gamma(\mathbf{y}) \right. \\ &\quad \left. \pm \frac{1}{|\sigma|} \int_\sigma g(\mathbf{z}(\mathbf{x}, \mathbf{y})) d\gamma(\mathbf{y}) \right]^2 d\mathbf{x} \leq \frac{2}{|\sigma|^2} \int_K \left[\int_\sigma (g(\mathbf{x}) - g(\mathbf{z}(\mathbf{x}, \mathbf{y}))) d\gamma(\mathbf{y}) \right]^2 d\mathbf{x} \\ &\quad + \frac{2}{|\sigma|^2} \int_K \left[\int_\sigma (g(\mathbf{z}(\mathbf{x}, \mathbf{y})) - g(\mathbf{y})) d\gamma(\mathbf{y}) \right]^2 d\mathbf{x} \leq \frac{2}{|\sigma|} (A + B), \end{aligned}$$

where

$$\begin{aligned} A &:= \int_K \int_\sigma (g(\mathbf{x}) - g(\mathbf{z}(\mathbf{x}, \mathbf{y})))^2 d\gamma(\mathbf{y}) d\mathbf{x}, \\ B &:= \int_K \int_\sigma (g(\mathbf{z}(\mathbf{x}, \mathbf{y})) - g(\mathbf{y}))^2 d\gamma(\mathbf{y}) d\mathbf{x}. \end{aligned}$$

Similarly,

$$(g_K - g_\sigma)^2 \leq \frac{2}{|K||\sigma|} (A + B).$$

We denote by $D_i g$ the partial derivative of g with respect to x_i , $i \in \{1, 2, 3\}$, and estimate A and B separately. For this purpose, we suppose that $g \in C^1(K)$ and use the density of $C^1(K)$ in $H^1(K)$ to extend the estimates to $g \in H^1(K)$.

We first estimate A . We have

$$\begin{aligned} A &= \int_0^\alpha \int_{J(x_1)} \int_{J(x_1, x_2)} \int_{J(0)} \int_{J(0, y_2)} \left(g(x_1, x_2, x_3) \right. \\ &\quad \left. - g(x_1, z_2(x_1, y_2), z_3(x_1, y_3)) \right)^2 dy_3 dy_2 dx_3 dx_2 dx_1. \end{aligned}$$

Let us suppose that $x_3 \geq z_3$. This implies that $[x_1, x_2, z_3(x_1, y_3)] \in K$, since the cross-section of K and the plane $x_1 = \text{const}$ is a triangle whose bottom edge is horizontal and the longest of its three edges. We deduce the inequality

$$\begin{aligned} &\left(g(x_1, x_2, x_3) - g(x_1, z_2(x_1, y_2), z_3(x_1, y_3)) \right)^2 = \left(g(x_1, x_2, x_3) - g(x_1, x_2, z_3(x_1, y_3)) \right. \\ &\quad \left. + g(x_1, x_2, z_3(x_1, y_3)) - g(x_1, z_2(x_1, y_2), z_3(x_1, y_3)) \right)^2 = \left(\int_{z_3(x_1, y_3)}^{x_3} D_3 g(x_1, x_2, s) ds \right. \\ &\quad \left. + \int_{z_2(x_1, y_2)}^{x_2} D_2 g(x_1, s, z_3(x_1, y_3)) ds \right)^2 \leq 2 \text{diam}(K) \int_{J(x_1, x_2)} [D_3 g(x_1, x_2, s)]^2 ds \\ &\quad + 2 \text{diam}(K) \left(1 - \frac{x_1}{\alpha} \right) \int_{z_2(x_1, y_2)}^{x_2} [D_2 g(x_1, s, z_3(x_1, y_3))]^2 ds, \end{aligned}$$

where we have used the Newton integration formula and the Cauchy–Schwarz inequality. Defining $D_i g$, $i \in \{1, 2, 3\}$, by 0 outside of K and considering also $x_3 < z_3$, we come to

$$A \leq 2\text{diam}(K)(A_1 + A_2 + A_3 + A_4)$$

with

$$\begin{aligned} A_1 &:= \int_0^\alpha \int_{J(x_1)} \int_{J(x_1, x_2)} \int_{J(0)} \int_{J(0, y_2)} \int_{J(x_1, x_2)} [D_3 g(x_1, x_2, s)]^2 ds dy_3 dy_2 dx_3 dx_2 dx_1, \\ A_2 &:= \int_0^\alpha \int_{J(x_1)} \int_{J(x_1, x_2)} \int_{J(0)} \int_{J(0, y_2)} \left(1 - \frac{x_1}{\alpha}\right) \\ &\quad \int_{z_2(x_1, y_2)}^{x_2} [D_2 g(x_1, s, z_3(x_1, y_3))]^2 ds dy_3 dy_2 dx_3 dx_2 dx_1, \\ A_3 &:= \int_0^\alpha \int_{J(x_1)} \int_{J(x_1, x_2)} \int_{J(0)} \int_{J(0, y_2)} \int_{z_2(x_1, y_2)}^{x_2} [D_2 g(x_1, s, x_3)]^2 ds dy_3 dy_2 dx_3 dx_2 dx_1, \\ A_4 &:= \int_0^\alpha \int_{J(x_1)} \int_{J(x_1, x_2)} \int_{J(0)} \int_{J(0, y_2)} \left(1 - \frac{x_1}{\alpha}\right) \\ &\quad \int_{z_3(x_1, y_3)}^{x_3} [D_3 g(x_1, z_2(x_1, y_2), s)]^2 ds dy_3 dy_2 dx_3 dx_2 dx_1. \end{aligned}$$

We easily see that

$$A_1 \leq \text{diam}(K)|\sigma| \int_K [D_3 g(\mathbf{x})]^2 d\mathbf{x}.$$

Next, we estimate A_2 . Using the Fubini theorem and the change of variables $z_3 = z_3(x_1, y_3)$, we have

$$\begin{aligned} &\int_{J(0, y_2)} \left(1 - \frac{x_1}{\alpha}\right) \int_{z_2(x_1, y_2)}^{x_2} [D_2 g(x_1, s, z_3(x_1, y_3))]^2 ds dy_3 \\ &= \int_{z_2(x_1, y_2)}^{x_2} \int_{J(x_1, z_2(x_1, y_2))} [D_2 g(x_1, s, z_3)]^2 dz_3 ds \\ &\leq \int_{J(x_1)} \int_{J(x_1, s)} [D_2 g(x_1, s, z_3)]^2 dz_3 ds, \end{aligned}$$

where the estimate follows by extending the integration region. Hence

$$A_2 \leq \text{diam}(K)|\sigma| \int_K [D_2 g(\mathbf{x})]^2 d\mathbf{x}.$$

Using the Fubini theorem, we similarly estimate A_3 and A_4 ,

$$\begin{aligned} A_3 &\leq \text{diam}(K)|\sigma| \int_K [D_2 g(\mathbf{x})]^2 d\mathbf{x}, \\ A_4 &\leq \text{diam}(K)|\sigma| \int_K [D_3 g(\mathbf{x})]^2 d\mathbf{x}, \end{aligned}$$

which finally yields

$$A \leq 4\text{diam}(K)^2 |\sigma| \int_K |\nabla g(\mathbf{x})|^2 d\mathbf{x}. \quad (2.23)$$

We now turn to the study of B . We write it as

$$B = \int_0^\alpha \int_{J(x_1)} \int_{J(x_1, x_2)} \int_{J(0)} \int_{J(0, y_2)} \left(g(x_1, z_2(x_1, y_2), z_3(x_1, y_3)) - g(0, y_2, y_3) \right)^2 dy_3 dy_2 dx_3 dx_2 dx_1.$$

Using the Newton integration formula and the Cauchy–Schwarz and Hölder inequalities, we have

$$\begin{aligned} & \left(g(x_1, z_2(x_1, y_2), z_3(x_1, y_3)) - g(0, y_2, y_3) \right)^2 = \left(\int_0^{x_1} \left[D_1 g(s, z_2(s, y_2), z_3(s, y_3)) \right. \right. \\ & \left. \left. + D_2 g(s, z_2(s, y_2), z_3(s, y_3)) \frac{\beta - y_2}{\alpha} + D_3 g(s, z_2(s, y_2), z_3(s, y_3)) \frac{\gamma - y_3}{\alpha} \right] ds \right)^2 \\ & \leq \alpha \left(1 + \left(\frac{\beta - y_2}{\alpha} \right)^2 + \left(\frac{\gamma - y_3}{\alpha} \right)^2 \right) \int_0^{x_1} \sum_{i=1}^3 [D_i g(s, z_2(s, y_2), z_3(s, y_3))]^2 ds. \end{aligned}$$

Hence

$$B \leq \alpha \left(1 + \left(\frac{\beta - y_2}{\alpha} \right)^2 + \left(\frac{\gamma - y_3}{\alpha} \right)^2 \right) \sum_{i=1}^3 B_i$$

with

$$B_i = \int_0^\alpha \int_{J(x_1)} \int_{J(x_1, x_2)} \int_{J(0)} \int_{J(0, y_2)} \int_0^{x_1} [D_i g(s, z_2(s, y_2), z_3(s, y_3))]^2 ds dy_3 dy_2 dx_3 dx_2 dx_1,$$

$i \in \{1, 2, 3\}$. Using the Fubini theorem, we have

$$B_i = \int_{J(0)} \int_{J(0, y_2)} \int_0^\alpha [D_i g(s, z_2(s, y_2), z_3(s, y_3))]^2 \int_s^\alpha \int_{J(x_1)} \int_{J(x_1, x_2)} dx_3 dx_2 dx_1 ds dy_3 dy_2.$$

Hence

$$B_i \leq \frac{|\sigma|}{2\alpha} \int_0^\alpha \int_{J(0)} \int_{J(0, y_2)} [D_i g(s, z_2(s, y_2), z_3(s, y_3))]^2 (\alpha - s)^2 dy_3 dy_2 ds,$$

where we have used the estimate

$$\int_{J(x_1)} \int_{J(x_1, x_2)} dx_3 dx_2 \leq |\sigma| \left(1 - \frac{x_1}{\alpha} \right)$$

on the area of the cross-section of K and the plane $x_1 = \text{const}$. Now using the change of variables $z_3 = z_3(s, y_3)$ and $z_2 = z_2(s, y_2)$ gives

$$\begin{aligned} & \int_{J(0)} \int_{J(0, y_2)} [D_i g(s, z_2(s, y_2), z_3(s, y_3))]^2 (\alpha - s)^2 dy_3 dy_2 \\ & = \alpha^2 \int_{J(s)} \int_{J(s, z_2)} [D_i g(s, z_2, z_3)]^2 dz_3 dz_2 \end{aligned}$$

and thus

$$B_i \leq \frac{|\sigma|\alpha}{2} \int_K [D_i g(\mathbf{x})]^2 d\mathbf{x},$$

which finally yields, noticing that $\alpha^2 + (\beta - y_2)^2 + (\gamma - y_3)^2 = |\mathbf{a} - \mathbf{y}|^2 \leq \text{diam}(K)^2$,

$$B \leq \frac{|\sigma|}{2} \text{diam}(K)^2 \int_K |\nabla g(\mathbf{x})|^2 d\mathbf{x}. \quad (2.24)$$

Now combining (2.23) and (2.24) leads to the assertion of the lemma for $d = 3$. \square

2.5 Discrete Friedrichs inequality

We prove in this section the discrete Friedrichs inequality, using the results of the previous sections. We first give several auxiliary lemmas.

Lemma 2.5.1. *Let $d = 2$. Then*

$$|I(g)|_{1,\mathcal{T},*}^2 \leq \frac{C_d}{\kappa_{\mathcal{T}}^2} |g|_{1,\mathcal{T}}^2 \quad \forall g \in W(\mathcal{T}_h),$$

where C_d is given by (2.26) below.

PROOF:

Let $K \in \mathcal{T}_h$ and $\sigma_D, \sigma_E \in \mathcal{E}_K$. We define g_K by (2.18) and deduce from the inequality $(a - b)^2 \leq 2a^2 + 2b^2$ and from (2.20) that

$$(g_E - g_D)^2 \leq 2(g_E - g_K)^2 + 2(g_D - g_K)^2 \leq 4c_d \frac{\text{diam}(K)^2}{|K|} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x}. \quad (2.25)$$

Using this, the definition of $|\cdot|_{1,\mathcal{T},*}$, $|\sigma_{D,E}| \leq 2/3 \text{diam}(K)$, (2.9) and (2.8), the fact that each $K \in \mathcal{T}_h$ contains exactly three dual edges, and Assumption (A), we have

$$\begin{aligned} |I(g)|_{1,\mathcal{T},*}^2 &= \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|}{v_{D,E}} (g_E - g_D)^2 \\ &\leq 4c_d \sum_{K \in \mathcal{T}_h} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \sigma_{D,E} \subset K} \frac{|\sigma_{D,E}|^2}{v_{D,E} |\sigma_{D,E}|} \frac{\text{diam}(K)^2}{|K|} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \\ &\leq 8c_d \sum_{K \in \mathcal{T}_h} \left[\frac{\text{diam}(K)^2}{|K|} \right]^2 \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \leq \frac{8c_d}{\kappa_{\mathcal{T}}^2} \sum_{K \in \mathcal{T}_h} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x}. \quad \square \end{aligned}$$

Lemma 2.5.2. *There holds*

$$|I(g)|_{1,\mathcal{T},\dagger}^2 \leq \frac{C_d}{\kappa_{\mathcal{T}}} |g|_{1,\mathcal{T}}^2 \quad \forall g \in W(\mathcal{T}_h),$$

where

$$C_d = 8c_d \text{ for } d = 2, \quad C_d = \frac{27}{4}c_d \text{ for } d = 3, \quad (2.26)$$

and c_d is given by (2.22).

PROOF:

Using the definition of $|\cdot|_{1,\mathcal{T},\dagger}$, (2.25), $|\sigma_{D,E}| \leq C_d^* \text{diam}(K_{D,E})^{d-1}$ with $C_d^* = 2/3$ if $d = 2$ and $C_d^* = 9/32$ if $d = 3$, the fact that each $K \in \mathcal{T}_h$ contains $\binom{d+1}{2} = \frac{(d+1)d}{2}$ dual sides, and Assumption (A), we have

$$\begin{aligned} |I(g)|_{1,\mathcal{T},\dagger}^2 &= \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|}{\text{diam}(K_{D,E})} (g_E - g_D)^2 \\ &\leq 4c_d \sum_{K \in \mathcal{T}_h} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \sigma_{D,E} \subset K} \frac{|\sigma_{D,E}| \text{diam}(K)}{|K|} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \\ &\leq 2c_d(d+1)dC_d^* \sum_{K \in \mathcal{T}_h} \frac{\text{diam}(K)^d}{|K|} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \leq \frac{C_d}{\kappa_{\mathcal{T}}} \sum_{K \in \mathcal{T}_h} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x}. \quad \square \end{aligned}$$

Lemma 2.5.3. (Interpolation estimate) *There holds*

$$\|g - I(g)\|_{0,\Omega}^2 \leq c_d h^2 |g|_{1,\mathcal{T}}^2 \quad \forall g \in W(\mathcal{T}_h).$$

PROOF:

We have

$$\begin{aligned} \|g - I(g)\|_{0,\Omega}^2 &= \sum_{K \in \mathcal{T}_h} \sum_{\sigma_D \in \mathcal{E}_K} \int_{K \cap D} [g(\mathbf{x}) - g_D]^2 \, d\mathbf{x} \\ &\leq c_d \sum_{K \in \mathcal{T}_h} \sum_{\sigma_D \in \mathcal{E}_K} [\text{diam}(K \cap D)]^2 \int_{K \cap D} |\nabla g(\mathbf{x})|^2 \, d\mathbf{x} \\ &\leq c_d h^2 \sum_{K \in \mathcal{T}_h} \int_K |\nabla g(\mathbf{x})|^2 \, d\mathbf{x}, \end{aligned}$$

using the estimate (2.21) for the simplex $K \cap D$ and $\text{diam}(K \cap D) \leq h$. \square

We state below the first of the two main results of this chapter.

Theorem 2.5.4. (Discrete Friedrichs inequality) *There holds*

$$\|g\|_{0,\Omega}^2 \leq C_F |g|_{1,\mathcal{T}}^2 \quad \forall g \in W_0(\mathcal{T}_h), \quad \forall h > 0$$

with

$$C_F = \frac{C_d}{\kappa_{\mathcal{T}}^2} |\Omega| + 2c_d h^2 \quad \text{for } d = 2, \quad C_F = 2C_d \frac{C_{d,\mathcal{T}}}{\kappa_{\mathcal{T}}} \left[\inf_{\mathbf{b} \in \mathbb{R}^d} \{\text{diam}_{\mathbf{b}}(\Omega)\} \right]^2 + 2c_d h^2 \quad \text{for } d = 2, 3,$$

where $C_{d,\mathcal{T}}$ is given by (2.12) when Assumption (B) is satisfied and by (2.15) in the general case, c_d is given by (2.22), and C_d is given by (2.26).

PROOF:

One has

$$\|g\|_{0,\Omega}^2 \leq 2\|g - I(g)\|_{0,\Omega}^2 + 2\|I(g)\|_{0,\Omega}^2.$$

The error $\|g - I(g)\|_{0,\Omega}^2$ of the approximation follows from Lemma 2.5.3. Note that $I(g) \in Y_0(\mathcal{D}_h)$ and hence the discrete Friedrichs inequality for piecewise constant functions given by Theorem 2.3.1 together with Lemma 2.5.1 yield

$$\|I(g)\|_{0,\Omega}^2 \leq \frac{C_d}{2\kappa_{\mathcal{T}}^2} |\Omega| |g|_{1,\mathcal{T}}^2$$

for the case where $d = 2$. Similarly, using the discrete Friedrichs inequality for piecewise constant functions given by Theorem 2.3.5 together with Lemma 2.5.2, one has

$$\|I(g)\|_{0,\Omega}^2 \leq C_d \frac{C_{d,\mathcal{T}}}{\kappa_{\mathcal{T}}} [\text{diam}_{\mathbf{b}}(\Omega)]^2 |g|_{1,\mathcal{T}}^2$$

for an arbitrary vector $\mathbf{b} \in \mathbb{R}^d$ for the case where $d = 2, 3$. \square

Remark 2.5.5. (Dependence of C_F on Ω) *We have $h^2 \leq |\Omega|/\kappa_{\mathcal{T}}$ by Assumption (A) and $h \leq \theta_{\mathcal{T}} \text{diam}_{\mathbf{b}}(\Omega)$ by the consequence (2.5) of Assumption (A). Hence the constant in the discrete Friedrichs inequality only depends on the area of Ω if $d = 2$ and on the square of the infimum of the diameters of Ω in one direction if $d = 2, 3$. This dependence is optimal: [91, Theorem 1.1] gives the same dependence for the Friedrichs inequality and $H_0^1(\Omega) \subset W_0(\mathcal{T}_h)$.*

Remark 2.5.6. (Dependence of C_F on the shape regularity parameter) *One can see that C_F depends on $1/\kappa_{\mathcal{T}}^2$ if $d = 2$ and when it is expressed using $|\Omega|$. We are able to establish the same result also when C_F is expressed using $\inf_{\mathbf{b} \in \mathbb{R}^d} \text{diam}_{\mathbf{b}}(\Omega)$, but only when the meshes are not locally refined (when Assumption (B) is satisfied). Indeed, C_F in this case depends on $C_{d,\mathcal{T}}/\kappa_{\mathcal{T}}$ and the constant $C_{d,\mathcal{T}}$ given by (2.12) is of the form $[2^d(d-1)\zeta_{\mathcal{T}}^d(2C+1)]/\kappa_{\mathcal{T}}$; this follows by replacing the inequality (2.13) by $h \leq C \text{diam}_{\mathbf{b}}(\Omega)$ for some suitable constant C . Example 2.6.3 below shows that this dependence is optimal. However, in the case where the meshes are only shape-regular, we were only able to establish (2.15).*

Remark 2.5.7. (Discrete Friedrichs inequality for domains only bounded in one direction) *We see that the constant C_F only depends on the infimum of the diameters of Ω in one direction. Thus the discrete Friedrichs inequality may be extended onto domains only bounded in one direction, as it is the case for the Friedrichs inequality (cf. [91, Remark 1.1]).*

Remark 2.5.8. (Discrete Friedrichs inequality for functions only fixed to zero on a particular part of the boundary) *Let $\Gamma \subset \partial\Omega$ (given by a set of boundary sides) be such that there exists a vector $\mathbf{b} \in \mathbb{R}^d$ such that the first intersection of $\mathcal{B}_{\mathbf{x}}$ and $\partial\Omega$ lies in Γ for all $\mathbf{x} \in \Omega$, where $\mathcal{B}_{\mathbf{x}}$ is a straight semi-line defined by the origin \mathbf{x} and the vector \mathbf{b} . We notice that the discrete Friedrichs inequality can be immediately extended onto functions only fixed to zero on Γ . This follows easily from the proof of Theorem 2.3.5 (the zero condition is only used on boundary sides lying in Γ). The dependence of C_F on the shape regularity parameter is thus given by $C_{d,\mathcal{T}}/\kappa_{\mathcal{T}}$, cf. Remark 2.5.6. The constant C_F in this case depends on the square of the infimum of $\text{diam}_{\mathbf{b}}(\Omega)$ over suitable vectors \mathbf{b} (compare with the general case treated in the next remark).*

Remark 2.5.9. (Discrete Friedrichs inequality for functions only fixed to zero on a general part of the boundary) *The discrete Friedrichs inequality can also be extended onto functions only fixed to zero on an arbitrary set of boundary sides, see Lemma 2.7.2 and Remark 2.7.3 below. Then, for convex domains, C_F depends on the square of the diameter of Ω ; for nonconvex domains, the dependence of C_F on Ω is more complicated. The dependence of C_F on the shape regularity parameter again reveals given by $C_{d,\mathcal{T}}/\kappa_{\mathcal{T}}$.*

2.6 Discrete Friedrichs inequality for Crouzeix–Raviart finite elements

We show in this section how the proofs from the previous sections simplify for the case of Crouzeix–Raviart finite elements in two space dimensions. Let us consider the space $X(\mathcal{T}_h)$ introduced in Section 2.2. The basis of this space is spanned by the shape functions φ_D , $D \in \mathcal{D}_h$, such that $\varphi_D(Q_E) = \delta_{DE}$, $E \in \mathcal{D}_h$, δ being the Kronecker delta.

Lemma 2.6.1. *Let $d = 2$. Then for all $c \in X(\mathcal{T}_h)$,*

$$\|c\|_{0,\Omega} = \|I(c)\|_{0,\Omega}.$$

PROOF:

Let us write $c = \sum_{D \in \mathcal{D}_h} c_D \varphi_D$. Using that the quadrature formula $\int_K \psi \, d\mathbf{x} \approx \frac{1}{3}|K| \sum_{\sigma_D \in \mathcal{E}_K} \psi(Q_D)$ is exact for quadratic functions on triangles and (2.8), we have

$$\int_{\Omega} c^2(\mathbf{x}) \, d\mathbf{x} = \sum_{K \in \mathcal{T}_h} \int_K c^2(\mathbf{x}) \, d\mathbf{x} = \sum_{K \in \mathcal{T}_h} \frac{1}{3}|K| \sum_{\sigma_D \in \mathcal{E}_K} c^2(Q_D) = \sum_{D \in \mathcal{D}_h} c_D^2 |D|. \quad \square$$

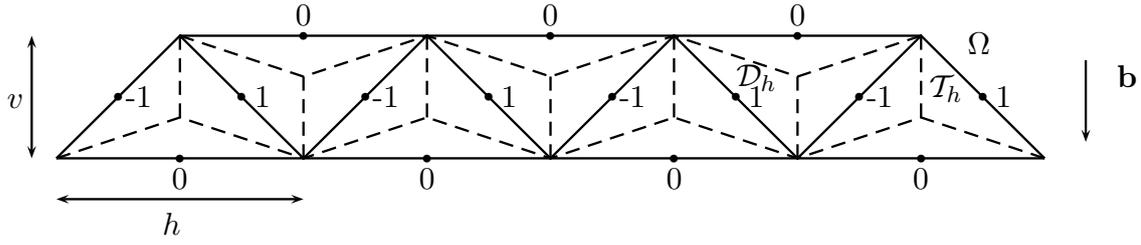


Figure 2.2: Domain Ω , triangulation \mathcal{T}_h , dual mesh \mathcal{D}_h , and values of a function $c \in X(\mathcal{T}_h)$ for the optimality example

Lemma 2.6.2. (Discrete Friedrichs inequality for Crouzeix–Raviart finite elements in two space dimensions) *Let $d = 2$. Then*

$$\|c\|_{0,\Omega}^2 \leq C_F |c|_{1,\mathcal{T}}^2 \quad \forall c \in X_0(\mathcal{T}_h), \forall h > 0,$$

where

$$C_F = \frac{1}{4\kappa_{\mathcal{T}}^2} |\Omega| \quad \text{or} \quad C_F = \frac{C_{d,\mathcal{T}}}{2\kappa_{\mathcal{T}}} \left[\inf_{\mathbf{b} \in \mathbb{R}^d} \{\text{diam}_{\mathbf{b}}(\Omega)\} \right]^2.$$

PROOF:

Let $c \in X_0(\mathcal{T}_h)$, $c = \sum_{D \in \mathcal{D}_h} c_D \varphi_D$. Note that by the definition of $X_0(\mathcal{T}_h)$, $c_D = 0$ for all $D \in \mathcal{D}_h^{\text{ext}}$. Using respectively Lemma 2.6.1 and Theorem 2.3.1 or Theorem 2.3.5, we get

$$\|c\|_{0,\Omega}^2 \leq \frac{|\Omega|}{2} |I(c)|_{1,\mathcal{T},*}^2, \quad \|c\|_{0,\Omega}^2 \leq C_{d,\mathcal{T}} [\text{diam}_{\mathbf{b}}(\Omega)]^2 |I(c)|_{1,\mathcal{T},\dagger}^2$$

for an arbitrary vector $\mathbf{b} \in \mathbb{R}^2$. Finally, we deduce that

$$\begin{aligned} |I(c)|_{1,\mathcal{T},*}^2 &= \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{|\sigma_{D,E}|^2}{v_{D,E} |\sigma_{D,E}|} \left(\nabla c|_{K_{D,E}} \cdot (Q_E - Q_D) \right)^2 \\ &\leq \frac{2}{3} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{\text{diam}(K_{D,E})^2}{|K_{D,E}|} |\nabla c|_{K_{D,E}}|^2 d_{D,E}^2 \\ &\leq \frac{1}{2\kappa_{\mathcal{T}}^2} \sum_{K \in \mathcal{T}_h} |\nabla c|_K|^2 |K| = \frac{1}{2\kappa_{\mathcal{T}}^2} |c|_{1,\mathcal{T}}^2, \end{aligned}$$

using (2.9) and (2.8), $|\sigma_{D,E}| \leq 2/3 \text{diam}(K_{D,E})$, the fact that the gradient of c is elementwise constant and that each $K \in \mathcal{T}_h$ contains exactly three dual edges, $d_{D,E} \leq \text{diam}(K_{D,E})/2$, and Assumption (A). Similarly, $|I(c)|_{1,\mathcal{T},\dagger}^2 \leq 1/(2\kappa_{\mathcal{T}}) |c|_{1,\mathcal{T}}^2$. \square

Example 2.6.3. (Optimality of the dependence of C_F on the shape regularity parameter) *Let us consider a domain Ω , its triangulation \mathcal{T}_h , a vector \mathbf{b} , and a function $c \in X(\mathcal{T}_h)$ given by the values 0, 1, -1 as depicted in Figure 2.2. Using Lemma 2.6.1, we immediately have*

$$\|c\|_{0,\Omega}^2 = \sum_{K \in \mathcal{T}_h} \frac{1}{3} |K| (0 + 1 + 1) = \frac{2}{3} |\Omega|.$$

On each $K \in \mathcal{T}_h$, $|\nabla c|_K| = 4/h$, hence $|c|_{1,\mathcal{T}}^2 = 16/h^2 |\Omega|$. Using Remark 2.5.8, the discrete Friedrichs inequality given by Lemma 2.6.2 holds true. The term occurring on its right hand

side, independent of the shape regularity parameter, is $1/2[\text{diam}_{\mathbf{b}}(\Omega)]^2|c|_{1,\mathcal{T}}^2 = 8v^2/h^2|\Omega|$. This term can be arbitrarily smaller than $\|c\|_{0,\Omega}^2$, letting $h \rightarrow +\infty$ or $v \rightarrow 0$. Next, $\kappa_{\mathcal{T}} = v/(2h)$. Note that \mathcal{T}_h satisfies Assumption (B) and hence $C_{d,\mathcal{T}} \approx 1/\kappa_{\mathcal{T}}$. In fact, by a simple estimation of the term A from Lemma 2.3.3, one has $C_{d,\mathcal{T}} = 1/\kappa_{\mathcal{T}}$ in this case and thus $C_{d,\mathcal{T}}/\kappa_{\mathcal{T}} = 1/\kappa_{\mathcal{T}}^2 = 4h^2/v^2$. One immediately sees that the multiplication by this term is necessary.

Corollary 2.6.4. (Discrete Friedrichs inequality for Crouzeix–Raviart finite elements on equilateral triangles) *Let $d = 2$ and let \mathcal{T}_h consist of equilateral triangles. Then*

$$\|c\|_{0,\Omega}^2 \leq C_F |c|_{1,\mathcal{T}}^2 \quad \forall c \in X_0(\mathcal{T}_h), \forall h > 0,$$

where

$$C_F = \frac{|\Omega|}{2} \quad \text{or} \quad C_F = \left[\inf_{\mathbf{b} \in \mathbb{R}^d} \{\text{diam}_{\mathbf{b}}(\Omega)\} + 2h \right]^2.$$

PROOF:

Let c be as in the previous lemma. For equilateral triangles, one has $d_{D,E} = v_{D,E}$ and thus the norms $|\cdot|_{1,\mathcal{T},*}$ and $|\cdot|_{1,\mathcal{T},\dagger}$ coincide. By (2.9) and (2.8), $|\sigma_{D,E}|v_{D,E} = 2/3|K|$, $\cos^2(\alpha) + \cos^2(\alpha + \pi/3) + \cos^2(\alpha + 2\pi/3) = 3/2$, so that

$$\sum_{K \in \mathcal{T}_h} \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}, \sigma_{D,E} \subset K} \frac{|\sigma_{D,E}|}{d_{D,E}} \left| \nabla c|_K \right|^2 d_{D,E}^2 \cos^2(\nabla c|_K, Q_E - Q_D) = \sum_{K \in \mathcal{T}_h} \left| \nabla c|_K \right|^2 |K|.$$

Now using respectively Lemma 2.6.1 and Theorem 2.3.1 or Remark 2.3.2 yields the assertion. \square

Remark 2.6.5. (C_F for Crouzeix–Raviart finite elements on equilateral triangles) *Let $d = 2$. Then the constant in the Friedrichs inequality may be expressed as $c_F = |\Omega|/2$ or $c_F = [\inf_{\mathbf{b} \in \mathbb{R}^d} \{\text{diam}_{\mathbf{b}}(\Omega)\}]^2$, cf. [91, Theorem 1.1]. Corollary 2.6.4 shows that for Crouzeix–Raviart finite elements and equilateral triangles, we are able to achieve the same result (up to h) also for the constant C_F from the discrete Friedrichs inequality. We however remark that there exist sharper estimates in the continuous case, see e.g. [107].*

2.7 Discrete Poincaré inequality for piecewise constant functions

As in the case of the discrete Friedrichs inequality, we start with the discrete Poincaré inequality for piecewise constant functions. [61, Lemma 10.2] states the discrete Poincaré inequality for piecewise constant functions on meshes satisfying the orthogonality property. We present in this section an analogy of this lemma for the mesh \mathcal{D}_h , where the orthogonality property is not necessarily satisfied.

Lemma 2.7.1. *Let ω be an open convex subset of Ω , $\omega \neq \emptyset$, and let*

$$m_{\omega}(c) := \frac{1}{|\omega|} \int_{\omega} c(\mathbf{x}) \, d\mathbf{x}.$$

Then for all $c \in Y(\mathcal{D}_h)$,

$$\|c - m_{\omega}(c)\|_{0,\omega}^2 \leq \frac{|B_{\Omega}|}{|\omega|} C_{d,\mathcal{T}} [\text{diam}(\Omega)]^2 |c|_{1,\mathcal{T},\dagger}^2,$$

where B_{Ω} is the ball of \mathbb{R}^d with center 0 and radius $\text{diam}(\Omega)$ and $C_{d,\mathcal{T}}$ is given by (2.12) when Assumption (B) is satisfied and by (2.15) in the general case.

The proof of this lemma follows the proof of the first step of [61, Lemma 10.2], using the techniques introduced in Section 2.3 for meshes where the orthogonality property is not satisfied.

Lemma 2.7.2. *Let ω be a polygonal open convex subset of Ω and let $I \subset \partial\omega$ with $|I| > 0$. Let $E \in \mathcal{D}_h$ be such that $I \cap E$ is not an empty set and not just a point (such dual element always exists). Then for all $c \in Y(\mathcal{D}_h)$ with $c_E = 0$,*

$$\|c\|_{0,\omega}^2 \leq C_{d,\mathcal{T}}[\text{diam}(\Omega)]^2 |c|_{1,\mathcal{T},\dagger}^2,$$

where $C_{d,\mathcal{T}}$ is given by (2.12) when Assumption (B) is satisfied and by (2.15) in the general case.

PROOF:

The proof is similar to that of Theorem 2.3.5. There exist a set of vectors \mathbf{b}_i and of nonempty nonoverlapping subsets ω_i of ω , $i = 1, \dots, M$ (M may be equal to $+\infty$) with the following properties: (i) \mathbf{b}_i is such that $\mathcal{C}_{\mathbf{b}_i} \cap \omega \neq \emptyset$, where $\mathcal{C}_{\mathbf{b}_i}$ is the cylinder whose basis is $I \cap E$ and generator vector is $-\mathbf{b}_i$; (ii) $\omega_i = \mathcal{C}_{\mathbf{b}_i} \cap \omega \setminus \cup_{j=1}^{i-1} \omega_j$; (iii) $\cup_{i=1}^M \omega_i = \omega$. Note that the fact that ω is convex is important. For all $\mathbf{x} \in \omega_i$, we set $\mathcal{B}_{\mathbf{x}}^i$ as the straight semi-line defined by the origin \mathbf{x} and the vector \mathbf{b}_i . Let $\mathbf{y}(\mathbf{x}) = I \cap E \cap \mathcal{B}_{\mathbf{x}}^i$. Let the function $\chi_{\sigma}(\mathbf{x})$ be given by (2.16) for each $\sigma \in \mathcal{F}_h^{\text{int}}$. Let $D \in \mathcal{D}_h$, $D \cap \omega_i \neq \emptyset$, be fixed. We then have (2.17) for a.e. $\mathbf{x} \in D \cap \omega_i$, as in Theorem 2.3.5. Integrating (2.17) over $D \cap \omega_i$, summing over all $D \in \mathcal{D}_h$ such that $D \cap \omega_i \neq \emptyset$, and using Lemma 2.3.3 when Assumption (B) is satisfied and Lemma 2.3.4 in the general case yields

$$\sum_{D \in \mathcal{D}_h} |c_D|^2 |D \cap \omega_i| \leq C_{d,\mathcal{T}} \text{diam}(\Omega) \sum_{\sigma_{D,E} \in \mathcal{F}_h^{\text{int}}} \frac{(c_E - c_D)^2}{\text{diam}(K_{D,E})} \int_{\omega_i} \chi_{\sigma_{D,E}}(\mathbf{x}) \, d\mathbf{x}.$$

Using the inequality

$$\int_{\omega_i} \chi_{\sigma_{D,E}}(\mathbf{x}) \, d\mathbf{x} \leq |\sigma_{D,E} \cap \omega_i| \text{diam}(\omega),$$

$\text{diam}(\omega) \leq \text{diam}(\Omega)$, and summing over all i concludes the proof. \square

Remark 2.7.3. *Let Ω be convex. Then taking $\omega = \Omega$ in Lemma 2.7.2, we have an extension of Theorem 2.3.5 onto functions from $Y(\mathcal{D}_h)$ that vanish on only one boundary dual element.*

Remark 2.7.4. *Lemma 2.7.2 is an alternative to the second step of the proof of [61, Lemma 10.2].*

Theorem 2.7.5. (Discrete Poincaré inequality for piecewise constant functions) *Let*

$$m_{\Omega}(c) := \frac{1}{|\Omega|} \int_{\Omega} c(\mathbf{x}) \, d\mathbf{x}.$$

Then for all $c \in Y(\mathcal{D}_h)$,

$$\|c - m_{\Omega}(c)\|_{0,\Omega}^2 \leq C_{\Omega} C_{d,\mathcal{T}} [\text{diam}(\Omega)]^2 |c|_{1,\mathcal{T},\dagger}^2,$$

where

$$C_{\Omega} = \frac{|B_{\Omega}|}{|\Omega|} \tag{2.27}$$

when Ω is convex and

$$C_\Omega = 2 \sum_{i=1}^n \frac{|B_\Omega|}{|\Omega_i|} + 16(n-1)^2 \frac{|\Omega|}{|\Omega_i|_{\min}} \left(\frac{|B_\Omega|}{|\Omega_i|_{\min}} + 1 \right) \quad (2.28)$$

when Ω is not convex but there exists a finite number of disjoint open convex polygonal sets Ω_i such that $\overline{\Omega} = \cup_{i=1}^n \overline{\Omega_i}$. Here, $|\Omega_i|_{\min} = \min_{i=1, \dots, n} \{|\Omega_i|\}$, B_Ω is the ball of \mathbb{R}^d with center 0 and radius $\text{diam}(\Omega)$, and $C_{d, \mathcal{T}}$ is given by (2.12) when Assumption (B) is satisfied and by (2.15) in the general case.

PROOF:

When Ω is convex, the assertion of this theorem coincides with that of Lemma 2.7.1 for $\omega = \Omega$. When Ω is not convex, we have Lemmas 2.7.1 and 2.7.2 for each Ω_i . Then the third step of the proof of [61, Lemma 10.2] yields the assertion of the theorem. \square

Remark 2.7.6. One has

$$\|c\|_{0, \Omega}^2 \leq 2\|c - m_\Omega(c)\|_{0, \Omega}^2 + 2\|m_\Omega(c)\|_{0, \Omega}^2.$$

Hence Theorem 2.7.5 implies the discrete Poincaré inequality for piecewise constant functions in the more common form

$$\|c\|_{0, \Omega}^2 \leq 2C_\Omega C_{d, \mathcal{T}} [\text{diam}(\Omega)]^2 |c|_{1, \mathcal{T}, \dagger}^2 + \frac{2}{|\Omega|} \left(\int_\Omega c(\mathbf{x}) \, d\mathbf{x} \right)^2 \quad \forall c \in Y(\mathcal{D}_h), \forall h > 0.$$

2.8 Discrete Poincaré inequality

We state below the second of the two main results of this chapter.

Theorem 2.8.1. (Discrete Poincaré inequality) *There holds*

$$\|g\|_{0, \Omega}^2 \leq C_P |g|_{1, \mathcal{T}}^2 + \frac{4}{|\Omega|} \left(\int_\Omega g(\mathbf{x}) \, d\mathbf{x} \right)^2 \quad \forall g \in W(\mathcal{T}_h), \forall h > 0$$

with

$$C_P = 4C_d C_\Omega \frac{C_{d, \mathcal{T}}}{\kappa_{\mathcal{T}}} [\text{diam}(\Omega)]^2 + 8c_d h^2,$$

where C_Ω is given by (2.27) when Ω is convex and by (2.28) otherwise, $C_{d, \mathcal{T}}$ is given by (2.12) when Assumption (B) is satisfied and by (2.15) in the general case, c_d is given by (2.22), and C_d is given by (2.26).

PROOF:

One has

$$\|g\|_{0, \Omega}^2 \leq 4\|g - I(g)\|_{0, \Omega}^2 + 4\|I(g) - m_\Omega[I(g)]\|_{0, \Omega}^2 + 4\|m_\Omega[I(g)] - m_\Omega(g)\|_{0, \Omega}^2 + 4\|m_\Omega(g)\|_{0, \Omega}^2,$$

where $m_\Omega(f) = 1/|\Omega| \int_\Omega f(\mathbf{x}) \, d\mathbf{x}$. The discrete Poincaré inequality for piecewise constant functions given by Theorem 2.7.5 and Lemma 2.5.2 imply

$$\|I(g) - m_\Omega[I(g)]\|_{0, \Omega}^2 \leq C_d C_\Omega \frac{C_{d, \mathcal{T}}}{\kappa_{\mathcal{T}}} [\text{diam}(\Omega)]^2 |g|_{1, \mathcal{T}}^2.$$

We have

$$\|m_\Omega[I(g)] - m_\Omega(g)\|_{0, \Omega}^2 \leq \|g - I(g)\|_{0, \Omega}^2$$

by the Cauchy–Schwarz inequality. Finally, the error $\|g - I(g)\|_{0, \Omega}^2$ of the approximation follows from Lemma 2.5.3. \square

Remark 2.8.2. (Dependence of C_P on Ω) *Let Ω be a cube. We then have $h \leq \text{diam}(\Omega)$ and $C_\Omega \leq \pi\sqrt{3}/2$ and hence the constant in the discrete Poincaré inequality in this case only depends on the square of the diameter of Ω . This dependence is optimal: [91, Theorem 1.3] gives the same dependence for the Poincaré inequality and $H^1(\Omega) \subset W(\mathcal{T}_h)$.*

Remark 2.8.3. (Dependence of C_P on the shape regularity parameter) *Our results indicate that the dependence of C_P on the shape regularity parameter is given by $C_{d,\mathcal{T}}/\kappa_{\mathcal{T}}$, cf. Remark 2.5.6.*

Chapitre 3

Equivalence between lowest-order mixed finite element and multi-point finite volume methods

We consider in this chapter the lowest-order Raviart–Thomas mixed finite element method for elliptic diffusion problems on simplicial meshes in two or three space dimensions. This method produces saddle-point problems for scalar and flux unknowns. We show how to easily eliminate the flux unknowns, which implies equivalence between this method and a particular multi-point finite volume scheme, without any approximate numerical integration. The matrix of the final linear system is sparse, positive definite for a large class of problems, but in general nonsymmetric. We next show that these ideas also apply to mixed and upwind-mixed finite element discretizations of nonlinear parabolic convection–reaction–diffusion problems. We present a set of numerical experiments confirming important computational savings while using the equivalent finite volume form of the lowest-order mixed finite element method and compare it to a finite volume and combined finite volume–finite element schemes.

3.1 Introduction

Let us consider the elliptic problem

$$\mathbf{u} = -\mathbf{S}\nabla p \quad \text{in } \Omega, \quad (3.1a)$$

$$\nabla \cdot \mathbf{u} = q \quad \text{in } \Omega, \quad (3.1b)$$

$$p = p_D \quad \text{on } \Gamma_D, \quad \mathbf{u} \cdot \mathbf{n} = u_N \quad \text{on } \Gamma_N, \quad (3.1c)$$

where $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a polygonal domain (open, bounded, and connected set), \mathbf{S} is a bounded, symmetric (this is however not necessary), and uniformly positive definite tensor, $p_D \in H^{\frac{1}{2}}(\Gamma_D)$, $u_N \in H^{-\frac{1}{2}}(\Gamma_N)$, $q \in L_2(\Omega)$, $\Gamma_D \cap \Gamma_N = \emptyset$, $\overline{\Gamma_D} \cup \overline{\Gamma_N} = \partial\Omega$, and $|\Gamma_D| \neq 0$, where $|\Gamma_D|$ is the measure of the set Γ_D .

Let \mathcal{T}_h be a simplicial triangulation of Ω (consisting of triangles if $d = 2$ and of tetrahedra if $d = 3$) such that each boundary side (edge if $d = 2$, face if $d = 3$) lies entirely either in Γ_D or in Γ_N . Let us denote by \mathcal{E}_h the set of all non-Neumann sides of \mathcal{T}_h . Let finally $\tilde{\mathbf{u}} \in \mathbf{H}(\text{div}, \Omega)$ be such that $\tilde{\mathbf{u}} \cdot \mathbf{n} = u_N$ on Γ_N in the appropriate sense. The approximation of the problem (3.1a)–(3.1c) by means of the mixed finite element method consists in finding $\mathbf{u}_h = \mathbf{u}_{0,h} + \tilde{\mathbf{u}}$, $\mathbf{u}_{0,h} \in \mathbf{V}(\mathcal{E}_h)$, and $p_h \in \Phi(\mathcal{T}_h)$ such that (see [33, 108])

$$\begin{aligned} (\mathbf{S}^{-1}\mathbf{u}_{0,h}, \mathbf{v}_h)_\Omega - (\nabla \cdot \mathbf{v}_h, p_h)_\Omega &= -\langle \mathbf{v}_h \cdot \mathbf{n}, p_D \rangle_{\partial\Omega} \\ &\quad - (\mathbf{S}^{-1}\tilde{\mathbf{u}}, \mathbf{v}_h)_\Omega \quad \forall \mathbf{v}_h \in \mathbf{V}(\mathcal{E}_h), \end{aligned} \quad (3.2a)$$

$$-(\nabla \cdot \mathbf{u}_{0,h}, \phi_h)_\Omega = -(q, \phi_h)_\Omega + (\nabla \cdot \tilde{\mathbf{u}}, \phi_h)_\Omega \quad \forall \phi_h \in \Phi(\mathcal{T}_h), \quad (3.2b)$$

where $(\mathbf{u}_h, \mathbf{v}_h)_\Omega = \int_\Omega \mathbf{u}_h \cdot \mathbf{v}_h \, d\mathbf{x}$, $\langle \mathbf{v}_h \cdot \mathbf{n}, \varphi \rangle_{\partial\Omega} = \int_{\partial\Omega} \mathbf{v}_h \cdot \mathbf{n} \varphi \, d\gamma(\mathbf{x})$, and $\mathbf{V}(\mathcal{E}_h)$ and $\Phi(\mathcal{T}_h)$ are suitable finite-dimensional spaces defined on \mathcal{T}_h . The associated matrix problem is saddle-point and can be written in the form

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^t \\ \mathbb{B} & 0 \end{pmatrix} \begin{pmatrix} U \\ P \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}. \quad (3.3)$$

In the lowest-order Raviart–Thomas method [105] and its three-dimensional Nédélec variant [92] the scalar unknowns P are associated with the elements of \mathcal{T}_h and U are the fluxes through the sides of \mathcal{E}_h . Using the hybridization technique stemming from the ideas of [69], one can decrease the number of unknowns to the Lagrange multipliers associated with non-Dirichlet sides and obtain a symmetric and positive definite matrix, cf. [15, 33]. In fact, the hybridization is very close to the piecewise linear nonconforming finite element method, cf. [38] or Lemma 1.8.1 or a more detailed study in Section 4.6 of this thesis. The fluxes can then be recovered using the technique first proposed in [88]. Especially in three space dimensions, there are much less elements than sides, and hence the long-standing interest in reducing the unknowns to only the scalar unknowns P . This is indeed possible, using approximate numerical integration, see [110] for rectangles and \mathbf{S} diagonal and [8, 18] for rectangles and triangles and \mathbf{S} diagonal and for a limited class of tetrahedra and $\mathbf{S} = Id$. Using the expanded mixed finite element method, these techniques can be extended also onto full-matrix diffusion tensors \mathbf{S} for rectangular parallelepipeds [13] and for “smooth” coefficients and meshes consisting of triangles, quadrilaterals, and hexahedra [12]. To our knowledge, the only technique for reducing the number of unknowns to the number of elements without any numerical integration is proposed and studied in [37, 121, 122]. In two space dimensions, it works on unstructured triangular meshes, but in three space dimensions, it only works on a limited class of structured tetrahedral meshes. One associates here to each element a new unknown.

We present in Section 3.2 of this chapter a new method which permits to exactly and efficiently reduce the system (3.3) to a system for the (original) scalar unknowns P only. We show that, under a condition of the invertibility of some local matrices associated with vertices and only depending on the mesh and on the diffusion tensor, one can express the flux through a given side using the scalar unknowns, sources, and possibly boundary conditions associated with the elements sharing one of the vertices of this side. Recall that expressing the flux through a given side using the scalar unknowns in neighboring elements is the principle of multi-point finite volume schemes, cf. [1, 7, 44, 58, 65]. Hence the lowest-order Raviart–Thomas mixed finite element method is in the given case equivalent to a particular multi-point finite volume scheme, and this without any numerical integration. We call this scheme a *condensed mixed finite element scheme*. We then discuss the modifications of the proposed scheme if the local matrices are not invertible, consisting namely in considering different sets of elements for the expression of the flux through a given side.

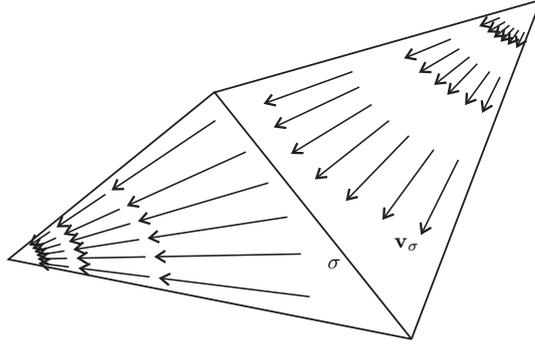
The condensation of the lowest-order Raviart–Thomas method leads to linear systems with sparse but in general nonsymmetric matrices, as we show in Section 3.3. The system matrix is positive definite under a condition on the mesh and on the tensor \mathbf{S} , which can be reduced to a shape criterion allowing for fairly general elements if \mathbf{S} is piecewise constant and scalar. For example, one can deform a square $(0, 1) \times (0, 1)$, discretized by regular right-angled triangles, until the triangle elements contain angles greater than 130 degrees, see Example 3.3.8 below. The fulfillment of this condition in particular implies the invertibility of the local matrices from the previous paragraph. Finally, in Section 3.4, we apply the proposed condensation to mixed (cf. [14, 52]) and upwind-mixed (cf. [46, 47, 77]) finite element discretizations of nonlinear parabolic convection–reaction–diffusion problems.

The essential idea of what we propose can be formulated as follows: given a second-order problem, first decompose it into scalar and flux unknowns and guarantee the fulfillment of the inf–sup (Babuška [16]–Brezzi [30]) condition. Then eliminate the added fluxes. One can in this way obtain the precision of the mixed finite element method for the price of the finite volume one. This is confirmed by numerical experiments carried out in Section 3.5. Especially for nonlinear parabolic convection–reaction–diffusion problems, one can reduce the CPU time of standard mixed solution approaches by a factor of 2 to 4. We refer to a more detailed discussion in Section 3.5.4. Finally, the proposed condensation can be easily implemented in a new self-standing code or in existing mixed finite element codes. Extension to higher-order schemes is an ongoing work.

3.2 The equivalence

We first define the spaces $\mathbf{V}(\mathcal{E}_h)$ and $\Phi(\mathcal{T}_h)$ in this section. We then establish the equivalence between the lowest-order mixed finite element and a particular multi-point finite volume method.

Let us consider simplices $K, L \in \mathcal{T}_h$ sharing an interior side σ . Let V_K be the vertex of K opposite to σ and V_L the vertex of L opposite to σ . Then the RTN (Raviart–Thomas–Nédélec) basis function $\mathbf{v}_\sigma \in \mathbf{V}(\mathcal{E}_h)$ associated with the side σ can be written in the form $\mathbf{v}_\sigma(\mathbf{x}) = \frac{1}{d|K|}(\mathbf{x} - V_K)$, $\mathbf{x} \in K$, $\mathbf{v}_\sigma(\mathbf{x}) = \frac{1}{d|L|}(V_L - \mathbf{x})$, $\mathbf{x} \in L$, $\mathbf{v}_\sigma(\mathbf{x}) = 0$ otherwise. We refer to Figure 3.1 for a schematic visualization of a RTN basis function in two space dimensions. We fix the orientation of \mathbf{v}_σ , i.e. the order of K and L . For a Dirichlet boundary side σ , the support of \mathbf{v}_σ only consists of $K \in \mathcal{T}_h$ such that $\sigma \in \mathcal{E}_K$, where \mathcal{E}_K stands for the sides of the element K . A basis function $\phi_K \in \Phi(\mathcal{T}_h)$ associated with an element $K \in \mathcal{T}_h$ is equal to 1 on K and to 0 otherwise.

Figure 3.1: RTN basis function \mathbf{v}_σ associated with the edge σ

Let us denote by \mathcal{V}_h the set of all vertices and consider $V \in \mathcal{V}_h$. We call the set of all elements of \mathcal{T}_h sharing this vertex a *cluster* associated with V and denote it by \mathcal{C}_V . Let us denote by $\mathcal{E}_{\mathcal{C}_V}$ the set of all non-Neumann sides of \mathcal{C}_V , by $\mathcal{F}_{\mathcal{C}_V}$ the set of all non-Neumann sides sharing V , and by $\mathcal{G}_{\mathcal{C}_V}$ the set of other non-Neumann sides of \mathcal{C}_V . Let finally $\mathcal{C}_V^{\text{el}}$ denote the set of elements from the cluster that contain exactly one side from $\mathcal{G}_{\mathcal{C}_V}$. We denote by δ_K the side from $\mathcal{E}_K \cap \mathcal{G}_{\mathcal{C}_V}$ for $K \in \mathcal{C}_V^{\text{el}}$. We have $\mathcal{E}_{\mathcal{C}_V} = \mathcal{F}_{\mathcal{C}_V} \cup \mathcal{G}_{\mathcal{C}_V}$, $\mathcal{F}_{\mathcal{C}_V} \cap \mathcal{G}_{\mathcal{C}_V} = \emptyset$, and $|\mathcal{C}_V^{\text{el}}| = |\mathcal{G}_{\mathcal{C}_V}|$, where we denote by $|A|$ the cardinality of a set A . An example of a cluster \mathcal{C}_V lying in the interior of the domain Ω is given in Figure 3.2. In this case, $\mathcal{F}_{\mathcal{C}_V}$ are simply the sides sharing V , $\mathcal{G}_{\mathcal{C}_V}$ the other sides of \mathcal{C}_V , and $\mathcal{C}_V^{\text{el}} = \mathcal{C}_V$. The situation is more delicate near the boundary, especially if there are Neumann boundary conditions, cf. Figure 3.3 below. This is also the reason for the quite complex notation introduced. The basic principle of the condensation will however be clear from Figure 3.2. Finally, we are not interested in the particular and trivial cases where $\mathcal{F}_{\mathcal{C}_V} = \emptyset$ or $\mathcal{G}_{\mathcal{C}_V} = \emptyset$.

Our aim is to express $\mathbf{u}_{0,h}$ with the aid of p_h , or, equivalently, the fluxes U with the aid of the scalar unknowns P . For this purpose, we consider the equations (3.2a) for the basis functions \mathbf{v}_γ , $\gamma \in \mathcal{F}_{\mathcal{C}_V}$. We remark that the support of all \mathbf{v}_γ , $\gamma \in \mathcal{F}_{\mathcal{C}_V}$, is included in \mathcal{C}_V and that $\mathbf{u}_{0,h}|_{\mathcal{C}_V} = \sum_{\sigma \in \mathcal{E}_{\mathcal{C}_V}} U_\sigma \mathbf{v}_\sigma$. This yields, using also that $p_h|_K = P_K$,

$$\sum_{\sigma \in \mathcal{E}_{\mathcal{C}_V}} U_\sigma (\mathbf{v}_\sigma, \mathbf{S}^{-1} \mathbf{v}_\gamma)_{\mathcal{C}_V} - \sum_{K \in \mathcal{C}_V} P_K (\nabla \cdot \mathbf{v}_\gamma, 1)_K = -\langle \mathbf{v}_\gamma \cdot \mathbf{n}, p_D \rangle_{\partial\Omega} - (\mathbf{S}^{-1} \tilde{\mathbf{u}}, \mathbf{v}_\gamma)_{\mathcal{C}_V} \quad \forall \gamma \in \mathcal{F}_{\mathcal{C}_V}, \quad (3.4)$$

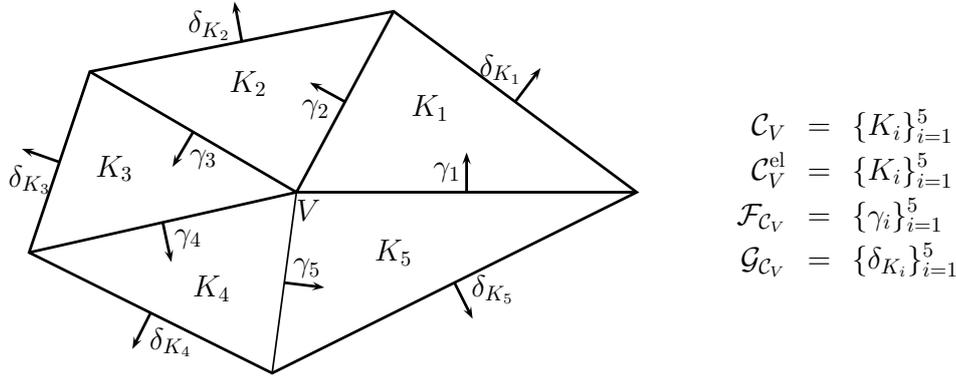
i.e. $|\mathcal{F}_{\mathcal{C}_V}|$ equations for the $|\mathcal{E}_{\mathcal{C}_V}|$ unknown fluxes U_σ , $\sigma \in \mathcal{E}_{\mathcal{C}_V}$, where we consider the scalar unknowns P_K , $K \in \mathcal{C}_V$, as parameters. Note that in practice, $p_D|_\sigma \approx \langle p_D, 1 \rangle_\sigma / |\sigma|$, $\sigma \subset \Gamma_D$, and $\tilde{\mathbf{u}} \approx \sum_{\sigma \subset \Gamma_N} \langle u_N, 1 \rangle_\sigma \mathbf{v}_\sigma$, so that the above system is completely discrete. The remaining $|\mathcal{G}_{\mathcal{C}_V}|$ equations are given by (3.2b) for all ϕ_K , $K \in \mathcal{C}_V^{\text{el}}$,

$$- \sum_{\sigma \in \mathcal{E}_K, \sigma \not\subset \Gamma_N} U_\sigma (\nabla \cdot \mathbf{v}_\sigma, 1)_K = -(q, 1)_K + (\nabla \cdot \tilde{\mathbf{u}}, 1)_K \quad \forall K \in \mathcal{C}_V^{\text{el}}. \quad (3.5)$$

The matrix problem associated with the set of equations (3.4)–(3.5) can be written in the form

$$\begin{pmatrix} \mathbb{A}_{1,V} & \mathbb{A}_{2,V} \\ \mathbb{B}_{1,V} & \mathbb{B}_{2,V} \end{pmatrix} \begin{pmatrix} U_V^{\mathcal{F}} \\ U_V^{\mathcal{G}} \end{pmatrix} = \begin{pmatrix} F_V - \mathbb{B}_V^t P_V \\ G_V \end{pmatrix}, \quad (3.6)$$

where $U_V^{\mathcal{F}} = \{U_\sigma\}_{\sigma \in \mathcal{F}_{\mathcal{C}_V}}$, $U_V^{\mathcal{G}} = \{U_\sigma\}_{\sigma \in \mathcal{G}_{\mathcal{C}_V}}$, and $P_V = \{P_K\}_{K \in \mathcal{C}_V}$.

Figure 3.2: Example of a cluster \mathcal{C}_V in the interior of Ω

We now notice that the matrix $\mathbb{B}_{2,V}$ is square, diagonal, and its entries are equal to ± 1 (this follows from the fact that each $K \in \mathcal{C}_V^{\text{el}}$ contains exactly one side from $\mathcal{G}_{\mathcal{C}_V}$ and using that $(\nabla \cdot \mathbf{v}_\sigma, 1)_K = \pm 1$ for $\sigma \in \mathcal{E}_K$). Hence we can eliminate the $U_V^{\mathcal{G}}$ unknowns and come to

$$\mathbb{M}_V U_V^{\mathcal{F}} = F_V - \mathbb{B}_V^t P_V - \mathbb{A}_{2,V} \mathbb{B}_{2,V}^{-1} G_V \quad (3.7)$$

for each vertex $V \in \mathcal{V}_h$. Let us call the matrix

$$\mathbb{M}_V := \mathbb{A}_{1,V} - \mathbb{A}_{2,V} \mathbb{B}_{2,V}^{-1} \mathbb{B}_{1,V} \quad (3.8)$$

a *local condensation matrix* associated with the vertex V . We now summarize the results in the following theorem:

Theorem 3.2.1. (Equivalence between MFEM and a particular multi-point FVM)

Let the matrices \mathbb{M}_V given by (3.8) be invertible for all $V \in \mathcal{V}_h$. Then the lowest-order Raviart–Thomas mixed finite element method on simplicial meshes is equivalent to a multi-point finite volume scheme, where the flux through each side can be expressed using the scalar unknowns, sources, and possibly boundary conditions associated with the elements sharing one of the vertices of this side.

Remark 3.2.2. (Comparison with a classical multi-point FVM)

In “classical” multi-point finite volume schemes, cf. [1, 7, 44, 58, 65], one attempts to express the flux through a given side only using the scalar unknowns associated with the neighboring elements. There are two essential differences between these classical multi-point finite volume schemes and a particular multi-point finite volume scheme—the mixed finite element method. First, in the mixed finite element method, not only the scalar unknowns, but also the sources and possibly boundary conditions associated with the neighboring elements are used to express the flux through a given side. Second, to obtain this expression, one has to solve a local linear problem. In this last feature, the condensed mixed finite element scheme is similar to the “multipoint flux-approximation” scheme proposed and tested in [1, 2].

Let $V \in \mathcal{V}_h$. Let us define a mapping $\Psi_V : \mathbb{R}^{|\mathcal{F}_{\mathcal{C}_V}|} \rightarrow \mathbb{R}^{|\mathcal{E}_h|}$, extending a vector $U_V^{\mathcal{F}} = \{U_\sigma\}_{\sigma \in \mathcal{F}_{\mathcal{C}_V}}$ of values associated with the sides from $\mathcal{F}_{\mathcal{C}_V}$ to a vector of values associated with all non-Neumann sides \mathcal{E}_h by

$$[\Psi_V(U_V^{\mathcal{F}})]_\sigma := \begin{cases} U_\sigma & \text{if } \sigma \in \mathcal{F}_{\mathcal{C}_V} \\ 0 & \text{if } \sigma \notin \mathcal{F}_{\mathcal{C}_V} \end{cases} .$$

Since there is no possibility of confusion, we keep the same notation also for a mapping $\mathbb{R}^{|\mathcal{F}_{C_V}| \times |\mathcal{F}_{C_V}|} \rightarrow \mathbb{R}^{|\mathcal{E}_h| \times |\mathcal{E}_h|}$, extending a local matrix \mathbb{M}_V to a full-size one by zeros by

$$[\Psi_V(\mathbb{M}_V)]_{\sigma,\gamma} := \begin{cases} (\mathbb{M}_V)_{\sigma,\gamma} & \text{if } \sigma \in \mathcal{F}_{C_V} \text{ and } \gamma \in \mathcal{F}_{C_V} \\ 0 & \text{if } \sigma \notin \mathcal{F}_{C_V} \text{ or } \gamma \notin \mathcal{F}_{C_V} \end{cases}.$$

We finally in the same fashion define a mapping $\Theta_V : \mathbb{R}^{|\mathcal{F}_{C_V}| \times |\mathcal{C}_V^{\text{el}}|} \rightarrow \mathbb{R}^{|\mathcal{E}_h| \times |\mathcal{T}_h|}$, filling a full-size representation of a matrix \mathbb{J}_V by zeros on the rows associated with the sides that are not from \mathcal{F}_{C_V} and on the columns associated with the elements that are not from $\mathcal{C}_V^{\text{el}}$,

$$[\Theta_V(\mathbb{J}_V)]_{\sigma,K} := \begin{cases} (\mathbb{J}_V)_{\sigma,K} & \text{if } \sigma \in \mathcal{F}_{C_V} \text{ and } K \in \mathcal{C}_V^{\text{el}} \\ 0 & \text{if } \sigma \notin \mathcal{F}_{C_V} \text{ or } K \notin \mathcal{C}_V^{\text{el}} \end{cases}.$$

Let the local condensation matrices \mathbb{M}_V be invertible for all $V \in \mathcal{V}_h$. Let us define \mathbb{J}_V by $\mathbb{J}_V := \mathbb{M}_V^{-1} \mathbb{A}_{2,V} \mathbb{B}_{2,V}^{-1}$. We then can rewrite (3.7) as

$$\Psi_V(U_V^{\mathcal{F}}) = \Psi_V(\mathbb{M}_V^{-1})(F - \mathbb{B}^t P) - \Theta_V(\mathbb{J}_V)G. \quad (3.9)$$

We now notice that

$$\sum_{V \in \mathcal{V}_h} \frac{1}{d} \Psi_V(U_V^{\mathcal{F}}) = U, \quad (3.10)$$

which expresses that if we go through all $V \in \mathcal{V}_h$ and observe the sides in the sets \mathcal{F}_{C_V} , each $\sigma \in \mathcal{E}_h$ appears just d -times (each side has d vertices). Hence we can sum (3.9) over all vertices and divide it by d to find that

$$U = \tilde{\mathbb{A}}^{-1}(F - \mathbb{B}^t P) - \mathbb{J}G, \quad (3.11)$$

where

$$\tilde{\mathbb{A}}^{-1} := \frac{1}{d} \sum_{V \in \mathcal{V}_h} \Psi_V(\mathbb{M}_V^{-1}), \quad \mathbb{J} := \frac{1}{d} \sum_{V \in \mathcal{V}_h} \Theta_V(\mathbb{J}_V). \quad (3.12)$$

Finally, inserting this expression into the second equation of (3.3), we obtain a system for only the scalar unknowns

$$-\mathbb{B} \tilde{\mathbb{A}}^{-1} \mathbb{B}^t P = G - \mathbb{B} \tilde{\mathbb{A}}^{-1} F + \mathbb{B} \mathbb{J} G. \quad (3.13)$$

We now give two remarks.

Remark 3.2.3. (Comparison with the direct elimination of the fluxes) From (3.3), $U = \mathbb{A}^{-1}(F - \mathbb{B}^t P)$. There are two essential differences in comparison with (3.11). First, the matrix $\tilde{\mathbb{A}}^{-1}$ is sparse, whereas \mathbb{A}^{-1} tends to be full. Second, $\tilde{\mathbb{A}}^{-1}$ is obtained for the price of the inverse of $|\mathcal{V}_h|$ local matrices, whereas obtaining \mathbb{A}^{-1} is in general very expensive.

Remark 3.2.4. (Implementation into existing mixed finite element codes) The local problems (3.6) correspond to the rows of (3.3) associated with the sides from \mathcal{F}_{C_V} and elements from $\mathcal{C}_V^{\text{el}}$. Hence obtaining the final problem (3.13) from (3.3) is immediate.

It appears that in some particular cases, the matrix \mathbb{M}_V is not invertible, cf. Example 3.3.10 below. We give sufficient conditions on the mesh \mathcal{T}_h and on the diffusion tensor \mathbf{S} ensuring that \mathbb{M}_V are invertible for all $V \in \mathcal{V}_h$ below as byproducts of Lemmas 3.3.6 and 3.3.9. Finally, we discuss in Section 3.3.3 the approaches how to modify the proposed technique in order to overcome this difficulty.

3.3 Properties of the condensed mixed finite element scheme

We study in this section the properties of the system matrix of the condensed mixed finite element scheme important from the computational point of view, namely its sparsity pattern, symmetry, and positive definiteness. It shows that all these properties are closely related to the properties of the local condensation matrices, which we shall study hereafter. We finally discuss variants and extensions of the proposed technique and open questions.

3.3.1 Properties of the system matrix

Theorem 3.3.1. (Stencil of the system matrix) *Let \mathbb{M}_V be invertible for all $V \in \mathcal{V}_h$. Then on a row of the final system matrix $\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t$ corresponding to an element $K \in \mathcal{T}_h$, the only possible nonzero entries are on columns corresponding to $L \in \mathcal{T}_h$ such that K and L share a common vertex.*

PROOF:

The assertion of this theorem follows from the fact that by (3.7), the flux through a side σ is expressed only using the scalar unknowns of the elements $K \in \mathcal{T}_h$ such that K and σ share a common vertex. \square

Theorem 3.3.2. (Positive definiteness of the system matrix) *Let \mathbb{M}_V be positive definite for all $V \in \mathcal{V}_h$. Then the final system matrix $\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t$ is also positive definite.*

PROOF:

Since \mathbb{B} has a full row rank, $\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t$ is positive definite as soon as $\tilde{\mathbb{A}}^{-1}$ is positive definite, i.e. when

$$X^t \tilde{\mathbb{A}}^{-1} X > 0 \quad \text{for all } X \in \mathbb{R}^{|\mathcal{T}_h|}, X \neq 0.$$

Let $V \in \mathcal{V}_h$. We define a mapping $\Pi_V : \mathbb{R}^{|\mathcal{E}_h|} \rightarrow \mathbb{R}^{|\mathcal{F}_{C_V}|}$, restricting a vector of values associated with all non-Neumann sides to a vector of values associated with the sides from \mathcal{F}_{C_V} . Let $X \in \mathbb{R}^{|\mathcal{E}_h|}$, $X \neq 0$. Then

$$X^t \tilde{\mathbb{A}}^{-1} X = \frac{1}{d} \sum_{V \in \mathcal{V}_h} X^t \Psi_V(\mathbb{M}_V^{-1}) X = \frac{1}{d} \sum_{V \in \mathcal{V}_h} [\Pi_V(X)]^t \mathbb{M}_V^{-1} \Pi_V(X) > 0,$$

using the positive definiteness of the local condensation matrices \mathbb{M}_V and consequently of \mathbb{M}_V^{-1} for all $V \in \mathcal{V}_h$ and the fact that in the above sum, all the terms are non-negative and at least d of them are positive. \square

Theorem 3.3.3. (Symmetry of the system matrix) *Let \mathbb{M}_V be invertible and symmetric for all $V \in \mathcal{V}_h$. Then the final system matrix $\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t$ is also symmetric.*

PROOF:

If \mathbb{M}_V and consequently \mathbb{M}_V^{-1} are symmetric for all $V \in \mathcal{V}_h$, their extensions $\Psi_V(\mathbb{M}_V^{-1})$ are symmetric as well. Hence $\tilde{\mathbb{A}}^{-1}$, a sum of symmetric matrices by (3.12), is symmetric. Finally, if $\tilde{\mathbb{A}}^{-1}$ is symmetric, $\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t$ is symmetric as well. \square

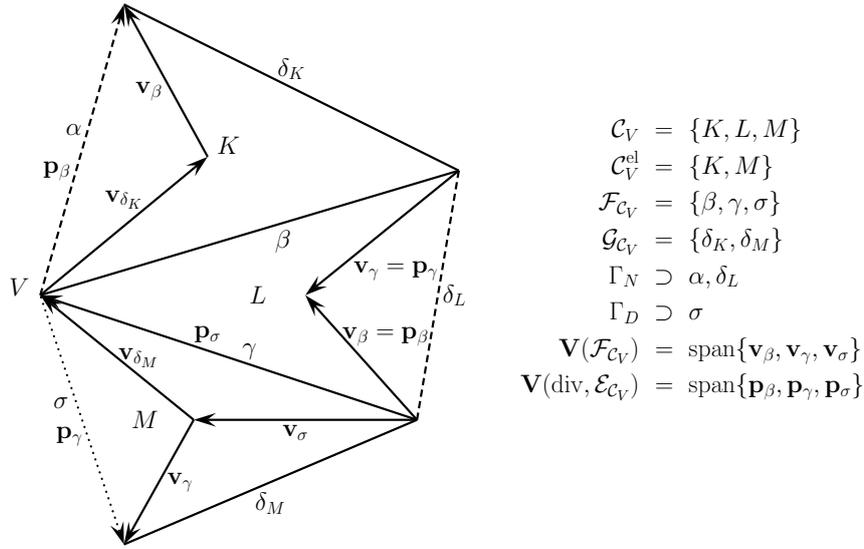


Figure 3.3: Example of a boundary cluster \mathcal{C}_V and schematic representation of the basis functions of the spaces $\mathbf{V}(\mathcal{F}_{\mathcal{C}_V})$ and $\mathbf{V}(\text{div}, \mathcal{E}_{\mathcal{C}_V})$

3.3.2 Properties of the local condensation matrices

The local condensation matrix \mathbb{M}_V (3.8) for $V \in \mathcal{V}_h$ stems from the equations (3.4)–(3.5). It does not depend on the right-hand side, and hence it is connected with the following problem: find $\mathbf{u} \in \mathbf{V}(\mathcal{E}_{\mathcal{C}_V})$ such that

$$(\mathbf{u}, \mathbf{S}^{-1}\mathbf{v})_{\mathcal{C}_V} = 0 \quad \forall \mathbf{v} \in \mathbf{V}(\mathcal{F}_{\mathcal{C}_V}), \quad (3.14a)$$

$$(\nabla \cdot \mathbf{u}, \phi_K)_K = 0 \quad \forall K \in \mathcal{C}_V^{\text{el}}. \quad (3.14b)$$

Here, $\mathbf{V}(\mathcal{E}_{\mathcal{C}_V})$ is the space spanned by the RTN basis functions \mathbf{v}_σ associated with the non-Neumann sides $\mathcal{E}_{\mathcal{C}_V}$ of the cluster \mathcal{C}_V and $\mathbf{V}(\mathcal{F}_{\mathcal{C}_V})$ is its restriction with the basis functions \mathbf{v}_σ associated with the sides from $\mathcal{F}_{\mathcal{C}_V}$. The problem (3.14a)–(3.14b) is further equivalent to the following Petrov–Galerkin problem: find $\mathbf{u} \in \mathbf{V}(\text{div}, \mathcal{E}_{\mathcal{C}_V})$ such that

$$(\mathbf{u}, \mathbf{S}^{-1}\mathbf{v})_{\mathcal{C}_V} = 0 \quad \forall \mathbf{v} \in \mathbf{V}(\mathcal{F}_{\mathcal{C}_V}),$$

where $\mathbf{V}(\text{div}, \mathcal{E}_{\mathcal{C}_V})$ is the subspace of $\mathbf{V}(\mathcal{E}_{\mathcal{C}_V})$ of the functions whose divergence is equal to 0 on all elements $K \in \mathcal{C}_V^{\text{el}}$. The space $\mathbf{V}(\text{div}, \mathcal{E}_{\mathcal{C}_V})$ is spanned by basis functions \mathbf{p}_σ associated with the sides from $\mathcal{F}_{\mathcal{C}_V}$, which have the same support as the RTN basis functions \mathbf{v}_σ and whose fluxes through the associated sides equal to those of \mathbf{v}_σ (this namely fixes their orientation). In particular, for $K \in \mathcal{C}_V^{\text{el}}$ and $\sigma \in \mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}_V}$, $\mathbf{p}_\sigma|_K = \mathbf{v}_\sigma - \frac{(\nabla \cdot \mathbf{v}_\sigma, 1)_K}{(\nabla \cdot \mathbf{v}_{\delta_K}, 1)_K} \mathbf{v}_{\delta_K}$. Note that this is a constant function given by $\frac{1}{d|K|} \mathbf{q}_\sigma|_K$, where $\mathbf{q}_\sigma|_K$ is the vector of the edge of K that is not included in the sides σ and δ_K . For $K \in \mathcal{C}_V \setminus \mathcal{C}_V^{\text{el}}$, $\mathbf{p}_\sigma|_K = \mathbf{v}_\sigma|_K$. We refer to Figure 3.3 for a schematic visualization for $d = 2$.

Lemma 3.3.4. (Form of the local condensation matrices) *The local condensation matrix \mathbb{M}_V for $V \in \mathcal{V}_h$ is given by*

$$(\mathbb{M}_V)_{\gamma, \sigma} = (\mathbf{p}_\sigma, \mathbf{S}^{-1}\mathbf{v}_\gamma)_{\mathcal{C}_V},$$

where \mathbf{p}_σ and \mathbf{v}_σ , $\sigma \in \mathcal{F}_{\mathcal{C}_V}$, are the basis functions of the spaces $\mathbf{V}(\text{div}, \mathcal{E}_{\mathcal{C}_V})$ and $\mathbf{V}(\mathcal{F}_{\mathcal{C}_V})$, respectively, defined above.

PROOF:

We can rewrite (3.14a)–(3.14b) as

$$\sum_{\sigma \in \mathcal{E}_{\mathcal{C}_V}} U_\sigma(\mathbf{v}_\sigma, \mathbf{S}^{-1}\mathbf{v}_\gamma)_{\mathcal{C}_V} = 0 \quad \forall \gamma \in \mathcal{F}_{\mathcal{C}_V}, \quad (3.15a)$$

$$\sum_{\sigma \in \mathcal{E}_K, \sigma \notin \Gamma_N} U_\sigma(\nabla \cdot \mathbf{v}_\sigma, 1)_K = 0 \quad \forall K \in \mathcal{C}_V^{\text{el}}, \quad (3.15b)$$

where $\mathbf{u} = \sum_{\sigma \in \mathcal{E}_{\mathcal{C}_V}} U_\sigma \mathbf{v}_\sigma$. Expressing U_{δ_K} from (3.15b) gives

$$U_{\delta_K} = \frac{- \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}_V}} U_\sigma(\nabla \cdot \mathbf{v}_\sigma, 1)_K}{(\nabla \cdot \mathbf{v}_{\delta_K}, 1)_K}.$$

Inserting this into (3.15a), we have

$$\begin{aligned} \sum_{\sigma \in \mathcal{E}_{\mathcal{C}_V}} U_\sigma(\mathbf{v}_\sigma, \mathbf{S}^{-1}\mathbf{v}_\gamma)_{\mathcal{C}_V} &= \sum_{K \in \text{supp}(\mathbf{v}_\gamma)} \left\{ \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}_V}} U_\sigma(\mathbf{v}_\sigma, \mathbf{S}^{-1}\mathbf{v}_\gamma)_K + U_{\delta_K}(\mathbf{v}_{\delta_K}, \mathbf{S}^{-1}\mathbf{v}_\gamma)_K \right\} \\ &= \sum_{K \in \text{supp}(\mathbf{v}_\gamma)} \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}_V}} U_\sigma \left(\mathbf{v}_\sigma - \frac{(\nabla \cdot \mathbf{v}_\sigma, 1)_K}{(\nabla \cdot \mathbf{v}_{\delta_K}, 1)_K} \mathbf{v}_{\delta_K}, \mathbf{S}^{-1}\mathbf{v}_\gamma \right)_K, \end{aligned}$$

where we have defined for simplification $\mathbf{v}_{\delta_K} = 0$ if $K \in \mathcal{C}_V \setminus \mathcal{C}_V^{\text{el}}$ (i.e. if $\mathcal{E}_K \cap \mathcal{G}_{\mathcal{C}_V} = \emptyset$). Hence, using the definition of the basis functions of the spaces $\mathbf{V}(\text{div}, \mathcal{E}_{\mathcal{C}_V})$ and $\mathbf{V}(\mathcal{F}_{\mathcal{C}_V})$, the assertion of the lemma follows. \square

Remark 3.3.5. (Implementation) Let \mathbf{S} be piecewise constant and let $\Gamma_N = \emptyset$. Then

$$\begin{aligned} (\mathbb{M}_V)_{\gamma, \sigma} &= \sum_{K \in \mathcal{C}_V; \sigma, \gamma \in \mathcal{E}_K} (\mathbf{p}_\sigma, \mathbf{S}^{-1}\mathbf{v}_\gamma)_K = \sum_{K \in \mathcal{C}_V; \sigma, \gamma \in \mathcal{E}_K} \frac{(\nabla \cdot \mathbf{v}_\gamma, 1)_K}{d^2 |K|^2} (\mathbf{S}^{-1}\mathbf{q}_\sigma, \mathbf{x} - V_{\gamma, K})_K \\ &= \sum_{K \in \mathcal{C}_V; \sigma, \gamma \in \mathcal{E}_K} \frac{1}{d^2 |K|} \mathbf{S}|_K^{-1} \mathbf{q}_\sigma|_K \cdot \mathbf{w}_\gamma|_K, \end{aligned}$$

where $\sigma, \gamma \in \mathcal{F}_{\mathcal{C}_V}$ and $\mathbf{w}_\gamma|_K := (\nabla \cdot \mathbf{v}_\gamma, 1)_K (\mathbf{x}_K - V_{\gamma, K})$ with \mathbf{x}_K the barycentre of K and $V_{\gamma, K}$ the vertex of K opposite to the side γ , cf. Figure 3.4. We have used the facts that $\{K \in \mathcal{C}_V; \sigma, \gamma \in \mathcal{E}_K\} = \text{supp}(\mathbf{p}_\sigma) \cap \text{supp}(\mathbf{v}_\gamma)$ and that $\mathbf{x}_K = (\mathbf{x}, 1)_K / |K|$. Hence, to implement the condensed mixed finite element scheme when in addition $q = 0$, everything we need are the edge and vertex–barycentre vectors in each simplex and the measure of each simplex.

We now give two lemmas that guarantee the positive definiteness of the local condensation matrices, the assumption of Theorem 3.3.2. Since positive definiteness implies invertibility, the local condensation matrices are under the following conditions namely invertible, which guarantees the feasibility of the condensation in the proposed form. The given conditions are sufficient but not necessary to ensure the positive definiteness—they can be used as a simple elementwise or sidewise criterion, in order to avoid the direct checking of the positive definiteness of the local condensation matrices.

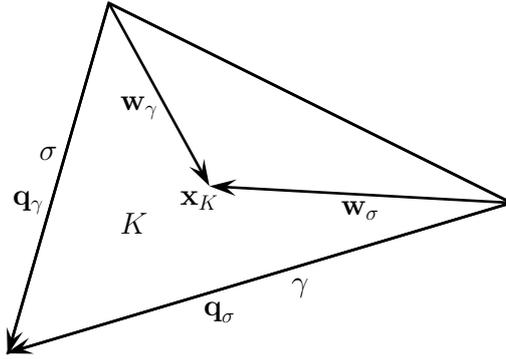


Figure 3.4: Triangle K for the simplified elementwise positive definiteness criterion

Lemma 3.3.6. (Positive definiteness of the local condensation matrices—elementwise criterion) *Let the matrices $\mathbb{E}_{V,K} \in \mathbb{R}^{|\mathcal{E}_K \cap \mathcal{F}_{C_V}| \times |\mathcal{E}_K \cap \mathcal{F}_{C_V}|}$ given by*

$$(\mathbb{E}_{V,K})_{\gamma,\sigma} := (\mathbf{p}_\sigma, \mathbf{S}^{-1} \mathbf{v}_\gamma)_K,$$

where \mathbf{p}_σ and \mathbf{v}_σ , $\sigma \in \mathcal{E}_K \cap \mathcal{F}_{C_V}$, are the basis functions of the spaces $\mathbf{V}(\text{div}, \mathcal{E}_{C_V})$ and $\mathbf{V}(\mathcal{F}_{C_V})$, respectively, be positive definite for all $K \in \mathcal{T}_h$ and for all vertices V of K . Then the local condensation matrices \mathbb{M}_V are positive definite for all $V \in \mathcal{V}_h$.

PROOF:

Let $V \in \mathcal{V}_h$ and let $X \in \mathbb{R}^{|\mathcal{F}_{C_V}|}$, $X \neq 0$. We then have, with $\mathbf{p} = \sum_{\sigma \in \mathcal{F}_{C_V}} X_\sigma \mathbf{p}_\sigma$, $\mathbf{v} = \sum_{\sigma \in \mathcal{F}_{C_V}} X_\sigma \mathbf{v}_\sigma$,

$$\begin{aligned} X^t \mathbb{M}_V X &= (\mathbf{p}, \mathbf{S}^{-1} \mathbf{v})_{C_V} = \sum_{K \in \mathcal{C}_V^{\text{el}}} (\mathbf{p}, \mathbf{S}^{-1} \mathbf{v})_K + \sum_{K \in \mathcal{C}_V \setminus \mathcal{C}_V^{\text{el}}} (\mathbf{p}, \mathbf{S}^{-1} \mathbf{v})_K \\ &= \sum_{K \in \mathcal{C}_V^{\text{el}}} [\Pi_{V,K}(X)]^t \mathbb{E}_{V,K} \Pi_{V,K}(X) + \sum_{K \in \mathcal{C}_V \setminus \mathcal{C}_V^{\text{el}}} (\mathbf{v}, \mathbf{S}^{-1} \mathbf{v})_K > 0, \end{aligned}$$

where the mapping $\Pi_{V,K} : \mathbb{R}^{|\mathcal{F}_{C_V}|} \rightarrow \mathbb{R}^{|\mathcal{E}_K \cap \mathcal{F}_{C_V}|}$ restricts a vector of values associated with the sides from \mathcal{F}_{C_V} to a vector of values associated with the sides from $\mathcal{E}_K \cap \mathcal{F}_{C_V}$, and using that the two last terms are non-negative and at least one of them is positive. \square

Remark 3.3.7. (Simplified elementwise positive definiteness criterion in two space dimensions) *Let $d = 2$ and let \mathbf{S} be piecewise constant. Let $\mathbf{q}_\sigma, \mathbf{q}_\gamma, \mathbf{w}_\sigma, \mathbf{w}_\gamma$ be the constant edge and vertex-barycentre vectors of a triangle K as in Figure 3.4. Then a simplified criterion for the positive definiteness of the local condensation matrices is*

$$\left| \mathbf{S}|_K^{-1} \mathbf{q}_\sigma \cdot \mathbf{w}_\gamma + \mathbf{S}|_K^{-1} \mathbf{q}_\gamma \cdot \mathbf{w}_\sigma \right|^2 < 4(\mathbf{S}|_K^{-1} \mathbf{q}_\sigma \cdot \mathbf{w}_\sigma)(\mathbf{S}|_K^{-1} \mathbf{q}_\gamma \cdot \mathbf{w}_\gamma)$$

for all $K \in \mathcal{T}_h$ and for all denotation σ, γ of two edges of K . Notice that $\mathbf{q}_\sigma \cdot \mathbf{w}_\gamma = 0$ for an equilateral triangle and that this quantity grows in the absolute value while deforming the triangle. On the contrary, $\mathbf{q}_\sigma \cdot \mathbf{w}_\sigma$ decreases with the angle between \mathbf{q}_σ and \mathbf{w}_σ and it is positive only if this angle is less than $\pi/2$. This criterion is a consequence of Remark 3.3.5 and of Lemma 3.3.6 with a tightened up criterion for triangles with Neumann edges.

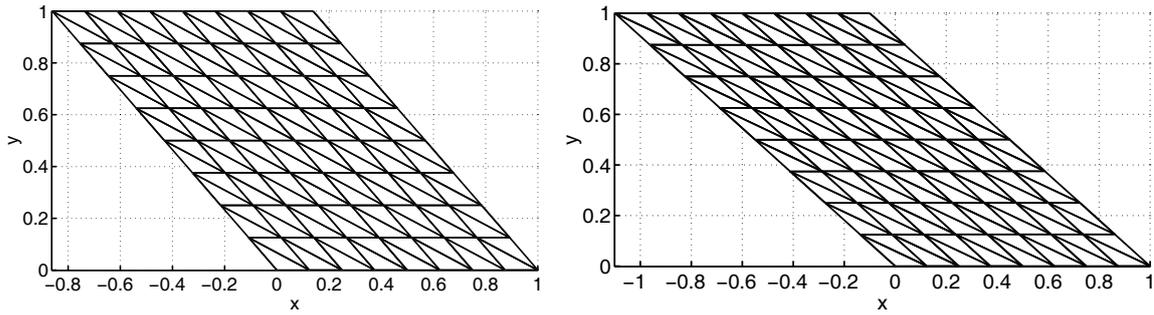


Figure 3.5: Theoretical (left) and experimental (right) limit mesh for the positive definiteness of the system matrix for a deformed square and $\mathbf{S} = Id$

Example 3.3.8. (Positive definiteness for a triangulation of a deformed square) Let $\mathbf{S} = Id$, let Ω be a square $(0, 1) \times (0, 1)$, and let \mathcal{T}_h be its triangulation by regular right-angled triangles. Let us deform the domain and the mesh by shifting horizontally the upper edge of the square. Remark 3.3.7 gives that the local condensation matrices (and consequently the system matrix) are positive definite up to the mesh given in Figure 3.5 on the left-hand side. The experimental limit mesh is still a bit less restrictive and is given in Figure 3.5 on the right-hand side.

Lemma 3.3.9. (Positive definiteness of the local condensation matrices—sidewise criterion) Let \mathbf{S} be piecewise constant and let $\Gamma_N = \emptyset$. Let for all $\gamma \in \mathcal{E}_h$ and for all vertices V of γ ,

$$\sum_{K \in \text{supp}(\mathbf{v}_\gamma)} \frac{1}{d^2|K|} \mathbf{S}|_K^{-1} \mathbf{q}_\gamma|_K \cdot \mathbf{w}_\gamma|_K > \sum_{K \in \text{supp}(\mathbf{v}_\gamma)} \frac{1}{d^2|K|} \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{F}_{C_V}, \sigma \neq \gamma} \left| \frac{1}{2} \mathbf{S}|_K^{-1} (\mathbf{q}_\sigma|_K \cdot \mathbf{w}_\gamma|_K + \mathbf{q}_\gamma|_K \cdot \mathbf{w}_\sigma|_K) \right|,$$

where the constant edge and vertex–barycentre vectors $\mathbf{q}_\sigma|_K$, $\mathbf{w}_\sigma|_K$, respectively, are derived from the basis functions of the spaces $\mathbf{V}(\text{div}, \mathcal{E}_{C_V})$ and $\mathbf{V}(\mathcal{F}_{C_V})$ \mathbf{p}_σ and \mathbf{v}_σ as in Remark 3.3.5. Then the local condensation matrices \mathbb{M}_V are positive definite for all $V \in \mathcal{V}_h$.

PROOF:

The assumption of the lemma ensures that the matrices $\frac{1}{2}(\mathbb{M}_V + \mathbb{M}_V^t)$ for all $V \in \mathcal{V}_h$ have positive diagonal entries and are strictly diagonally dominant and hence they are positive definite. To conclude, it suffices to note that the matrix \mathbb{M}_V is positive definite if and only if its symmetric part $\frac{1}{2}(\mathbb{M}_V + \mathbb{M}_V^t)$ is positive definite. \square

Example 3.3.10. (Singular local condensation matrix) We give in Figure 3.6 an example of a mesh where the local condensation matrix \mathbb{M}_V is singular for $\mathbf{S} = Id$. All the triangles sharing the vertex V have exactly one edge σ such that $\mathbf{q}_\sigma \cdot \mathbf{w}_\sigma = 0$ with the notation of Figure 3.4. Hence, in particular, the assumptions of Lemma 3.3.6 are not verified. This singularity is not local—it suffices to modify the coordinates of one point to make \mathbb{M}_V invertible.

We now state under which conditions the assumption of Theorem 3.3.3 is satisfied.

Lemma 3.3.11. (Symmetry of the local condensation matrices) Let \mathcal{T}_h consist of equilateral simplices and let \mathbf{S} be a piecewise constant scalar function. Then \mathbb{M}_V are symmetric for all $V \in \mathcal{V}_h$.

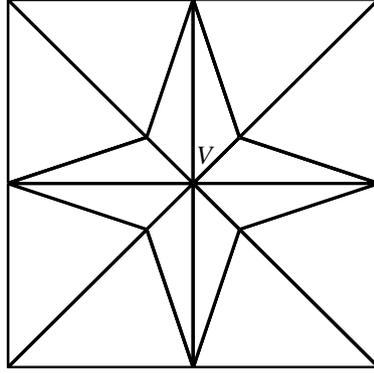


Figure 3.6: A mesh where the local condensation matrix \mathbb{M}_V is singular

PROOF:

We have

$$(\mathbb{M}_V)_{\gamma,\sigma} = \left(\mathbf{v}_\sigma - \frac{(\nabla \cdot \mathbf{v}_\sigma, 1)_K}{(\nabla \cdot \mathbf{v}_{\delta_K}, 1)_K} \mathbf{v}_{\delta_K}, \mathbf{S}^{-1} \mathbf{v}_\gamma \right)_K,$$

where $K \in \text{supp}(\mathbf{p}_\sigma) \cap \text{supp}(\mathbf{v}_\gamma)$, $\sigma, \gamma \in \mathcal{F}_{\mathcal{C}_V}$, $\sigma \neq \gamma$. If $K \in \mathcal{C}_V \setminus \mathcal{C}_V^{\text{el}}$ and thus $\mathbf{v}_{\delta_K} = 0$ by the definition, $(\mathbb{M}_V)_{\gamma,\sigma}$ clearly equals to $(\mathbb{M}_V)_{\sigma,\gamma}$ for a general \mathbf{S} by its symmetry. If $K \in \mathcal{C}_V^{\text{el}}$, $(\mathbb{M}_V)_{\gamma,\sigma} = (\mathbb{M}_V)_{\sigma,\gamma}$ as soon as

$$(\mathbf{S}^{-1} \mathbf{v}_{\delta_K}, \mathbf{v}_\gamma (\nabla \cdot \mathbf{v}_\gamma, 1)_K)_K = (\mathbf{S}^{-1} \mathbf{v}_{\delta_K}, \mathbf{v}_\sigma (\nabla \cdot \mathbf{v}_\sigma, 1)_K)_K,$$

which is the case of an equilateral simplex and \mathbf{S} a piecewise constant scalar function. \square

Remark 3.3.12. (Equilateral simplices and a piecewise constant scalar diffusion tensor) Let \mathcal{T}_h consist of equilateral simplices, let \mathbf{S} be a piecewise constant scalar function, and let $\Gamma_N = \emptyset$. Then it follows from Remark 3.3.5 that $(\mathbb{M}_V)_{\gamma,\sigma} = 0$ if $\sigma \neq \gamma$ (since the vectors $\mathbf{q}_\sigma|_K$ and $\mathbf{w}_\gamma|_K$ are orthogonal), and hence the local condensation matrices are diagonal. Thus to express the flux through an interior side γ in this case, we only need the scalar unknowns associated with the two elements that share this side. As a consequence, the final system matrix has only a 4-point stencil in two space dimensions and a 5-point stencil in three space dimensions and is moreover symmetric and positive definite. A simple computation gives that this matrix is equivalent to that of the standard finite volume scheme [61] when $\mathbf{S} = \text{Id}$. Note however that the right-hand side is generally different in a presence of a source term!

3.3.3 Variants, extensions, and open problems

The essential idea of the proposed elimination, briefly said, consists in considering such sets of elements that it is possible to eliminate the fluxes through the exterior sides of these sets by the divergence equations on the exterior elements. The clusters defined by all elements sharing a given vertex represent just the most basic possibility. We now precise on this point.

Let \mathcal{C} be a set of elements of \mathcal{T}_h and let $\mathcal{G}_\mathcal{C}$ be the set of sides of \mathcal{C} between an element $K \in \mathcal{C}$ and $L \notin \mathcal{C}$. Let each $K \in \mathcal{C}$ contain at most one $\sigma \in \mathcal{G}_\mathcal{C}$ and let us denote the subset of \mathcal{C} of elements containing a $\sigma \in \mathcal{G}_\mathcal{C}$ by \mathcal{C}^{el} . Clearly, $|\mathcal{C}^{\text{el}}| = |\mathcal{G}_\mathcal{C}|$, and we denote by δ_K the side of $K \in \mathcal{C}^{\text{el}}$ such that $\delta_K \in \mathcal{G}_\mathcal{C}$. Finally, let $\mathcal{E}_\mathcal{C}$ stand for all non-Neumann sides of \mathcal{C} and $\mathcal{F}_\mathcal{C}$ for $\mathcal{E}_\mathcal{C} \setminus \mathcal{G}_\mathcal{C}$. A particular example is the cluster \mathcal{C}_V associated with a vertex V . We have the spaces $\mathbf{V}(\mathcal{F}_{\mathcal{C}_V})$ and $\mathbf{V}(\text{div}, \mathcal{E}_{\mathcal{C}_V})$ as in Section 3.3.2 and the following generalization of Lemma 3.3.6:

Lemma 3.3.13. (Positive definiteness of local condensation matrices on general clusters) *Let the matrices $\mathbb{E}_{\mathcal{C},K} \in \mathbb{R}^{|\mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}}| \times |\mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}}|}$ given by*

$$(\mathbb{E}_{\mathcal{C},K})_{\gamma,\sigma} := (\mathbf{p}_{\sigma}, \mathbf{S}^{-1} \mathbf{v}_{\gamma})_K,$$

where \mathbf{p}_{σ} and \mathbf{v}_{σ} , $\sigma \in \mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}}$, are the basis functions of the spaces $\mathbf{V}(\text{div}, \mathcal{E}_{\mathcal{C}})$ and $\mathbf{V}(\mathcal{F}_{\mathcal{C}})$, respectively, be positive definite for all $K \in \mathcal{C}^{\text{el}}$. Then the local condensation matrix $\mathbb{M}_{\mathcal{C}}$ associated with the cluster \mathcal{C} , $(\mathbb{M}_{\mathcal{C}})_{\gamma,\sigma} = (\mathbf{p}_{\sigma}, \mathbf{S}^{-1} \mathbf{v}_{\gamma})_{\mathcal{C}}$, is positive definite.

PROOF:

Let $X \in \mathbb{R}^{|\mathcal{F}_{\mathcal{C}}|}$, $X \neq 0$. We then have, with $\mathbf{p} = \sum_{\sigma \in \mathcal{F}_{\mathcal{C}}} X_{\sigma} \mathbf{p}_{\sigma}$, $\mathbf{v} = \sum_{\sigma \in \mathcal{F}_{\mathcal{C}}} X_{\sigma} \mathbf{v}_{\sigma}$,

$$\begin{aligned} X^t \mathbb{M}_{\mathcal{C}} X &= (\mathbf{p}, \mathbf{S}^{-1} \mathbf{v})_{\mathcal{C}} = \sum_{K \in \mathcal{C}^{\text{el}}} (\mathbf{p}, \mathbf{S}^{-1} \mathbf{v})_K + \sum_{K \in \mathcal{C} \setminus \mathcal{C}^{\text{el}}} (\mathbf{p}, \mathbf{S}^{-1} \mathbf{v})_K \\ &= \sum_{K \in \mathcal{C}^{\text{el}}} [\Pi_{\mathcal{C},K}(X)]^t \mathbb{E}_{\mathcal{C},K} \Pi_{\mathcal{C},K}(X) + \sum_{K \in \mathcal{C} \setminus \mathcal{C}^{\text{el}}} (\mathbf{v}, \mathbf{S}^{-1} \mathbf{v})_K > 0, \end{aligned}$$

where the mapping $\Pi_{\mathcal{C},K} : \mathbb{R}^{|\mathcal{F}_{\mathcal{C}}|} \rightarrow \mathbb{R}^{|\mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}}|}$ restricts a vector of values associated with the sides from $\mathcal{F}_{\mathcal{C}}$ to a vector of values associated with the sides from $\mathcal{E}_K \cap \mathcal{F}_{\mathcal{C}}$, and using that the two last terms are non-negative and at least one of them is positive. \square

The above lemma shows that the positive definiteness of local condensation matrices only depends on the elements from \mathcal{C}^{el} . Hence, in particular, shall the local condensation matrix associated with a cluster of a vertex V be singular, we can resort to a wider cluster. This namely functions in the case of Example 3.3.10. Finally, to expose the problem in its full complexity, it appears that it is not necessary to consider the divergence equations on the elements of \mathcal{C} sharing a side with an element $L \notin \mathcal{C}$. Let again \mathcal{C} be a set of elements of \mathcal{T}_h and let $\mathcal{G}_{\mathcal{C}}$ be the set of sides of \mathcal{C} between an element $K \in \mathcal{C}$ and $L \notin \mathcal{C}$. Let $\mathcal{E}_{\mathcal{C}}$ stand for all non-Neumann sides of \mathcal{C} and $\mathcal{F}_{\mathcal{C}}$ for $\mathcal{E}_{\mathcal{C}} \setminus \mathcal{G}_{\mathcal{C}}$. We notice that on the rows of the submatrix \mathbb{A} of (3.3) associated with the sides from $\mathcal{F}_{\mathcal{C}}$ and on the rows of the submatrix \mathbb{B} associated with the elements from \mathcal{C} , the only nonzero entries are on the columns associated with the sides from $\mathcal{E}_{\mathcal{C}}$. Hence, to carry out the condensation, it is sufficient if the submatrix consisting of the above rows has a rank equal to $|\mathcal{E}_{\mathcal{C}}|$. The main open problem, which resembles the existence of “singular triangles” in [37, 121], is whether there always has to exist a system of clusters covering \mathcal{T}_h with the above property. Next, in the case of clusters associated with vertices, we have the simple expression (3.10) for the fluxes through all non-Neumann sides. For general clusters, however, we have to associate a weight α_{σ}^i to each side $\sigma \in \mathcal{E}_h$ and i -th out of b clusters \mathcal{C} where σ belongs to $\mathcal{F}_{\mathcal{C}}$, such that $\sum_{i=1}^b \alpha_{\sigma}^i = 1$, in order to have $\sum_{i=1}^b \alpha_{\sigma}^i U_{\sigma}^i = U_{\sigma}$, where U_{σ}^i is the expression of the flux through σ from the i -th cluster. Another interesting open problem is whether one could influence the stencil, symmetry, and positive definiteness of the system matrix by a suitable choice of these weights. For the moment, we have only focused on the basic case. Throughout all the tests presented in Section 3.5, which involve general meshes and inhomogeneous and anisotropic (nonconstant full-matrix) diffusion tensors, we have used the local condensation matrices associated with vertices. These were always invertible, although not always positive definite.

In the lowest-order Raviart–Thomas mixed finite element method on rectangular meshes or in the lowest-order Brezzi–Douglas–Marini mixed finite element method [31, 32] on simplicial meshes, it is either not possible to create subsets \mathcal{C} of \mathcal{T}_h such that each element of \mathcal{C} shares at

most one side with an element $L \notin \mathcal{C}$, or the number of degrees of freedom of vector unknowns per side is greater than the number of degrees of freedom of scalar unknowns per element. Hence the basic form of the condensation with clusters around vertices does not apply. On the other hand, for both Raviart–Thomas and Brezzi–Douglas–Marini mixed finite elements of second order on simplicial meshes, the two above properties are satisfied. The extension of the basic condensation to this case, which may lead to an interesting relation between these second-order mixed finite element methods and the discontinuous Galerkin method, is an ongoing work.

3.4 Application to nonlinear parabolic problems

We show in this section that the above ideas easily apply also to the discretization of nonlinear parabolic convection–reaction–diffusion problems. We consider in particular the problem

$$\frac{\partial \beta(p)}{\partial t} + \nabla \cdot \mathbf{u} + F(p) = q \quad \text{in } \Omega \times (0, T), \quad (3.16a)$$

$$\mathbf{u} = -\mathbf{S}\nabla p + \psi(p)\mathbf{w} \quad \text{in } \Omega \times (0, T), \quad (3.16b)$$

$$p(\cdot, 0) = p_0 \quad \text{in } \Omega, \quad (3.16c)$$

$$p = p_D \quad \text{on } \Gamma_D \times (0, T), \quad (3.16d)$$

$$\mathbf{u} \cdot \mathbf{n} = u_N \quad \text{on } \Gamma_N \times (0, T), \quad (3.16e)$$

where β , ψ , and F are nonlinear functions, \mathbf{S} is again a bounded, symmetric, and uniformly positive definite tensor, \mathbf{w} is a velocity field, and q represents a source term.

Let again $\tilde{\mathbf{u}}$ be such that $\tilde{\mathbf{u}} \cdot \mathbf{n} = u_N$ on Γ_N in the appropriate sense. We split up the time interval $(0, T)$ such that $0 = t_0 < \dots < t_n < \dots < t_N = T$ and define $\Delta t_n := t_n - t_{n-1}$, $n \in \{1, 2, \dots, N\}$, and $p_h^0|_K$ by $(p_0, 1)_K/|K|$ for all $K \in \mathcal{T}_h$. The fully implicit lowest-order Raviart–Thomas mixed finite element approximation of the problem (3.16a)–(3.16e), cf. [14], consists in finding on each time level t_n , $n \in \{1, 2, \dots, N\}$, the functions $\mathbf{u}_h^n = \mathbf{u}_{0,h}^n + \tilde{\mathbf{u}}^n$, $\mathbf{u}_{0,h}^n \in \mathbf{V}(\mathcal{E}_h)$, and $p_h^n \in \Phi(\mathcal{T}_h)$ such that

$$\begin{aligned} (\mathbf{S}^{-n} \mathbf{u}_{0,h}^n, \mathbf{v}_h)_\Omega - (\nabla \cdot \mathbf{v}_h, p_h^n)_\Omega - (\psi(p_h^n) \mathbf{w}^n, \mathbf{S}^{-n} \mathbf{v}_h)_\Omega &= -\langle \mathbf{v}_h \cdot \mathbf{n}, p_D^n \rangle_{\partial\Omega} \\ -(\mathbf{S}^{-n} \tilde{\mathbf{u}}^n, \mathbf{v}_h)_\Omega \quad \forall \mathbf{v}_h \in \mathbf{V}(\mathcal{E}_h), \end{aligned} \quad (3.17a)$$

$$\begin{aligned} \left(\frac{\beta(p_h^n) - \beta(p_h^{n-1})}{\Delta t_n}, \phi_h \right)_\Omega + (\nabla \cdot \mathbf{u}_{0,h}^n, \phi_h)_\Omega + (F(p_h^n), \phi_h)_\Omega &= (q, \phi_h)_\Omega \\ -(\nabla \cdot \tilde{\mathbf{u}}^n, \phi_h)_\Omega \quad \forall \phi_h \in \Phi(\mathcal{T}_h), \end{aligned} \quad (3.17b)$$

where

$$\begin{aligned} \mathbf{S}^{-n} &:= \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \mathbf{S}^{-1}(\cdot, t) dt, \quad \mathbf{w}^n := \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \mathbf{w}(\cdot, t) dt, \\ p_D^n &:= \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} p_D(\cdot, t) dt, \quad \tilde{\mathbf{u}}^n := \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \tilde{\mathbf{u}}(\cdot, t) dt \quad n \in \{1, 2, \dots, N\}. \end{aligned}$$

Note that if $\beta = F = \psi = 0$, the matrix form of the problem (3.17a)–(3.17b) is given by (3.3), where the second equation is multiplied by -1 . Such system matrix is not symmetric, but is positive definite, which is a favorable starting form for (3.17a)–(3.17b).

Everything we have to say about the application of the proposed condensation to the system (3.17a)–(3.17b) is that the terms where the unknown discrete velocity function $\mathbf{u}_{0,h}^n$

appears are exactly the same as in the linear elliptic diffusion case, see (3.2a)–(3.2b). Hence one can eliminate $\mathbf{u}_{0,h}^n$ on each discrete time level as in Section 3.2. This time, the flux unknowns are *nonlinear* functions of the scalar unknowns, convection velocity field, sources, and boundary conditions. The system (3.17a)–(3.17b), linearized by e.g. the Newton method, can be written in the matrix form as

$$\begin{pmatrix} \mathbb{A} & \mathbb{C} \\ \mathbb{B} & \mathbb{D} \end{pmatrix} \begin{pmatrix} U \\ P \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}. \quad (3.18)$$

Let $V \in \mathcal{V}_h$ be a vertex and \mathcal{C}_V the associated cluster and let us consider the linearized equations (3.17a) for the basis functions \mathbf{v}_γ , $\gamma \in \mathcal{F}_{\mathcal{C}_V}$, and the linearized equations (3.17b) for all ϕ_K , $K \in \mathcal{C}_V^{\text{el}}$. This gives

$$\begin{pmatrix} \mathbb{A}_{1,V} & \mathbb{A}_{2,V} \\ \mathbb{B}_{1,V} & \mathbb{B}_{2,V} \end{pmatrix} \begin{pmatrix} U_V^{\mathcal{F}} \\ U_V^{\mathcal{G}} \end{pmatrix} = \begin{pmatrix} F_V - \mathbb{C}_V P_{1,V} \\ G_V - \mathbb{D}_V P_{2,V} \end{pmatrix}. \quad (3.19)$$

In fact, in the present case, $P_{1,V} = P_{2,V} = \{P_K\}_{K \in \mathcal{C}_V}$. We shall need the form (3.19) below for the upwind-mixed method. The matrix $\mathbb{B}_{2,V}$ is still diagonal, and hence we easily have

$$\mathbb{M}_V U_V^{\mathcal{F}} = F_V - \mathbb{C}_V P_{1,V} - \mathbb{A}_{2,V} \mathbb{B}_{2,V}^{-1} (G_V - \mathbb{D}_V P_{2,V}), \quad (3.20)$$

where the local condensation matrix associated with the vertex V , $\mathbb{M}_V = \mathbb{A}_{1,V} - \mathbb{A}_{2,V} \mathbb{B}_{2,V}^{-1} \mathbb{B}_{1,V}$, is the same as in the linear elliptic diffusion case. Hence its invertibility and the feasibility of the condensation in this form is determined by the rules studied in Section 3.3. Shall \mathbb{M}_V be invertible for all $V \in \mathcal{V}_h$, we have

$$U = \tilde{\mathbb{A}}^{-1} (F - \mathbb{C}P) - \mathbb{J} (G - \mathbb{D}P),$$

using (3.10). Here $\tilde{\mathbb{A}}^{-1}$ and \mathbb{J} are given by (3.12). It now suffices to insert this expression for U into the second equation of (3.18) to obtain the final system for the scalar unknowns P only,

$$(-\mathbb{B} \tilde{\mathbb{A}}^{-1} \mathbb{C} + \mathbb{B} \mathbb{J} \mathbb{D} + \mathbb{D})P = G - \mathbb{B} \tilde{\mathbb{A}}^{-1} F + \mathbb{B} \mathbb{J} G. \quad (3.21)$$

This transcription enables in particular a straightforward implementation of the proposed condensation in any mixed finite element code.

Remark 3.4.1. (Assemblage of $\tilde{\mathbb{A}}^{-1}$ and \mathbb{J}) *We note that the matrices $\tilde{\mathbb{A}}^{-1}$ and \mathbb{J} only depend on the matrices \mathbb{A}, \mathbb{B} of (3.18). Hence, if these matrices do not change (i.e. when the diffusion tensor \mathbf{S} is constant with respect to time), the assemblage of $\tilde{\mathbb{A}}^{-1}$ and \mathbb{J} can be done only once before the start of the calculation. On each time and linearization step, one then needs only $\mathbb{C}, \mathbb{D}, F$, and G from (3.18) to assemble the final linear system (3.21).*

We now finally turn to the upwind-mixed lowest-order Raviart–Thomas method, cf. [46, 47, 77]. For this purpose, we first rewrite (3.16a)–(3.16b) as

$$\begin{aligned} \frac{\partial \beta(p)}{\partial t} + \nabla \cdot \mathbf{r} + \nabla \cdot (\psi(p) \mathbf{w}) + F(p) &= q \quad \text{in } \Omega \times (0, T), \\ \mathbf{r} &= -\mathbf{S} \nabla p \quad \text{in } \Omega \times (0, T). \end{aligned}$$

Whereas the initial and Dirichlet boundary conditions (3.16c) and (3.16d) stay the same, we rewrite the Robin boundary condition (3.16e) as a Neumann one,

$$\mathbf{r} \cdot \mathbf{n} = v_N \quad \text{on } \Gamma_N \times (0, T).$$

Let again $\tilde{\mathbf{r}}$ be such that $\tilde{\mathbf{r}} \cdot \mathbf{n} = v_N$ on Γ_N in the appropriate sense and define $\tilde{\mathbf{r}}^n := \frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \tilde{\mathbf{r}}(\cdot, t) dt$, $n \in \{1, 2, \dots, N\}$. The fully implicit upwind-mixed finite element method then reads: on each time level t_n , $n \in \{1, 2, \dots, N\}$, find the functions $\mathbf{r}_h^n = \mathbf{r}_{0,h}^n + \tilde{\mathbf{r}}^n$, $\mathbf{r}_{0,h}^n \in \mathbf{V}(\mathcal{E}_h)$, and $p_h^n \in \Phi(\mathcal{T}_h)$ such that

$$\begin{aligned} (\mathbf{S}^{-n} \mathbf{r}_{0,h}^n, \mathbf{v}_h)_\Omega - (\nabla \cdot \mathbf{v}_h, p_h^n)_\Omega &= -\langle \mathbf{v}_h \cdot \mathbf{n}, p_D^n \rangle_{\partial\Omega} \\ -(\mathbf{S}^{-n} \tilde{\mathbf{r}}^n, \mathbf{v}_h)_\Omega &\quad \forall \mathbf{v}_h \in \mathbf{V}(\mathcal{E}_h), \end{aligned} \quad (3.23a)$$

$$\begin{aligned} \left(\frac{\beta(P_K^n) - \beta(P_K^{n-1})}{\Delta t_n}, \phi_K \right)_K + (\nabla \cdot \mathbf{r}_{0,h}^n, \phi_K)_K + \sum_{\sigma \in \mathcal{E}_K} \psi(\widehat{p}_\sigma^n) \mathbf{w}_{K,\sigma}^n + (F(P_K^n), \phi_K)_K \\ = (q, \phi_K)_K - (\nabla \cdot \tilde{\mathbf{r}}^n, \phi_K)_K \quad \forall K \in \mathcal{T}_h, \end{aligned} \quad (3.23b)$$

where $\mathbf{w}_{K,\sigma}^n = \langle \mathbf{w}^n \cdot \mathbf{n}, 1 \rangle_\sigma$ and \widehat{p}_σ^n is the upwind value defined by

$$\widehat{p}_\sigma^n := \begin{cases} P_K^n & \text{if } \mathbf{w}_{K,\sigma}^n \geq 0 \\ P_L^n & \text{if } \mathbf{w}_{K,\sigma}^n < 0 \end{cases}$$

if σ is an interior side between the elements K and L ,

$$\widehat{p}_\sigma^n := \begin{cases} P_K^n & \text{if } \mathbf{w}_{K,\sigma}^n \geq 0 \\ \langle p_D^n, 1 \rangle_\sigma / |\sigma| & \text{if } \mathbf{w}_{K,\sigma}^n < 0 \end{cases}$$

if σ is a Dirichlet boundary side, and $\widehat{p}_\sigma^n := P_K^n$ if σ is a Neumann boundary side. The linearization of the system (3.23a)–(3.23b) has again the form (3.18), with this time $\mathbb{C} = -\mathbb{B}^t$. The condensation applies again directly and in particular the final system has the form (3.21). The only difference is that because of the upstream weighting, $P_{1,V} \neq P_{2,V}$ in (3.19). In the expression for the fluxes through the $\mathcal{F}_{\mathcal{C}_V}$ sides, all the scalar unknowns associated with the elements sharing a side with an element from the cluster \mathcal{C}_V may appear. Hence also the stencil of the final matrix is in this case wider: on a row of the final matrix corresponding to an element $K \in \mathcal{T}_h$, the only possible nonzero entries are on columns corresponding to $L \in \mathcal{T}_h$ such that L shares a common side with an element $M \in \mathcal{T}_h$ such that M and K share a common vertex. Finally, a similar observation to Remark 3.4.1 holds also in this case. Shall \mathbb{A} and \mathbb{B} be constant, we only need to upload \mathbb{D} , F , and G on each time and linearization step, as in the finite volume method.

3.5 Numerical experiments

We give the results of several numerical experiments in two space dimensions in this section. We first compare the computational cost of the proposed condensation of the lowest-order mixed finite element method with standard mixed solution approaches for elliptic and parabolic problems. We then present a comparison of precision and efficiency between the condensed mixed finite element, finite volume, and combined finite volume–finite element methods for a nonlinear parabolic problem. We use the basic local condensation matrices associated with vertices; these are not always positive definite, but are always invertible.

We perform the simulations on unstructured triangular meshes, given as regular refinements (each triangle is refined into four triangles by joining its edges midpoints) of some mesh from Figure 3.7. In the mesh A, the minimal and maximal angles are equal to 29.1 and 82.7 degrees, in the mesh B to 29.1 and 84.8 degrees, in the mesh C to 15.3 and 135 degrees, and in the mesh D to 15.3 and 142 degrees, respectively. We denote the number of refinements by r ($r = 0$

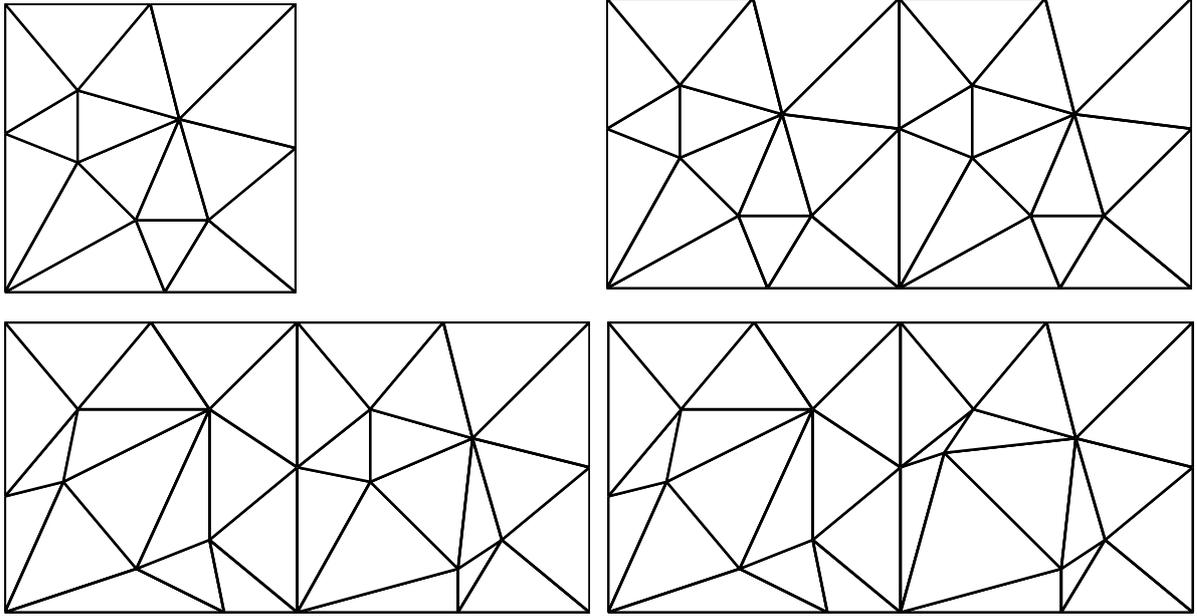


Figure 3.7: Initial meshes A (top left), B (top right), C (bottom left), and D (bottom right)

corresponds to the initial mesh). For a parabolic problem on a time interval $(0, T)$, the initial time step is equal to $T/2$ and is divided by two each time the space mesh is refined. Since the discrete maximum principle is not necessarily satisfied by the considered numerical schemes, we prolong each function $\beta(p)$ only defined for positive values also for negative values by $\beta(-p) = -\beta(p)$. In the tables with results, we shall use the abbreviation SPD for a symmetric positive definite matrix, NPD for a nonsymmetric but positive definite matrix (recall that a real matrix $\mathbb{M} \in \mathbb{R}^{M \times M}$ is positive definite if $X^t \mathbb{M} X > 0$ for all $X \in \mathbb{R}^M$, $X \neq 0$), and NID for a nonsymmetric indefinite matrix. We further use *st.* for the stencil, i.e. for the maximum number of nonzero entries on each matrix row, and *cond.* for the 2-norm condition number (defined for a matrix \mathbb{M} by $\|\mathbb{M}\|_2 \|\mathbb{M}^{-1}\|_2$, or equivalently by the ratio of its largest and smallest singular value).

We employ two iterative methods for the solution of systems of linear equations with sparse matrices. If the matrix is symmetric and positive definite, we use the conjugate gradients (CG) method [73, 103]. For nonsymmetric matrices, we employ the bi-conjugate gradients stabilized (Bi-CGStab, in tables abbreviated as Bi-CGS) method [103, 117]. In all considered cases, the nonsymmetric matrices are negative-stable (all their eigenvalues have positive real parts, which is in particular the case when the system matrix is positive definite), which is an essential requirement for a reasonably fast convergence of the Bi-CGStab method. To accelerate the convergence of these methods, we use incomplete Cholesky (IC) or incomplete LU (ILU) factorizations with a specified drop tolerance, cf. [111]. We denote the preconditioned methods by PCG and PBi-CGStab (PBi-CGS), respectively. In order to stop the iterative process, we monitor the relative residual $\|Y - \mathbb{M}\tilde{X}\|_2 / \|Y\|_2$, where \tilde{X} is the approximate solution to the system $\mathbb{M}X = Y$. We focus on iterative solvers since they permit an efficient solution of nonlinear problems, combined with the Newton method and a suitable preconditioning.

All the computations were done in a C++ code in double precision on a notebook with Intel Pentium 4-M 1.8 GHz processor and MS Windows XP operating system. Machine precision was in power of 10^{-16} . All the matrix operations were done with the help of MATLAB 6.1.

	Ref.	Unkn.	Matr.	St.	Cond.	Bi-CGS	Iter.	CG	Iter.
CMFE	3	1024	NPD	14	721	0.20	76.5		
	4	4096	NPD	14	2882	1.43	147.5		
	5	16384	NPD	14	11523	12.55	295.5		
	6	65536	NPD	14	46093	117.58	555.5		
MHFE	3	1504	SPD	5	1397	0.31	118.0	0.22	157
	4	6080	SPD	5	5616	2.43	230.5	1.75	316
	5	24448	SPD	5	22499	23.40	449.5	16.87	623
	6	98048	SPD	5	89995	227.04	864.0	162.09	1226

Table 3.1: Comparison of the computational cost of the condensed and hybridized mixed finite element methods, problem (3.24), mesh A

3.5.1 Condensed mixed finite element method for elliptic problems

For $\Omega = (0, 1) \times (0, 1)$, we consider the problem

$$-\Delta p = -2e^x e^y \quad (3.24)$$

with a Dirichlet boundary condition given by the exact solution $p(x, y) = e^x e^y$ and the initial mesh A from Figure 3.7. We compare the computational cost of the condensation of the mixed finite element method proposed in this chapter and of the hybridization of the mixed finite element method onto the Lagrange multipliers associated with the edges.

We compare the number of unknowns (number of triangles for the condensed and number of edges for the hybridized mixed finite element method), symmetry, positive definiteness, stencil, and the condition number of the system matrices in Table 3.1. Recall that the system matrix of the mixed-hybrid method is in the given case equivalent to that of the nonconforming finite element method, cf. [38], Lemma 1.8.1, or a more detailed study in Section 4.6 of this thesis. We further give the CPU time in seconds and the number of iterations necessary to decrease the relative residual below $1e-10$. We have used a zero start vector.

One needs about 1.35-times less CPU time for the condensed version than for the hybridized version of the mixed finite element method. If the system matrix may become nonsymmetric (convection–diffusion problems, nonsymmetric diffusion tensors), then e.g. the Bi-CGStab method will be necessary also for the mixed-hybrid method. One may expect more important computational savings in this case, as it is indicated by the use of this method in the present case (up to 2-times less the CPU time). Essential computational savings are however confirmed in the next section for (nonlinear) parabolic problems, where also a detailed study of the influence of the coefficients of the equation at hand, of the shapes of the elements of the mesh, and of the preconditioning is performed.

3.5.2 Condensed mixed finite element method for nonlinear parabolic problems

We compare in this section the condensed and standard mixed finite element methods for several nonlinear parabolic (convection–)reaction–diffusion problems, which may involve discontinuous coefficients and inhomogeneous and anisotropic diffusion tensors.

	Unkn.	Matr.	St.	Cond.	Bi-CGS	Iter.	CPU	ILU	PBi-CGS	Iter.
CMFE	128	NPD	14	37	0.02	29.5	0.02	0.01	0.01	2.0
	512	NPD	14	109	0.06	50.0	0.02	0.01	0.01	2.5
	2048	NPD	14	298	0.37	80.5	0.10	0.06	0.04	3.0
	8192	NPD	14	747	2.45	122.5	0.68	0.38	0.30	5.0
	32768	NPD	14	1753	14.24	175.0	4.75	2.95	1.80	7.0
					CG	IC			PCG	
MFE	204	SPD	5	290	0.05	110	0.02	0.01	0.01	5
	792	SPD	5	764	0.14	206	0.04	0.02	0.02	7
	3120	SPD	5	1770	0.95	333	0.18	0.08	0.10	11
	12384	SPD	5	3820	5.36	508	1.21	0.58	0.63	14
	49344	SPD	5	7974	34.45	743	8.17	3.83	4.34	18

Table 3.2: Comparison of the computational cost of the condensed and standard mixed finite element methods, first time and linearization step, problem (3.25), tensor (3.26), mesh B

A reaction–diffusion problem

For $\Omega = (0, 2) \times (0, 1)$ and a time interval $(0, 1)$, we consider the nonlinear reaction–diffusion problem

$$\frac{\partial(p + p^\alpha)}{\partial t} - \nabla \cdot (\mathbf{S}\nabla p) + 3p + \alpha p^\alpha = 0 \quad (3.25)$$

with $\alpha = 0.5$ and either

$$\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ in } \Omega \quad (3.26)$$

or

$$\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ for } x < 1, \quad \mathbf{S} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \text{ for } x > 1. \quad (3.27)$$

Initial and Dirichlet boundary conditions are given by the exact solution $p(x, y, t) = e^x e^y e^{-t} / e^3$. Notice that the flux of the solution given by $-\mathbf{S}\nabla p$ has a continuous normal trace across the discontinuity line $x = 1$ for the diffusion tensor (3.27). The derivative of the function p^α , $\alpha = 0.5$, blows up in 0 but the problem is not really degenerate parabolic, since the exact solution does not take the value of 0. We perform the simulations starting from the meshes B and C from Figure 3.7. We consider the condensation of the mixed finite element method (3.21) and the mixed finite element method (3.17a)–(3.17b). We notice that the system of equations of the mixed method has on each time and linearization step the form (3.18), where \mathbb{D} is a diagonal matrix. Hence a standard solution approach is to inverse \mathbb{D} , then solve for U the system $(\mathbb{A} - \mathbb{C}\mathbb{D}^{-1}\mathbb{B})U = F - \mathbb{C}\mathbb{D}^{-1}G$, and finally recover P from $P = \mathbb{D}^{-1}(G - \mathbb{B}U)$. In fact, in the present case, $\mathbb{C} = \mathbb{B}^t$, and thus the final system matrix is symmetric. It is noted in [74] that this approach is not suitable when the term occurring in the time derivative and the reaction term are too small in comparison with the other terms, which is however not the present case. On the contrary, according to [74], such solution approach is more reliable than the hybridization of the mixed finite element method for general diffusion tensors.

We compare the properties of the system matrices on the first time and Newton linearization steps in Table 3.2, considering the tensor (3.26) and the initial mesh B. We further give

	Unkn.	Matr.	St.	Cond.	Bi-CGS	Iter.	CPU	ILU	PBi-CGS	Iter.
CMFE	128	NPD	14	61	0.02	31.0	0.02	0.01	0.01	2.0
	512	NPD	14	225	0.07	62.0	0.02	0.01	0.01	2.5
	2048	NPD	14	676	0.49	119.0	0.12	0.07	0.05	3.5
	8192	NPD	14	1814	3.76	212.0	0.75	0.39	0.36	6.0
	32768	NPD	14	4603	26.62	335.5	5.31	3.02	2.29	8.0
				CG			IC	PCG		
MFE	204	SPD	5	1358	0.07	170	0.02	0.01	0.01	6
	792	SPD	5	4314	0.36	409	0.04	0.02	0.02	10
	3120	SPD	5	11506	2.23	836	0.28	0.12	0.16	16
	12384	SPD	5	28188	15.93	1456	1.61	0.68	0.93	20
	49344	SPD	5	65024	108.51	2279	11.75	5.76	5.99	27

Table 3.3: Comparison of the computational cost of the condensed and standard mixed finite element methods, first time and linearization step, problem (3.25), tensor (3.27), mesh C

the CPU time in seconds and the number of iterations necessary to decrease the relative residual below $1e-8$ by the Bi-CGStab and CG methods, respectively, without any preconditioning. We then use the incomplete LU and Cholesky factorizations, respectively, as preconditioners. The drop tolerance is chosen in such way that the sum of CPU times for the preconditioning and the solution of the preconditioned system was minimal. We give in Table 3.2 the separate times as well as their sum (CPU) and the number of iterations of the preconditioned method. We have always used a zero start vector. We report finally in Table 3.3 the same values for the case of the tensor (3.27) and the initial mesh C.

The CPU time of the condensed mixed finite element method is about 2-times shorter than the CPU time of the standard approach in the case of the tensor (3.26) and the initial mesh B. When full-matrix and discontinuous diffusion tensor (3.27) and a less regular mesh C are used, then the CPU time of the condensed version is more than 4-times shorter when no preconditioning is used and more than 2-times shorter with preconditioning. Note the important increase of the condition number of the system matrix of the standard mixed finite element method for the tensor (3.27).

A convection–reaction–diffusion problem

For $\Omega = (0, 2) \times (0, 1)$ and a time interval $(0, 1)$, we consider the nonlinear convection–reaction–diffusion problem

$$\frac{\partial(p + p^\alpha)}{\partial t} - \nabla \cdot (\mathbf{S}\nabla p) + \nabla \cdot (p\mathbf{w}) + \alpha p^\alpha = 0 \quad (3.28)$$

with $\alpha = 0.5$ and either

$$\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ in } \Omega, \quad \mathbf{w} = (3, 0) \text{ in } \Omega \quad (3.29)$$

or

$$\begin{aligned} \mathbf{S} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ for } x < 1, & \mathbf{S} &= \begin{pmatrix} 8 & -7 \\ -7 & 20 \end{pmatrix} \text{ for } x > 1, \\ \mathbf{w} &= (3, 0) \text{ for } x < 1, & \mathbf{w} &= (3, 12) \text{ for } x > 1. \end{aligned} \quad (3.30)$$

	Unkn.	Matr.	St.	Cond.	Bi-CGS	Iter.	CPU	ILU	PBi-CGS	Iter.
CMFE	128	NPD	14	39	0.02	27.0	0.02	0.01	0.01	2.0
	512	NPD	14	116	0.07	56.5	0.02	0.01	0.01	2.5
	2048	NPD	14	311	0.38	82.5	0.11	0.06	0.05	3.5
	8192	NPD	14	768	2.65	139.0	0.75	0.41	0.34	5.5
	32768	NPD	14	1782	17.14	191.5	4.85	2.95	1.90	7.0
MFE	204	NPD	5	405	0.06	95.5	0.02	0.01	0.01	2.0
	792	NPD	5	917	0.22	153.0	0.07	0.03	0.04	3.0
	3120	NPD	5	1949	1.36	282.0	0.34	0.14	0.20	4.0
	12384	NPD	5	4016	8.47	406.5	2.57	0.94	1.63	5.0
	49344	NPD	5	8181	51.18	553.0	17.63	6.94	10.69	6.0

Table 3.4: Comparison of the computational cost of the condensed and standard mixed finite element methods, first time and linearization step, problem (3.28), coefficients (3.29), mesh B

Initial and Dirichlet boundary conditions are again given by the exact solution $p(x, y, t) = e^x e^y e^{-t}/e^3$. Notice that in the case of the coefficients given by (3.30), the velocity field \mathbf{w} as well as the flux of the solution given by $-\mathbf{S}\nabla p + (p\mathbf{w})$ have a continuous normal trace across the discontinuity line $x = 1$. We perform the simulations on refinements of the meshes B, C, and D from Figure 3.7. The problem is not convection-dominated, and hence we can use the mixed finite element method (3.17a)–(3.17b). Notice that the associated linear system on each time and linearization step has again the form (3.18) with \mathbb{D} a diagonal matrix. Hence the same solution approach as in the previous section can be used. In this case however $\mathbb{C} \neq \mathbb{B}^t$, and thus the final system for U is nonsymmetric.

We compare the properties of the linear systems on the first time and Newton linearization steps for different combinations of coefficients and meshes in Tables 3.4, 3.5, and 3.6. The settings are the same as in the previous section, except for the fact that we have to use the Bi-CGStab method and the LU incomplete factorization also for the standard mixed approach in view of the nonsymmetry of its system matrices.

One can observe that the increase of the condition number of the system matrix of the condensed mixed finite element method with less regular coefficients and meshes is much less important than that of the standard mixed finite element method. Hence the CPU time of the unpreconditioned Bi-CGStab method for the condensed version is about 3-times shorter for the coefficients (3.29) and mesh B, but about 10-times shorter for the coefficients (3.30) and meshes C and D. Using the preconditioning by the LU incomplete factorization considerably smears the difference. The CPU time of the condensed version is then about 3.5-times shorter for the coefficients (3.29) and mesh B and 4-times shorter for the coefficients (3.30) and mesh C. The system matrix of the condensed mixed finite element method loses the positive definiteness property while changing from the mesh C to the mesh D when the coefficients (3.30) are considered. It appears that this transition is connected with a remarkable 50% increase in the CPU time in the case where the preconditioning is used. The system matrix of the standard mixed finite element method stays positive definite and the increase in the CPU time while changing from mesh C to mesh D with the coefficients (3.30) is about 10%. Hence the CPU time of the condensed version is about 3-times shorter for the coefficients (3.30) and mesh D, while using the preconditioning.

	Unkn.	Matr.	St.	Cond.	Bi-CGS	Iter.	CPU	ILU	PBi-CGS	Iter.
CMFE	128	NPD	14	470	0.04	70.0	0.02	0.01	0.01	2.0
	512	NPD	14	1665	0.21	149.5	0.03	0.01	0.02	2.5
	2048	NPD	14	4824	1.47	322.5	0.12	0.07	0.05	3.5
	8192	NPD	14	12523	8.66	474.5	0.88	0.56	0.32	5.0
	32768	NPD	14	31368	61.53	787.5	7.47	5.46	2.01	5.5
MFE	204	NPD	5	13849	0.23	412.5	0.02	0.01	0.01	2.0
	792	NPD	5	39935	1.38	1105.5	0.04	0.02	0.02	2.5
	3120	NPD	5	103342	12.12	2419.5	0.41	0.18	0.23	3.0
	12384	NPD	5	250923	103.42	5390.5	3.06	1.32	1.74	3.5
	49344	NPD	5	586375	617.26	7145.5	29.88	14.96	14.92	4.0

Table 3.5: Comparison of the computational cost of the condensed and standard mixed finite element methods, first time and linearization step, problem (3.28), coefficients (3.30), mesh C

	Unkn.	Matr.	St.	Cond.	Bi-CGS	Iter.	CPU	ILU	PBi-CGS	Iter.
CMFE	128	NID	14	613	0.03	74.0	0.02	0.01	0.01	2.0
	512	NID	14	2232	0.26	185.5	0.03	0.01	0.02	2.5
	2048	NID	14	6512	1.57	326.5	0.19	0.09	0.10	3.5
	8192	NID	14	16755	9.41	521.5	1.37	0.61	0.76	4.0
	32768	NID	14	41716	73.20	903.5	11.89	5.36	6.53	5.5
MFE	204	NPD	5	17156	0.25	441.5	0.02	0.01	0.01	2.0
	792	NPD	5	49995	1.32	1063.5	0.05	0.02	0.03	2.5
	3120	NPD	5	131073	13.62	2691.5	0.47	0.20	0.27	3.0
	12384	NPD	5	319842	87.93	4222.5	3.36	1.62	1.74	3.5
	49344	NPD	5	750271	686.64	7248.0	32.68	15.77	16.91	4.5

Table 3.6: Comparison of the computational cost of the condensed and standard mixed finite element methods, first time and linearization step, problem (3.28), coefficients (3.30), mesh D

	Unkn.	Matr.	St.	Cond.	Bi-CGS	Iter.	CPU	Per.	ILU	PBi-CGS	Iter.
CU-MFE	128	NPD	19	42	0.02	25.5	0.02		0.01	0.01	2.0
	512	NPD	19	120	0.09	57.0	0.02		0.01	0.01	2.5
	2048	NPD	19	318	0.46	88.0	0.11		0.06	0.05	3.0
	8192	NPD	19	777	2.99	138.5	0.68		0.36	0.32	5.0
	32768	NPD	19	1792	18.86	210.5	4.89		2.87	2.02	7.5
U-MFE	332	NPD	7	17	0.18	235.5	0.03	0.01	0.01	0.01	2.0
	1304	NPD	7	29	1.17	549.5	0.09	0.01	0.03	0.05	2.5
	5168	NPD	7	67	13.01	1540.5	0.48	0.03	0.15	0.30	3.5
	20576	NPD	7	168	124.06	3561.5	2.83	0.32	0.98	1.53	4.0
	82112	NPD	7	393	3233.05	16763.5	15.70	1.35	4.92	9.43	6.0

Table 3.7: Comparison of the computational cost of the condensed and standard upwind-mixed finite element methods, first time and linearization step, problem (3.28), coefficients (3.29), mesh B

A convection–reaction–diffusion problem and the upwind-mixed method

We consider here once more the problem (3.28) with coefficients (3.29) and mesh B. This time, we employ the upwind-mixed finite element method (3.23a)–(3.23b) and the corresponding condensed version.

We compare the properties of the linear systems on the first time and Newton linearization steps in Table 3.7. Although there is an increase in the stencil of the condensed upwind-mixed finite element method, the system matrix condition number and CPU times are very much like for the condensed mixed finite element method, cf. Table 3.4. The system for the upwind-mixed finite element method on each time and linearization step has again the form (3.18). The matrix \mathbb{D} is however in this case not diagonal, and hence we cannot easily eliminate the scalar unknowns P . We thus consider the whole matrix for the unknowns U and P . This matrix is very well conditioned, nonsymmetric, positive definite, and negative-stable, but the direct application of the Bi-CGStab method does not lead to satisfactory results, cf. Table 3.7. Also the direct LU incomplete factorization is almost impossible, since the LU factors tend to considerably increase the fill-in. A suitable solution approach however seems to be to first perform the column minimum degree permutation [70]. The matrix with permuted columns then has sparser LU incomplete factors, which can in turn be successfully used as preconditioners. We report in Table 3.7 the CPU times necessary for finding the column minimum degree permutation, LU incomplete factorization of the matrix with permuted columns, and for the solution of the preconditioned systems by the Bi-CGStab method, as well as the sum of these times (CPU). In the present case, the condensation reduces the CPU time by a factor better than 3.

3.5.3 Comparison of the condensed mixed finite element, finite volume, and combined finite volume–finite element methods

We compare in this section the precision and efficiency of the discretization schemes studied in this thesis for a nonlinear convection–reaction–diffusion problem.

Method	Unkn.	Matr.	St.	Cond.
FV	8192	NPD	4	1704
CMFE	8192	NPD	14	768
FV–FE	4001	NPD	8	281
FV–NCFE	12192	NPD	5	1501

Table 3.8: Characteristics of the system matrices on the first time and linearization steps, problem (3.28), coefficients (3.29), Dirichlet boundary conditions, mesh B, $r = 4$

Method	Unkn.	Matr.	St.	Cond.
CMFE	8192	NID	14	19589
FV–FE	4032	NPD	8	12965
FV–NCFE	12224	NPD	5	59029

Table 3.9: Characteristics of the system matrices on the first time and linearization steps, problem (3.28), coefficients (3.30), Dirichlet–Robin boundary conditions, mesh D, $r = 4$

The first considered scheme is the condensed mixed finite element (CMFE) scheme studied in this chapter. The second scheme is the combined finite volume–nonconforming finite element (FV–NCFE) scheme studied in Chapter 1. Since we will only consider piecewise constant diffusion tensors, this scheme coincides with the combined finite volume–mixed–hybrid finite element one, see Remark 1.3.5. The third scheme is the combined finite volume–finite element (FV–FE) scheme studied in Appendix 1.9 of Chapter 1. In fact, in contrast to Appendix 1.9, we shall start here from the triangular mesh and then construct the dual mesh by joining triangle barycentres with edge midpoints, as originally proposed in [67, 90]. Finally, for triangulations with acute angles and for scalar diffusion tensors, we will also consider the cell-centered finite volume (FV) scheme, cf. [61, 72]. Recall that in the FV and CMFE schemes, the degrees of freedom are associated with triangles, in the FV–FE scheme with vertices, and in the FV–NCFE scheme with edges.

We consider the problem (3.28) with various coefficients, meshes, and boundary conditions. For the FV, FV–FE, and FV–NCFE schemes, we use the local Péclet upstream weighting defined by (1.10), (1.11). Since the problem is not convection-dominated, we employ the CMFE scheme starting from the mixed finite element method (3.17a)–(3.17b). The system matrices properties on the first time and Newton linearization steps for the different methods and 4-times refined original meshes are listed in Tables 3.8 and 3.9. Initial conditions for the CMFE and FV methods are given by the mean values of the known exact solution over the triangles at time $t = 0$. For the FV–FE and FV–NCFE methods, we consider instead the point values of the exact solution at vertices and edge midpoints, respectively, at $t = 0$. The boundary conditions for the FV–FE method are given by the point values of the exact solution at the vertices lying on the boundary. For the other methods, we consider the point values in the boundary edges midpoints. A quintic (7-point) numerical integration formulae on triangles is used to evaluate the initial conditions for CMFE and FV methods, as well for error computation. It has no influence on the considered order of precision of the results.

Tables 3.10 and 3.11 give discrete relative and projection relative errors for all the compared schemes and up to five refinements of the original space-time grid, considering coefficients (3.29), mesh B, and Dirichlet boundary conditions. The discrete $L^\infty(0, T; L^2(\Omega))$

Method \ r	0	1	2	3	4	5
FV	0.15130	0.07538	0.03768	0.01884	0.00942	0.00471
PFV	0.02914	0.01159	0.00551	0.00276	0.00140	0.00070
CMFE	0.14519	0.07221	0.03607	0.01804	0.00902	0.00451
PCMFE	0.03480	0.01249	0.00558	0.00276	0.00139	0.00069
FV–FE	0.04892	0.01665	0.00693	0.00314	0.00149	0.00073
FV–NCFE	0.02642	0.01146	0.00554	0.00278	0.00140	0.00070

Table 3.10: Discrete $L^\infty(0, T; L^2(\Omega))$ relative errors, problem (3.28), coefficients (3.29), mesh B

Method \ r	0	1	2	3	4	5
FV	0.04957	0.02428	0.01215	0.00608	0.00304	0.00152
CMFE	0.02542	0.01099	0.00539	0.00273	0.00138	0.00070
FV–FE	0.13859	0.04922	0.01771	0.00655	0.00252	0.00102
FV–NCFE	0.03595	0.01495	0.00658	0.00306	0.00147	0.00072

Table 3.11: Discrete $L^\infty(0, T; L^2(\Omega))$ projection relative errors, problem (3.28), coefficients (3.29), mesh B

relative error is defined by

$$\max_{n \in \{1, 2, \dots, N\}} \frac{\|p_h^n(\cdot) - p(\cdot, t_n)\|_{0, \Omega}}{\|p(\cdot, t_n)\|_{0, \Omega}},$$

where p_h^n is the approximate solution at time t_n . For the FV–FE and FV–NCFE schemes, we consider piecewise linear approximations, whereas for the FV and CMFE schemes, only a piecewise constant solution is originally at disposal. Hence we use the following postprocessing technique to construct also piecewise linear approximations. In both the FV and CMFE schemes, one can easily evaluate the discrete diffusive fluxes $\{U_{K, \sigma}^{\text{dif}}\}_{\sigma \in \mathcal{E}_K}$ through the edges of each triangle. In the FV scheme, $U_{K, \sigma}^{\text{dif}} = (P_K - P_L)|\sigma|/d_{K, L}$, where K and L are the triangles sharing an interior edge σ and $d_{K, L}$ is the distance between \mathbf{x}_K and \mathbf{x}_L , the circumscribed circles centers (intersections of the edge orthogonal bisectors) of K and L . This follows naturally from the definition of the scheme, cf. [61, 72]. In the considered CMFE scheme, the established unknowns U_σ , $\sigma \in \mathcal{E}_h$, represent full (the sum of convective and diffusive) fluxes through the triangle edges (in the orientation of the RTN basis functions). To obtain only diffusive fluxes, we thus subtract the convective ones, which we approximate by $U_{K, \sigma}^{\text{conv}} \approx (P_K + P_L)\mathbf{w}_\sigma/2$, where again K and L are the triangles sharing the edge σ and $\mathbf{w}_\sigma = \langle \mathbf{w} \cdot \mathbf{n}, 1 \rangle_\sigma$, \mathbf{n} being the unit normal vector of the edge σ , outward to K . Under the same principles, there are slight modifications on the boundary. Then the three fluxes on each triangle define a linear vector function, the diffusive flux, $\mathbf{u}_h^{\text{dif}}|_K := \sum_{\sigma \in \mathcal{E}_K} U_{K, \sigma}^{\text{dif}} \mathbf{v}_\sigma$, where \mathbf{v}_σ is the RTN basis function associated with the side σ , oriented outward from K . Note that for the elliptic diffusion problem (3.1a)–(3.1c), this would be the solution \mathbf{u}_h , approximation of $-\mathbf{S}\nabla p$, for the mixed finite element method. Similar approximation holds in the finite volume method, see [60]. We evaluate the function $\mathbf{u}_h^{\text{dif}}$ in triangle barycentres, so as to make it a constant vector on each K . Now disposing by the approximation of the gradient and by one value per triangle, it is straightforward to reconstruct a linear approximation: we fix it in \mathbf{x}_K by the known value P_K for the finite volume method and by its mean value P_K over K for the mixed finite element method. Note that this approximation is elementwise linear but generally completely discon-

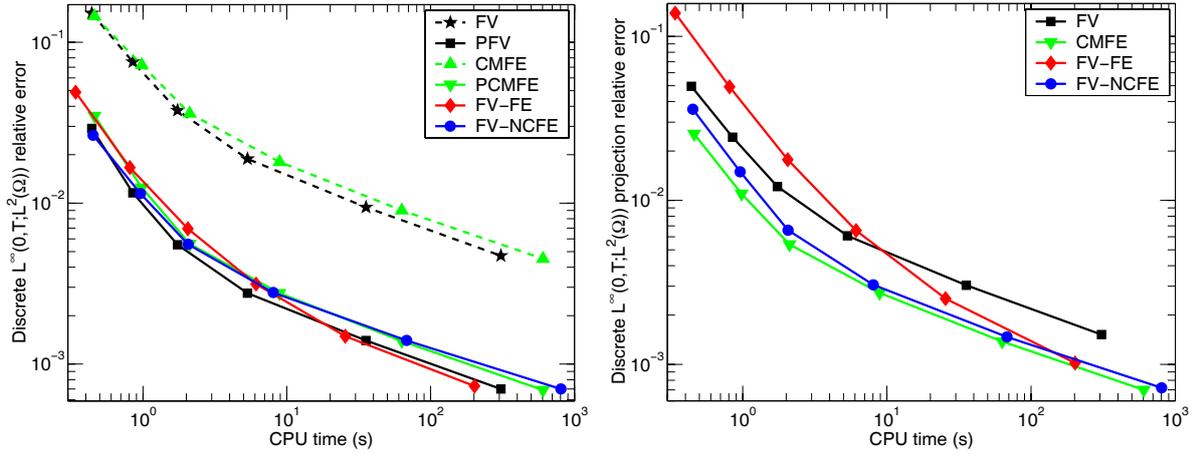


Figure 3.8: Efficiency comparison, problem (3.28), coefficients (3.29), mesh B

tinuous and that the price for its obtaining is negligible, since we do it locally. We denote by PFV and PCMFE the FV and CMFE schemes with these postprocessed solutions.

We define a discrete $L^\infty(0, T; L^2(\Omega))$ projection relative error by

$$\max_{n \in \{1, 2, \dots, N\}} \frac{\|\tilde{p}_h^n(\cdot) - \tilde{p}(\cdot, t_n)\|_{0, \Omega}}{\|p(\cdot, t_n)\|_{0, \Omega}},$$

where \tilde{p}_h^n is the piecewise constant approximate solution at time t_n . For the FV-FE and FV-NCFE schemes, we consider the solutions piecewise constant on the dual volumes, whereas for the FV and CMFE scheme, we use the original elementwise constant results. The function \tilde{p} is given by the mean values of the exact solution p on the dual volumes for the combined schemes and on the triangles for the FV and CMFE schemes. We finally give “efficiency comparisons” in Figure 3.8. We plot the approximation errors against the CPU times of the whole calculations. We have used the inexact Newton method, where on each linearization step, a limited number (three to four) of Bi-CGStab iterations, started from the previous linearization values, was performed. The LU incomplete factorization for preconditioning was done at the beginning of each linearization cycle. The drop tolerance was chosen in order to decrease the relative residual on the first linearization step after the limit number of iterations to about $1e-5$ (the stopping criterion was $1e-8$). The Newton linearization was initiated by the previous time step (initial) values and terminated whenever

$$\left(\sum_{i=1}^M (X_i^{n,k} - X_i^{n,k-1})^2 \right)^{\frac{1}{2}} / \left(\sum_{i=1}^M (X_i^{n,k})^2 \right)^{\frac{1}{2}} \leq 1e-8,$$

where $X^{n,k}$ is the vector of approximation values on time and linearization steps n and k , respectively. Three or four linearization steps on each time level were necessary.

As the FV-FE and FV-NCFE schemes produce piecewise linear approximations, their discrete $L^\infty(0, T; L^2(\Omega))$ relative errors are smaller than those of the FV and CMFE schemes that we obtain while employing the original elementwise solutions. Using instead the postprocessed values however completely eliminates this difference and on the finest mesh, all schemes give comparable results. This implies the highest efficiency for the FV-FE scheme because its lowest computational cost (for $r = 5$ and 64 time steps, the CPU times of the whole calculations were 202 seconds for the FV-FE scheme, 308 seconds for the FV scheme, 604 seconds for the CMFE scheme, and 804 seconds for the FV-NCFE scheme). The lowest discrete

Method \ r	0	1	2	3	4	5
CMFE	0.14313	0.07135	0.03564	0.01782	0.00891	0.00446
PCMFE	0.02608	0.00761	0.00259	0.00110	0.00053	0.00026
FV–FE	0.03961	0.01345	0.00537	0.00238	0.00111	0.00054
FV–NCFE	0.01990	0.00680	0.00293	0.00143	0.00072	0.00036

Table 3.12: Discrete $L^\infty(0, T; L^2(\Omega))$ relative errors, problem (3.28), coefficients (3.30), mesh D

Method \ r	0	1	2	3	4	5
CMFE	0.00821	0.00389	0.00199	0.00102	0.00052	0.00027
FV–FE	0.13895	0.04848	0.01713	0.00619	0.00230	0.00089
FV–NCFE	0.03122	0.01210	0.00475	0.00197	0.00087	0.00040

Table 3.13: Discrete $L^\infty(0, T; L^2(\Omega))$ projection relative errors, problem (3.28), coefficients (3.30), mesh D

$L^\infty(0, T; L^2(\Omega))$ projection relative error is produced by the mixed finite element method and the highest by the finite volume method. The differences are important enough to persist to the efficiency graph. The experimental order of convergence for fine meshes is $O(h, \Delta t)$, a little bit better for coarser meshes.

Tables 3.12 and 3.13 give the discrete relative and projection relative errors for the coefficients (3.30), mesh D, and Robin boundary conditions on $x = 0$ and Dirichlet boundary conditions otherwise. In the present case, the FV–FE scheme gives worse results than the FV–NCFE scheme, which is in turn outperformed by the elementwise linear postprocessed solution of the CMFE scheme. In a similar manner, the projection error of the mixed finite element method is very low, in spite of the discontinuous coefficients and inhomogeneous and anisotropic diffusion tensor. It can be seen from Table 3.9 that the conditioning of the system matrices is in this case considerably increased. However, this increase in conditioning is much more important for the FV–FE and FV–NCFE schemes than for the CMFE scheme. We have purposely chosen the mesh D, where the system matrix of the CMFE scheme fails to be positive definite. The results from Section 3.5.2 indicate that this may considerably increase the CPU time of the CMFE scheme in comparison with the mesh C, where the CMFE system matrix would be positive definite. Nevertheless, the CMFE scheme shows to be superior in this case also in terms of efficiency, followed by the two other schemes, cf. Figure 3.9. These differences are more important for the $L^\infty(0, T; L^2(\Omega))$ projection error. For the sake of completeness, we indicate that for $r = 5$ and 64 time steps, the CPU times were 242 seconds for the FV–FE scheme, 1017 seconds for the CMFE scheme, and 1153 seconds for the FV–NCFE scheme.

To conclude, we mention that we have only considered the discretization of not convection-dominated problems on fixed grids. When the problem at hand approaches the hyperbolic case, the presented schemes reduce to a finite volume scheme stabilized by an upstream weighting. They then only differ by the definition of the control volumes. A comparative study of the influence of the definition of the control volumes is given in [97]. It follows from this study that the most efficient choice is represented by triangular control volumes, then by control volumes associated with vertices, and finally by control volumes associated with edges. Finally, for a precise and efficient solution of convection-dominated problems, either local mesh refinement or the use of higher-order schemes would be necessary.

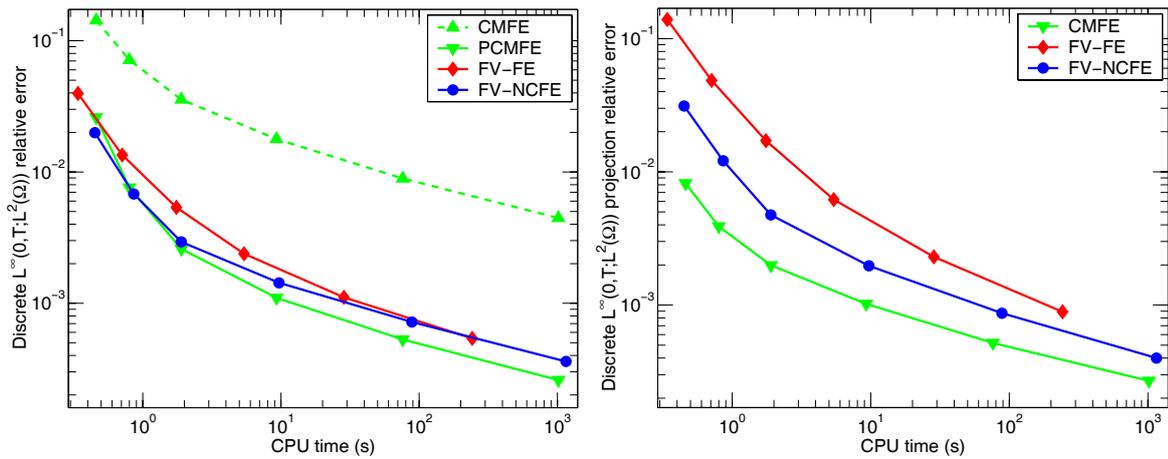


Figure 3.9: Efficiency comparison, problem (3.28), coefficients (3.30), mesh D

3.5.4 Conclusions

We have studied in this section the computational cost of the proposed condensation of the mixed finite element method for elliptic and (nonlinear) parabolic problems.

For elliptic diffusion problems in two space dimensions, the standard hybridization leads to systems for the number of edges unknowns with symmetric positive definite matrices with a 5-point stencil. In the proposed condensation, the number of unknowns is reduced to the number of elements (which is approximately 2/3 of the number of edges), but the system matrices are in general nonsymmetric, have a wider (about 13-point) stencil, and are positive definite only under a condition on the mesh and the diffusion tensor. This condition however allows for quite deformed triangles in the case of a piecewise constant scalar diffusion tensor. The CPU time speed-up for the test case was about 1.35. The finite volume reformulation of the mixed finite element method proposed and studied in [37, 121, 122] leads to symmetric matrices with the number of elements unknowns and a 4-point stencil. The matrices are positive definite for Delaunay triangulations and constant scalar diffusion tensors but indefinite otherwise. Hence the computational savings of the reformulation will be very probably more important than those of the condensation for Delaunay triangulations and constant scalar diffusion tensors. The situation should be much more favorable for the condensation when the mesh is not Delaunay or when the diffusion tensor is inhomogeneous and anisotropic. In three space dimensions, the finite volume reformulation is in general not possible, see [121]. In contrast, the condensation applies directly as in two space dimensions. Moreover, the number of unknowns is in this case only about 1/2 of that of the hybridization. Hence one can expect even more important computational savings than in the two-dimensional case.

We believe that the main importance of the proposed condensation lies in its application to mixed finite element discretizations of (nonlinear) parabolic convection–reaction–diffusion problems. The resulting matrices are still sparse, positive definite for a large class of meshes and diffusion tensors, nonsymmetric, and seem to be very well conditioned. Moreover, if the diffusion tensor is constant with respect to time, one can assemble and invert the local condensation matrices only once before the start of the calculation and then only work with the scalar unknowns as in the finite volume method, which still reduces the computational complexity. In two space dimensions, the number of unknowns is equal to approximately 2/3 of that of standard solution approaches in the mixed finite element method and to approximately 2/5 of

that of the upwind-mixed method. The CPU times necessary for the solution of the associated linear systems in the presented test cases were reduced by a factor 2 for parabolic reaction–diffusion problems. When convection is present, nonsymmetric matrices arise naturally also in the mixed and upwind-mixed schemes. The speed-up was in this case comprised between 3 and 4. The finite volume reformulation of the mixed finite element method is possible for parabolic reaction–diffusion problems, but leads in general to indefinite nonsymmetric systems with a limited gain in the terms of the computational cost, cf. [37, 122]. Hence the condensation seems to be much more attractive in this case. This is still emphasized by the fact that it can be very easily implemented into existing mixed finite element codes. Finally, the speed-up should be even more important in three space dimensions, where the number of unknowns of the condensation is about 1/2 of that of the mixed and 1/3 of that of the upwind-mixed schemes.

We have finally compared the condensed mixed finite element scheme with a finite volume and combined finite volume–finite element ones for nonlinear parabolic equations. For problems with discontinuous coefficients and inhomogeneous and anisotropic diffusion tensors, the mixed finite element method clearly gives better results. This, taking into account the gain in the CPU time due to the proposed condensation, makes it both more precise and efficient than the combined schemes. Also, for such problems, the combined finite volume–nonconforming/mixed-hybrid finite element scheme seems to be superior to the combined finite volume–finite element one, one of the possible reasons being that the latter scheme employs the arithmetic average of the heterogeneities. When no essential heterogeneities and discontinuities are present, the combined finite volume–finite element scheme may be the most efficient due to its low computational cost. Finally, when the diffusion operator is only a Laplacian and for Delaunay meshes, the finite volume method represents another cheap and efficient solution technique.

Chapitre 4

Mixed and nonconforming finite element methods on a fracture network

We investigate in this chapter the lowest-order Raviart–Thomas mixed finite element method for second-order elliptic problems posed over a system of intersecting two-dimensional polygons placed in three-dimensional Euclidean space. The domain is characteristic by the presence of intersection lines shared by three or more polygons. We first construct continuous and discrete function spaces ensuring the continuity of scalar functions and an appropriate continuity of the normal trace of vector functions across such intersection lines. We then propose a variant of the lowest-order Raviart–Thomas mixed finite element method for the given problem with the domain discretized into a triangular mesh and prove its well-posedness. We finally investigate the relation of the hybridization of the considered mixed finite element method to the piecewise linear nonconforming finite element method. We extend the results known in this direction onto networks of polygons, general diffusion tensors, and general boundary conditions. This enables us in particular to efficiently implement the mixed finite element method. We verify the theoretical results on a model problem with a known analytical solution and show the application of the proposed method to the simulation of underground water flow through a system of polygons representing a network of fractures that perturbs a rock massif.

4.1 Introduction

The motivation of this chapter is the need to simulate water flow through underground rock massifs. Such massifs are proposed as e.g. nuclear waste repositories and they are always disrupted by a system of geological faults, *fractures*. One of the possible modeling approaches is to approximate the fractures by a network of planar polygonal disks and to consider two-dimensional Darcy flow through such network, see e.g. [5, 24, 119]. This problem is mathematically a second-order elliptic problem posed over a system of intersecting two-dimensional polygons placed in three-dimensional Euclidean space. An example of such system is given in Figure 4.1. The system in this figure is already discretized into a triangular mesh (the colors represent various values of hydraulic conductivity associated with the elements). We can easily notice an essential property of a domain created by a system of polygons that is impossible in classical planar domains: there exist interelement edges in the triangulation which belong to three or more triangular elements.

We propose and investigate in this chapter a variant of the lowest-order Raviart–Thomas mixed finite element method [105] (cf. also [33, 108]) for systems of polygons. It turns out that the essential step is the definition of appropriate continuous and discrete function spaces: we have to ensure the continuity of the scalar primary unknown (pressure) across the intersection lines between polygons and an appropriate continuity of the normal trace of the flux of the primary unknown (the hydraulic conductivity tensor times the negative of the gradient of the pressure, i.e. the Darcy velocity) across these intersection lines. The well-posedness of the weak mixed formulation is then implied by the well-posedness of the weak primal formulation, which is easy to show. To demonstrate the existence and uniqueness of the mixed approximation, we define a global interpolation operator on the discrete velocity space and prove the commuting diagram property, which implies the discrete inf–sup (Babuška [16]–Brezzi [30]) condition.

We next investigate the relation of the hybridization of the lowest-order Raviart–Thomas mixed finite element method to the piecewise linear nonconforming finite element method. It is known that the matrices of these two methods coincide for an elliptic problem with an elementwise constant diffusion tensor and a homogeneous Dirichlet boundary condition, see [15] or a detailed study given in [38]. We extend these results onto systems of polygons, nonconstant diffusion tensors, and inhomogeneous mixed Dirichlet/Neumann boundary conditions. The implementation of the considered mixed finite element method via the nonconforming method is on the one hand very efficient and on the other hand, since a polygonal domain is a trivial instance of a system of polygons, is naturally valid also for standard planar domains. Such implementation in particular avoids the inverting of local matrices (cf. [33, Section V.1.2]), usually used when the relation with the nonconforming method is not known. Recall that inverting of local matrices is a potential source of significant numerical errors, cf. [74].

The outline of this chapter is as follows. In Section 4.2 we formulate the second-order elliptic problem on a system of polygons and in Section 4.3 we define continuous and discrete function spaces on such system. We state the weak primal formulation and the nonconforming finite element approximation in Section 4.4. Section 4.5 is devoted to the lowest-order Raviart–Thomas mixed finite element method: we state and show the existence and uniqueness of the weak mixed solution and of the mixed approximation, introduce the hybridization of the mixed approximation, and give error estimates. Finally, in Section 4.6 we investigate the relation between mixed and nonconforming methods and in the first part of Section 4.7 we present the results of a numerical experiment on a model problem with a known analytical solution. We refer to the second part of Section 4.7 for the description of the application of the proposed method to the simulation of fracture flow and for a comparison with other methods.

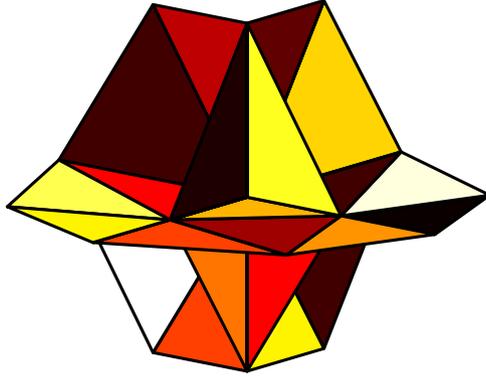


Figure 4.1: Example of a simple system of polygons (discretized into a triangular mesh)

4.2 Second-order elliptic problem on a system of polygons

We define the system of polygons \mathcal{S} and the second-order elliptic problem on this system in this section. We set

$$\mathcal{S} := \left\{ \bigcup_{\ell \in L} \overline{\alpha_\ell} \setminus \partial\mathcal{S} \right\}, \quad (4.1)$$

where α_ℓ is an open two-dimensional polygon placed in three-dimensional space and L is the index set of polygons. We suppose that the closures $\overline{\alpha_\ell}$ of polygons are all connected into the system; the connection is only possible through an edge, not through a point. For the purpose of the mathematical description, we require that $\alpha_i \cap \alpha_j = \emptyset$ if $i \neq j$ and that $\overline{\alpha_i} \cap \overline{\alpha_j}$, $i \neq j$, is either an edge or a point or an empty set. In order to fulfill this property it is enough to divide each polygon from a general system of polygons as that of Figure 4.1 into subpolygons along each intersection line that it contains. Finally, $\partial\mathcal{S}$ is the set of those boundary points of α_ℓ , $\ell \in L$, which do not create the connection with other polygons. We suppose that there is a two-dimensional orthogonal coordinate system given in each polygon. The system of the model problem from Section 4.7 viewed in Figure 4.3 below may serve as an example. In this case \mathcal{S} consists of three rectangles, denoted as four polygons α_1 – α_4 , and $\partial\mathcal{S}$ consists of twelve edges Γ_1 – Γ_{12} .

We seek p (a scalar function in each $\overline{\alpha_\ell}$) and \mathbf{u} (a two-dimensional vector in each $\overline{\alpha_\ell}$) which are the solutions of the problem

$$\mathbf{u} = -\mathbf{K}(\nabla p + \nabla z) \quad \text{in } \alpha_\ell, \ell \in L, \quad (4.2a)$$

$$\nabla \cdot \mathbf{u} = q \quad \text{in } \alpha_\ell, \ell \in L, \quad (4.2b)$$

$$p = p_D \quad \text{on } \Gamma_D, \quad \mathbf{u} \cdot \mathbf{n} = u_N \quad \text{on } \Gamma_N, \quad (4.2c)$$

where all the variables are expressed in local coordinates of the appropriate α_ℓ and also the differentiation is always done with respect to these local coordinates. In the context of groundwater flow the variable p denotes the pressure head, $p = \tilde{p}/\rho g$, where \tilde{p} is the fluid pressure, g is the gravitational acceleration constant and ρ is the fluid density, \mathbf{u} is the flow velocity, q represents stationary sources or sinks, z is the elevation, i.e. the upward vertical three-dimensional coordinate, and \mathbf{K} is the tensor of hydraulic conductivity. The equation (4.2a) is then the Darcy law, (4.2b) is the mass balance equation, and (4.2c) prescribes Dirichlet and Neumann boundary conditions. We suppose that $\Gamma_D \cap \Gamma_N = \emptyset$, $\overline{\Gamma_D} \cup \overline{\Gamma_N} = \partial\mathcal{S}$, and that the measure of Γ_D is nonzero. Note that \mathbf{n} in (4.2c) is the unit outward normal vector of the appropriate α_ℓ .

Let f be an edge such that there exist polygons α_i and α_j , $i \neq j$, such that $f = \overline{\alpha_i} \cap \overline{\alpha_j}$. We denote the set of such edges by \mathcal{E}^{int} and the set of all $i \in L$ such that $f \subset \partial\alpha_i$ by I_f . The system (4.2a)–(4.2c) is completed by requiring

$$p|_{\overline{\alpha_i}} = p|_{\overline{\alpha_j}} \quad \text{on } f \quad \forall f \in \mathcal{E}^{\text{int}}, \forall i, j \in I_f, \quad (4.3a)$$

$$\sum_{i \in I_f} \mathbf{u}|_{\overline{\alpha_i}} \cdot \mathbf{n}_{f, \alpha_i} = 0 \quad \text{on } f \quad \forall f \in \mathcal{E}^{\text{int}}, \quad (4.3b)$$

where \mathbf{n}_{f, α_i} is the unit outward normal vector of the edge f with respect to the polygon α_i . The equations (4.3a)–(4.3b) express the continuity of p across the interpolygon boundaries and the mass balance of \mathbf{u} across these boundaries (what is the outflow from one polygon has to be the inflow into the neighboring ones). We finally suppose that the second rank tensor \mathbf{K} is symmetric and uniformly positive definite in each α_ℓ , i.e.

$$\mathbf{K}(\mathbf{x})\boldsymbol{\eta} \cdot \boldsymbol{\eta} \geq c_{\mathbf{K}}\boldsymbol{\eta} \cdot \boldsymbol{\eta}, \quad c_{\mathbf{K}} > 0 \quad (4.4)$$

for any $\boldsymbol{\eta} \in \mathbb{R}^2$ and almost all $\mathbf{x} \in \alpha_\ell$, for all $\ell \in L$.

4.3 Function spaces for nonconforming and mixed finite elements

We give in this section the definitions of function spaces used in the sequel. We will use the spaces $H^1(\alpha_\ell)$ and $\mathbf{H}(\text{div}, \alpha_\ell)$ on separate polygons with certain matching conditions on the interpolygon boundaries in order to define the spaces $H^1(\mathcal{S})$ and $\mathbf{H}(\text{div}, \mathcal{S})$ on the whole system \mathcal{S} . We introduce also the discrete counterparts of these spaces.

4.3.1 Continuous function spaces

We use the product of the spaces L^p , $1 \leq p \leq \infty$, on separate polygons in order to define the spaces $L^p(\mathcal{S})$ and $\mathbf{L}^p(\mathcal{S})$ on the system \mathcal{S} ,

$$L^p(\mathcal{S}) := \prod_{\ell \in L} L^p(\alpha_\ell), \quad \mathbf{L}^p(\mathcal{S}) := L^p(\mathcal{S}) \times L^p(\mathcal{S}). \quad (4.5)$$

For each polygon α_ℓ , we denote by $H^1(\alpha_\ell)$ the Sobolev space of scalar functions with square-integrable weak derivatives, $H^1(\alpha_\ell) = \{\varphi \in L^2(\alpha_\ell); \nabla\varphi \in \mathbf{L}^2(\alpha_\ell)\}$. We define $H^1(\mathcal{S})$ as the space of functions whose restrictions to each α_ℓ are from $H^1(\alpha_\ell)$ and that coincide on the interpolygon boundaries in the sense of traces,

$$H^1(\mathcal{S}) := \left\{ v \in L^2(\mathcal{S}); v|_{\alpha_\ell} \in H^1(\alpha_\ell) \quad \forall \ell \in L, \right. \\ \left. (v|_{\alpha_i})|_f = (v|_{\alpha_j})|_f \quad \forall f \in \mathcal{E}^{\text{int}}, \forall i, j \in I_f \right\}. \quad (4.6)$$

We then have the space $H_D^1(\mathcal{S})$ of the functions from $H^1(\mathcal{S})$ vanishing on Γ_D and the spaces $H^{\frac{1}{2}}(\partial\mathcal{S})$, $H^{-\frac{1}{2}}(\partial\mathcal{S})$, $H^{\frac{1}{2}}(\Gamma_D)$, and $H^{-\frac{1}{2}}(\Gamma_N)$ as in the standard planar case.

For each polygon α_ℓ , we denote by $\mathbf{H}(\text{div}, \alpha_\ell)$ the space of vector functions with square-integrable weak divergences, $\mathbf{H}(\text{div}, \alpha_\ell) = \{\mathbf{v} \in \mathbf{L}^2(\alpha_\ell); \nabla \cdot \mathbf{v} \in L^2(\alpha_\ell)\}$. We define $\mathbf{H}(\text{div}, \mathcal{S})$ as the space of functions whose restrictions to each α_ℓ are from $\mathbf{H}(\text{div}, \alpha_\ell)$ and whose sum of normal traces over all polygons sharing a given edge $f \in \mathcal{E}^{\text{int}}$ is zero in the appropriate sense,

$$\mathbf{H}(\text{div}, \mathcal{S}) := \left\{ \mathbf{v} \in \mathbf{L}^2(\mathcal{S}); \mathbf{v}|_{\alpha_\ell} \in \mathbf{H}(\text{div}, \alpha_\ell) \quad \forall \ell \in L, \sum_{i \in I_f} \langle \mathbf{v}|_{\alpha_i} \cdot \mathbf{n}_{\partial\alpha_i}, \varphi_i \rangle_{\partial\alpha_i} = 0 \right. \\ \left. \forall \varphi_i \in H_{\partial\alpha_i \setminus f}^1(\alpha_i), \varphi_i|_f = \varphi_j|_f \quad \forall i, j \in I_f, \forall f \in \mathcal{E}^{\text{int}} \right\}. \quad (4.7)$$

Finally, we denote

$$\mathbf{H}_{0,N}(\operatorname{div}, \mathcal{S}) := \{ \mathbf{v} \in \mathbf{H}(\operatorname{div}, \mathcal{S}); \langle \mathbf{v} \cdot \mathbf{n}, \varphi \rangle_{\partial \mathcal{S}} = 0 \quad \forall \varphi \in H_D^1(\mathcal{S}) \}$$

as the space of functions from $\mathbf{H}(\operatorname{div}, \mathcal{S})$ such that their normal trace on Γ_N is equal to zero in the appropriate sense.

We use $(\cdot, \cdot)_{0, \alpha_\ell}$ to denote the L^2 scalar product, $\|\cdot\|_{0, \alpha_\ell}$ to denote the associated L^2 norm, $\|\cdot\|_{1, \alpha_\ell}$ to denote the $H^1(\alpha_\ell)$ norm, and $\|\cdot\|_{\mathbf{H}(\operatorname{div}, \alpha_\ell)}$ to denote the $\mathbf{H}(\operatorname{div}, \alpha_\ell)$ norm given by $\|\mathbf{v}\|_{\mathbf{H}(\operatorname{div}, \alpha_\ell)}^2 = \|\mathbf{v}\|_{0, \alpha_\ell}^2 + \|\nabla \cdot \mathbf{v}\|_{0, \alpha_\ell}^2$. The bracket $\langle \mathbf{v} \cdot \mathbf{n}, \varphi \rangle_{\partial \mathcal{S}}$ denotes the duality pairing between $H^{-\frac{1}{2}}(\partial \mathcal{S})$ and $H^{\frac{1}{2}}(\partial \mathcal{S})$ and may be written formally as $\int_{\partial \mathcal{S}} \mathbf{v} \cdot \mathbf{n} \varphi \, d\gamma(\mathbf{x})$. The norms on the spaces defined by (4.5), (4.6), and (4.7) are given by

$$\|\cdot\|_{\cdot, \mathcal{S}}^2 := \sum_{\ell \in L} \|\cdot\|_{\cdot, \alpha_\ell}^2. \quad (4.8)$$

Remark 4.3.1. (Continuity across the interpolygon boundaries) *The definitions (4.6) and (4.7) express weakly the conditions (4.3a) and (4.3b). Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain and let \mathcal{S} be its polygonal partition. Then the definitions (4.6) and (4.7) coincide with the standard characterizations of the spaces $H^1(\Omega)$ and $\mathbf{H}(\operatorname{div}, \Omega)$ (cf. [33, Propositions III.1.1 and III.1.2] or [108, Theorem 1.3]).*

Throughout this chapter, we shall suppose that $\mathbf{K}_{ij} \in L^\infty(\mathcal{S})$, $q \in L_2(\mathcal{S})$, $p_D \in H^{\frac{1}{2}}(\Gamma_D)$, and $u_N \in H^{-\frac{1}{2}}(\Gamma_N)$.

4.3.2 Discrete function spaces

Let us suppose a triangulation \mathcal{T}_h of the system \mathcal{S} such that the boundary edges lie entirely either in Γ_D or in Γ_N . We set

$$M_{-1}^0(\mathcal{T}_h) := \{ \phi \in L^2(\mathcal{S}); \phi|_e \text{ is constant } \forall e \in \mathcal{T}_h \}.$$

We denote the set of all edges of \mathcal{T}_h by \mathcal{E}_h , the set of all edges of \mathcal{T}_h except those from Γ_D by $\mathcal{E}_{h,D}$, and the set of all interior edges of \mathcal{T}_h by $\mathcal{E}_h^{\text{int}}$. We set

$$M_{-1}^0(\mathcal{E}_{h,D}) := \{ \mu : \mathcal{E}_h \rightarrow \mathbb{R}; \mu|_f \text{ is constant } \forall f \in \mathcal{E}_h, \\ \mu|_f = 0 \quad \forall f \subset \Gamma_D \}.$$

For the nonconforming approximation, we set

$$X_0^1(\mathcal{E}_h) := \{ \varphi \in L^2(\mathcal{S}); \varphi|_e \text{ is linear } \forall e \in \mathcal{T}_h, \varphi \text{ is continuous in } Q_f, f \in \mathcal{E}_h^{\text{int}} \},$$

where Q_f is the midpoint of the edge f . The basis of $X_0^1(\mathcal{E}_h)$ is spanned by shape functions φ_f , $f \in \mathcal{E}_h$, such that $\varphi_f(Q_g) = \delta_{fg}$, $g \in \mathcal{E}_h$, δ being the Kronecker delta. A simple computation gives

$$\nabla \varphi_f|_e = \frac{|f|}{|e|} \mathbf{n}_f \quad e \in \mathcal{T}_h, f \subset \partial e, \quad (4.9)$$

where $|e|$ is the area of the element e , $|f|$ is the length of the edge f , and \mathbf{n}_f is the unit normal vector of the edge f , outward to e . We finally set

$$X_0^1(\mathcal{E}_{h,D}) := \{ \varphi \in X_0^1(\mathcal{E}_h); \varphi(Q_f) = 0 \quad \forall f \subset \Gamma_D \}.$$

For a given triangular element $e \in \mathcal{T}_h$, we define $\mathbf{RT}^0(e)$ as the space of linear vector functions with the basis \mathbf{v}_i^e , $i = 1, 2, 3$,

$$\mathbf{v}_i^e(\mathbf{x}) := \frac{1}{2|e|} \begin{pmatrix} x - x_i \\ y - y_i \end{pmatrix} \text{ if } \mathbf{x} = (x, y)^t \in e, \quad \mathbf{v}_i^e(\mathbf{x}) := \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ if } \mathbf{x} \notin e, \quad (4.10)$$

where $(x_i, y_i)^t$ are the coordinates of the i -th vertex of e . Note that $\mathbf{v}_i^e \cdot \mathbf{n}_f$ is constant over each edge $f \subset \partial e$. The Raviart–Thomas space $\mathbf{RT}_{-1}^0(\mathcal{T}_h)$ of elementwise linear vector functions without any continuity requirement is defined by

$$\mathbf{RT}_{-1}^0(\mathcal{T}_h) := \{ \mathbf{v} \in \mathbf{L}^2(\mathcal{S}); \mathbf{v}|_e \in \mathbf{RT}^0(e) \quad \forall e \in \mathcal{T}_h \}. \quad (4.11)$$

We set the space $\mathbf{RT}_0^0(\mathcal{T}_h)$ of functions ensuring the normal trace continuity as

$$\begin{aligned} \mathbf{RT}_0^0(\mathcal{T}_h) &:= \left\{ \mathbf{v} \in \mathbf{RT}_{-1}^0(\mathcal{T}_h); \sum_{e \in \mathcal{T}_h; f \subset \partial e} \mathbf{v}|_e \cdot \mathbf{n}_{f,e} = 0 \text{ on } f \quad \forall f \in \mathcal{E}_h^{\text{int}} \right\} \\ &= \mathbf{RT}_{-1}^0(\mathcal{T}_h) \cap \mathbf{H}(\text{div}, \mathcal{S}). \end{aligned} \quad (4.12)$$

To characterize the discrete functions with zero normal trace on Γ_N , we finally set

$$\mathbf{RT}_{0,N}^0(\mathcal{T}_h) := \{ \mathbf{v} \in \mathbf{RT}_0^0(\mathcal{T}_h); \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \Gamma_N \} = \mathbf{RT}_{-1}^0(\mathcal{T}_h) \cap \mathbf{H}_{0,N}(\text{div}, \mathcal{S}).$$

4.4 Nonconforming finite element method

We introduce in this section a weak primal solution of the problem (4.2a)–(4.3b). We next define its piecewise linear nonconforming finite element approximation.

4.4.1 Weak primal solution

Let $\tilde{p} \in H^1(\mathcal{S})$ be such that $\tilde{p} = p_D$ on Γ_D in the sense of traces. We then define:

Definition 4.4.1. (Weak primal solution) *As the weak primal solution of the problem (4.2a)–(4.3b), we understand a function $p = p_0 + \tilde{p}$, $p_0 \in H_D^1(\mathcal{S})$, satisfying*

$$\begin{aligned} (\mathbf{K}\nabla p_0, \nabla \varphi)_{0,\mathcal{S}} &= (q, \varphi)_{0,\mathcal{S}} - \langle u_N, \varphi \rangle_{\partial \mathcal{S}} - (\mathbf{K}\nabla z, \nabla \varphi)_{0,\mathcal{S}} \\ &\quad - (\mathbf{K}\nabla \tilde{p}, \nabla \varphi)_{0,\mathcal{S}} \quad \forall \varphi \in H_D^1(\mathcal{S}). \end{aligned} \quad (4.13)$$

Existence and uniqueness of the weak primal solution follow from (4.4) and from the definition of the norms on \mathcal{S} given by (4.8) using the Lax–Milgram lemma.

4.4.2 Nonconforming finite element approximation

We define:

Definition 4.4.2. (Nonconforming finite element approximation) *As the piecewise linear nonconforming finite element approximation of the problem (4.13), we understand a function $p_h = p_{0,h} + \tilde{p}$, $p_{0,h} \in X_0^1(\mathcal{E}_{h,D})$, satisfying*

$$\begin{aligned} \sum_{e \in \mathcal{T}_h} (\mathbf{K}\nabla p_{0,h}, \nabla \varphi_h)_{0,e} &= \sum_{e \in \mathcal{T}_h} \{ (q, \varphi_h)_{0,e} - \langle u_N, \varphi_h \rangle_{\partial e \cap \partial \mathcal{S}} - (\mathbf{K}\nabla z, \nabla \varphi_h)_{0,e} \\ &\quad - (\mathbf{K}\nabla \tilde{p}, \nabla \varphi_h)_{0,e} \} \quad \forall \varphi_h \in X_0^1(\mathcal{E}_{h,D}). \end{aligned} \quad (4.14)$$

Existence and uniqueness of the nonconforming approximation follow by the same arguments as above.

4.5 Raviart–Thomas mixed finite element method

We first define in this section a weak mixed solution of the problem (4.2a)–(4.3b) and show its existence and uniqueness. We then study its lowest-order Raviart–Thomas mixed finite element approximation. We finally introduce its hybridization.

4.5.1 Weak mixed solution

Let $\tilde{\mathbf{u}} \in \mathbf{H}(\text{div}, \mathcal{S})$ be such that $\tilde{\mathbf{u}} \cdot \mathbf{n} = u_N$ on Γ_N in the appropriate sense. We then define:

Definition 4.5.1. (Weak mixed solution) *As the weak mixed solution of the problem (4.2a)–(4.3b), we understand functions $\mathbf{u} = \mathbf{u}_0 + \tilde{\mathbf{u}}$, $\mathbf{u}_0 \in \mathbf{H}_{0,N}(\text{div}, \mathcal{S})$, and $p \in L^2(\mathcal{S})$ such that*

$$\begin{aligned} (\mathbf{K}^{-1}\mathbf{u}_0, \mathbf{v})_{0,\mathcal{S}} - (\nabla \cdot \mathbf{v}, p)_{0,\mathcal{S}} &= -\langle \mathbf{v} \cdot \mathbf{n}, p_D \rangle_{\partial\mathcal{S}} + (\nabla \cdot \mathbf{v}, z)_{0,\mathcal{S}} \\ -\langle \mathbf{v} \cdot \mathbf{n}, z \rangle_{\partial\mathcal{S}} - (\mathbf{K}^{-1}\tilde{\mathbf{u}}, \mathbf{v})_{0,\mathcal{S}} &\quad \forall \mathbf{v} \in \mathbf{H}_{0,N}(\text{div}, \mathcal{S}), \end{aligned} \quad (4.15a)$$

$$-(\nabla \cdot \mathbf{u}_0, \phi)_{0,\mathcal{S}} = -(q, \phi)_{0,\mathcal{S}} + (\nabla \cdot \tilde{\mathbf{u}}, \phi)_{0,\mathcal{S}} \quad \forall \phi \in L^2(\mathcal{S}). \quad (4.15b)$$

Theorem 4.5.2. (Existence and uniqueness of the weak mixed solution) *The problem (4.15a)–(4.15b) has a unique solution.*

PROOF:

The coercivity of the bilinear form $(\mathbf{K}^{-1}\mathbf{u}, \mathbf{v})_{0,\mathcal{S}}$, $\mathbf{u}, \mathbf{v} \in \mathbf{H}_{0,N}(\text{div}, \mathcal{S})$, on the space $\mathbf{W} = \{\mathbf{v} \in \mathbf{H}_{0,N}(\text{div}, \mathcal{S}); (\nabla \cdot \mathbf{v}, \phi)_{0,\mathcal{S}} = 0 \quad \forall \phi \in L^2(\mathcal{S})\}$ is the consequence of the uniform positive definiteness of the tensor \mathbf{K} on each α_ℓ given by (4.4). We next show that for all $q \in L^2(\mathcal{S})$ there exists $\mathbf{v} \in \mathbf{H}_{0,N}(\text{div}, \mathcal{S})$ such that $(\nabla \cdot \mathbf{v}, \phi)_{0,\mathcal{S}} = (q, \phi)_{0,\mathcal{S}}$ for all $\phi \in L^2(\mathcal{S})$. This will guarantee that the divergence operator from $\mathbf{H}_{0,N}(\text{div}, \mathcal{S})$ to $L^2(\mathcal{S})$ is surjective (and hence the inf–sup condition). To show this, consider for given $q \in L^2(\mathcal{S})$ the problem of finding $p \in H_D^1(\mathcal{S})$ such that

$$(\nabla p, \nabla \varphi)_{0,\mathcal{S}} = (q, \varphi)_{0,\mathcal{S}} \quad \forall \varphi \in H_D^1(\mathcal{S}). \quad (4.16)$$

The existence and uniqueness of such p follow by the well-posedness of the weak primal formulation given in Section 4.4.1. We shall pose $\mathbf{v} = -\nabla p$. To justify such choice, we have to show that $\nabla p \in \mathbf{H}_{0,N}(\text{div}, \mathcal{S})$ and that $-\nabla \cdot \nabla p = q$ in the appropriate sense. The second assertion is a simple consequence of (4.16), considering $\varphi \in H_0^1(\alpha_\ell)$, $\ell \in L$, as test functions in (4.16). We now proceed to show the first assertion. Let us consider an edge $f \in \mathcal{E}^{\text{int}}$. We take $\varphi \in H_D^1(\mathcal{S})$ such that φ only has as a support the polygons sharing the edge f and such that φ is zero on $\partial\alpha_i \setminus f$ for all $i \in I_f$ in the sense of traces. The second assertion gives $\nabla p|_{\alpha_\ell} \in \mathbf{H}(\text{div}, \alpha_\ell)$, $\ell \in L$, and $-\sum_{i \in I_f} (\nabla \cdot \nabla p, \varphi)_{0,\alpha_i} = (q, \varphi)_{0,\mathcal{S}}$. Hence, using the Green theorem on each polygon in (4.16) with the considered φ as the test function,

$$\begin{aligned} 0 &= \sum_{i \in I_f} (\nabla p, \nabla \varphi)_{0,\alpha_i} - (q, \varphi)_{0,\mathcal{S}} = \sum_{i \in I_f} \langle \nabla p|_{\alpha_i} \cdot \mathbf{n}_{\partial\alpha_i}, \varphi \rangle_{\partial\alpha_i} \\ &\quad - \sum_{i \in I_f} (\nabla \cdot \nabla p, \varphi)_{0,\alpha_i} - (q, \varphi)_{0,\mathcal{S}} = \sum_{i \in I_f} \langle \nabla p|_{\alpha_i} \cdot \mathbf{n}_{\partial\alpha_i}, \varphi \rangle_{\partial\alpha_i}, \end{aligned}$$

which by the fact that $\varphi \in H^1(\mathcal{S})$ implies that $\nabla p \in \mathbf{H}(\text{div}, \mathcal{S})$, cf. the definition (4.7). Finally, $\nabla p \in \mathbf{H}_{0,N}(\text{div}, \mathcal{S})$ follows by the above technique applied to (4.16). The existence and uniqueness of the weak mixed solution follow by [33, Theorem II.1.1] or [108, Theorem 10.1]. \square

4.5.2 Properties of the discrete velocity space

We begin with the space $\mathbf{RT}^0(e)$ for a given $e \in \mathcal{T}_h$. Its basis is given by (4.10). The dual basis to this basis is given by the functionals N_j^e , $j = 1, 2, 3$, where

$$N_j^e(\mathbf{u}) = \int_{f_j^e} \mathbf{u} \cdot \mathbf{n}_{\partial e} \, d\gamma(\mathbf{x}) \quad \mathbf{u} \in \mathbf{RT}^0(e).$$

Each N_j^e expresses the flux of \mathbf{u} through one edge f_j^e of e . The local interpolation operator is then given by

$$\pi_e(\mathbf{u}) = \sum_{i=1}^3 N_i^e(\mathbf{u}) \mathbf{v}_i^e \quad \mathbf{u} \in (H^1(e))^2. \quad (4.17)$$

We now turn to the problem of finding the basis and the dual basis of $\mathbf{RT}_0^0(\mathcal{T}_h)$. Let us consider $\mathbf{u} \in \mathbf{RT}_0^0(\mathcal{T}_h)$. We set $\mathcal{N}_h = \{N_1, N_2, \dots, N_{|\mathcal{N}_h|}\}$, where for each boundary edge f such that $f \subset \partial e$, we define one functional N_f by

$$N_f(\mathbf{u}) := \int_f \mathbf{u}|_e \cdot \mathbf{n}_{\partial e} \, d\gamma(\mathbf{x}),$$

and for each interior edge f shared by the elements $e_1, e_2, \dots, e_{|I_f|}$, we define $|I_f| - 1$ functionals by

$$N_{f,j}(\mathbf{u}) := \frac{1}{|I_f|} \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} \, d\gamma(\mathbf{x}) - \frac{1}{|I_f|} \int_f \mathbf{u}|_{e_{j+1}} \cdot \mathbf{n}_{\partial e_{j+1}} \, d\gamma(\mathbf{x}), \quad j = 1, \dots, |I_f| - 1.$$

We use the same denotation I_f for the index set of polygons sharing a given edge $f \in \mathcal{E}^{\text{int}}$ in the continuous case and for the index set of elements sharing a given edge $f \in \mathcal{E}_h^{\text{int}}$ in the discrete case. We have the following lemma:

Lemma 4.5.3. (Basis of the dual space to $\mathbf{RT}_0^0(\mathcal{T}_h)$) \mathcal{N}_h is a basis of the dual space to $\mathbf{RT}_0^0(\mathcal{T}_h)$.

PROOF:

To prove the lemma it suffices to show that for all $\mathbf{u} \in \mathbf{RT}_0^0(\mathcal{T}_h)$, from $N_j(\mathbf{u}) = 0 \, \forall j = 1, \dots, |\mathcal{N}_h|$, it follows that $\mathbf{u} = 0$. Let us suppose that $N_j(\mathbf{u}) = 0 \, \forall j = 1, \dots, |\mathcal{N}_h|$. From the definition of the functionals N_f on boundary edges, we have $\int_f \mathbf{u}|_e \cdot \mathbf{n}_{\partial e} \, d\gamma(\mathbf{x}) = 0$ for all boundary edges f . Using the definition of the functionals $N_{f,j}$ on interior edges, we have $\int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} \, d\gamma(\mathbf{x}) = \int_f \mathbf{u}|_{e_j} \cdot \mathbf{n}_{\partial e_j} \, d\gamma(\mathbf{x})$ for all $j = 2, \dots, |I_f|$. Considering the equality $\sum_{i \in I_f} \int_f \mathbf{u}|_{e_i} \cdot \mathbf{n}_{\partial e_i} \, d\gamma(\mathbf{x}) = 0$ characterizing the continuity of the normal trace of the functions from $\mathbf{RT}_0^0(\mathcal{T}_h)$, cf. the definition (4.12), we come to $\int_f \mathbf{u}|_e \cdot \mathbf{n}_{\partial e} \, d\gamma(\mathbf{x}) = 0$ for all $f \in \mathcal{E}_h$ and all $e, f \subset \partial e$. Since $\mathbf{RT}_0^0(\mathcal{T}_h) \subset \mathbf{RT}_{-1}^0(\mathcal{T}_h)$, $\mathbf{u} = 0$ follows. \square

We set $\mathcal{V}_h = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{|\mathcal{N}_h|}\}$, the basis of $\mathbf{RT}_0^0(\mathcal{T}_h)$, in the following way: we define one basis function \mathbf{v}_f by $\mathbf{v}_f := \mathbf{v}_f^e$ for each boundary edge f . Here \mathbf{v}_f^e is the local basis function associated with the element e and its edge f . For each interior edge f shared by the elements $e_1, e_2, \dots, e_{|I_f|}$, we define $|I_f| - 1$ basis functions by

$$\mathbf{v}_{f,i} := \sum_{k=1, k \neq i+1}^{|I_f|} \mathbf{v}_f^{e_k} - (|I_f| - 1) \mathbf{v}_f^{e_{i+1}}, \quad i = 1, \dots, |I_f| - 1.$$

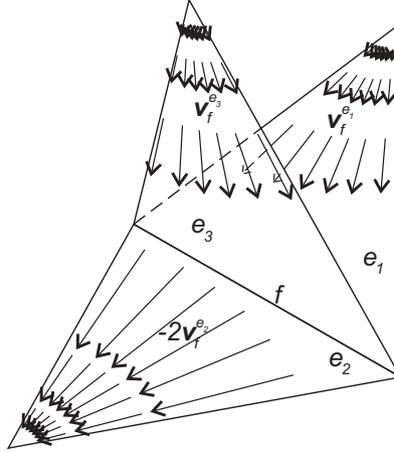


Figure 4.2: Velocity basis function for three elements sharing the same edge

Note that by the definition (4.12) of $\mathbf{RT}_0^0(\mathcal{T}_h)$, there is one condition imposed on each interior edge, so that the number of basis functions of $\mathbf{RT}_{-1}^0(\mathcal{T}_h)$ is decreased by one on each interior edge to obtain the appropriate continuity of the normal trace. When $|I_f| = 2$, we have the classical basis function. An example of one of the two basis functions for three elements with the same edge is given in Figure 4.2. We have the following lemma:

Lemma 4.5.4. (Duality) \mathcal{V}_h is the dual basis to \mathcal{N}_h .

PROOF:

We have to show that $N_j(\mathbf{v}_i) = \delta_{ij}$, $i, j = 1, \dots, |\mathcal{N}_h|$. We have from the definition of the basis functions of $\mathbf{RT}^0(e)$ that $N_f(\mathbf{v}_f) = 1$ for all boundary edges f , and simply $N_f(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathcal{V}_h$, $\mathbf{v} \neq \mathbf{v}_f$. Concerning the interior edges, we easily come to $N_{f,j}(\mathbf{v}_g) = 0$ for all $j = 1, \dots, |I_f| - 1$, f an interior edge, g a boundary edge, and to $N_{f,j}(\mathbf{v}_{g,i}) = 0$ for all $j = 1, \dots, |I_f| - 1$, $i = 1, \dots, |I_g| - 1$, f an interior edge, g another interior edge. We have

$$N_{f,j}(\mathbf{v}_{f,i}) = \frac{1}{|I_f|} \int_f \mathbf{v}_f^{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) - \frac{1}{|I_f|} \int_f \mathbf{v}_f^{e_{j+1}} \cdot \mathbf{n}_{\partial e_{j+1}} d\gamma(\mathbf{x}) = \frac{1}{|I_f|} - \frac{1}{|I_f|} = 0$$

for $i \neq j$ and

$$\begin{aligned} N_{f,i}(\mathbf{v}_{f,i}) &= \frac{1}{|I_f|} \int_f \mathbf{v}_f^{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) - \frac{1}{|I_f|} \int_f -(|I_f| - 1) \mathbf{v}_f^{e_{i+1}} \cdot \mathbf{n}_{\partial e_{i+1}} d\gamma(\mathbf{x}) \\ &= \frac{1}{|I_f|} + \frac{1}{|I_f|} (|I_f| - 1) = 1 \end{aligned}$$

for $i = 1, \dots, |I_f| - 1$, f an interior edge. Thus the proof is completed. \square

We are now ready to define the global interpolation operator. We introduce first a space smoother than $\mathbf{H}(\text{div}, \mathcal{S})$,

$$\begin{aligned} \mathbf{H}(\text{grad}, \mathcal{S}) &:= \left\{ v \in \mathbf{L}^2(\mathcal{S}); \mathbf{v}|_{\alpha_\ell} \in (H^1(\alpha_\ell))^2 \quad \forall \ell \in L, \right. \\ &\quad \left. \sum_{i \in I_f} \mathbf{v}|_{\alpha_i} \cdot \mathbf{n}_{f, \alpha_i} = 0 \text{ on } f \quad \forall f \in \mathcal{E}^{\text{int}} \right\}. \end{aligned} \quad (4.18)$$

We then define the global interpolation operator π_h by

$$\pi_h(\mathbf{u}) := \sum_{i=1}^{|\mathcal{N}_h|} N_i(\mathbf{u}) \mathbf{v}_i \quad \mathbf{u} \in \mathbf{H}(\text{grad}, \mathcal{S}). \quad (4.19)$$

We have the following relation between π_e and π_h :

Lemma 4.5.5. (Equality between local and global interpolation operators) *The local and global interpolation operators defined by (4.17) and (4.19), respectively, equal on each element, i.e.*

$$\pi_h(\mathbf{u})|_e = \pi_e(\mathbf{u}|_e) \quad \forall e \in \mathcal{T}_h, \forall \mathbf{u} \in \mathbf{H}(\text{grad}, \mathcal{S}).$$

PROOF:

As the basis functions \mathbf{v}_i , $i = 1, \dots, |\mathcal{N}_h|$, of $\mathbf{RT}_0^0(\mathcal{T}_h)$ are combined from the local basis functions \mathbf{v}_j^e on each element, we only have to verify that the coefficients of \mathbf{v}_j^e are the same. For boundary edges, the coefficients for both local and global interpolation operators are equally given by $\int_f \mathbf{u}|_e \cdot \mathbf{n}_{\partial e} d\gamma(\mathbf{x})$. For an interior edge f , we have for the global interpolation operator

$$\begin{aligned} & \left\{ \sum_{i=1}^{|\mathcal{I}_f|-1} N_{f,i}(\mathbf{u}) \mathbf{v}_{f,i} \right\} \Big|_{e_j} = \left\{ \sum_{i=1}^{|\mathcal{I}_f|-1} \left(\frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) \right. \right. \\ & \left. \left. - \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_{i+1}} \cdot \mathbf{n}_{\partial e_{i+1}} d\gamma(\mathbf{x}) \right) \left(\sum_{k=1, k \neq i+1}^{|\mathcal{I}_f|} \mathbf{v}_f^{e_k} - (|\mathcal{I}_f| - 1) \mathbf{v}_f^{e_{i+1}} \right) \right\} \Big|_{e_j} \\ & = \sum_{i=1, i \neq j-1}^{|\mathcal{I}_f|-1} \left(\frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) - \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_{i+1}} \cdot \mathbf{n}_{\partial e_{i+1}} d\gamma(\mathbf{x}) \right) \mathbf{v}_f^{e_j} \\ & \quad - (1 - \delta_{j1}) \left(\frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) - \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_j} \cdot \mathbf{n}_{\partial e_j} d\gamma(\mathbf{x}) \right) (|\mathcal{I}_f| - 1) \mathbf{v}_f^{e_j} \end{aligned}$$

using the definition of $N_{f,i}$ and $\mathbf{v}_{f,i}$, $i = 1, \dots, |\mathcal{I}_f| - 1$, $j = 1, \dots, |\mathcal{I}_f|$. Considering now only the coefficients of $\mathbf{v}_f^{e_j}$, we come to

$$\begin{aligned} & \sum_{i=1}^{|\mathcal{I}_f|-1} \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) - \sum_{i=1}^{|\mathcal{I}_f|-1} \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_{i+1}} \cdot \mathbf{n}_{\partial e_{i+1}} d\gamma(\mathbf{x}) \\ & = \left((|\mathcal{I}_f| - 1) \frac{1}{|\mathcal{I}_f|} + \frac{1}{|\mathcal{I}_f|} \right) \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) = \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) \end{aligned}$$

for $j = 1$, using the normal trace continuity of \mathbf{u} , which is expressed by $\sum_{i=1}^{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_i} \cdot \mathbf{n}_{\partial e_i} d\gamma(\mathbf{x}) = 0$. Similarly,

$$\begin{aligned} & (|\mathcal{I}_f| - 2) \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) + \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) \\ & + \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_j} \cdot \mathbf{n}_{\partial e_j} d\gamma(\mathbf{x}) - (|\mathcal{I}_f| - 1) \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_1} \cdot \mathbf{n}_{\partial e_1} d\gamma(\mathbf{x}) \\ & + (|\mathcal{I}_f| - 1) \frac{1}{|\mathcal{I}_f|} \int_f \mathbf{u}|_{e_j} \cdot \mathbf{n}_{\partial e_j} d\gamma(\mathbf{x}) = \int_f \mathbf{u}|_{e_j} \cdot \mathbf{n}_{\partial e_j} d\gamma(\mathbf{x}) \end{aligned}$$

for $j \geq 2$, and thus the proof is completed. \square

We conclude this section by the following theorem:

Theorem 4.5.6. (Commuting diagram property) *The commuting diagram property holds, i.e.*

$$\begin{array}{ccc} \mathbf{H}(\text{grad}, \mathcal{S}) & \xrightarrow{\text{div}} & L^2(\mathcal{S}) \\ \downarrow \pi_h & & \downarrow P_h \\ \mathbf{RT}_0^0(\mathcal{T}_h) & \xrightarrow{\text{div}} & M_{-1}^0(\mathcal{T}_h) \end{array},$$

where π_h is the global interpolation operator defined by (4.19) and P_h is the $L^2(\mathcal{S})$ -orthogonal projection onto $M_{-1}^0(\mathcal{T}_h)$.

PROOF:

The proof is immediate using the previous lemma and the validity of the commuting diagram property for the local interpolation operator, see e.g. [33, Proposition III.3.7] or [103, Section 3.4.2]. \square

4.5.3 Mixed finite element approximation

We are ready to define the mixed approximation:

Definition 4.5.7. (Mixed finite element approximation) *As the lowest-order Raviart–Thomas mixed finite element approximation of the problem (4.15a)–(4.15b), we understand functions $\mathbf{u}_h = \mathbf{u}_{0,h} + \tilde{\mathbf{u}}$, $\mathbf{u}_{0,h} \in \mathbf{RT}_{0,N}^0(\mathcal{T}_h)$, and $p_h \in M_{-1}^0(\mathcal{T}_h)$ satisfying*

$$\begin{aligned} (\mathbf{K}^{-1}\mathbf{u}_{0,h}, \mathbf{v}_h)_{0,\mathcal{S}} - (\nabla \cdot \mathbf{v}_h, p_h)_{0,\mathcal{S}} &= -\langle \mathbf{v}_h \cdot \mathbf{n}, p_D \rangle_{\partial\mathcal{S}} + (\nabla \cdot \mathbf{v}_h, z)_{0,\mathcal{S}} \\ -\langle \mathbf{v}_h \cdot \mathbf{n}, z \rangle_{\partial\mathcal{S}} - (\mathbf{K}^{-1}\tilde{\mathbf{u}}, \mathbf{v}_h)_{0,\mathcal{S}} & \quad \forall \mathbf{v}_h \in \mathbf{RT}_{0,N}^0(\mathcal{T}_h), \end{aligned} \quad (4.20a)$$

$$-(\nabla \cdot \mathbf{u}_{0,h}, \phi_h)_{0,\mathcal{S}} = -(q, \phi_h)_{0,\mathcal{S}} + (\nabla \cdot \tilde{\mathbf{u}}, \phi_h)_{0,\mathcal{S}} \quad \forall \phi_h \in M_{-1}^0(\mathcal{T}_h). \quad (4.20b)$$

The commuting diagram property expressed by Theorem 4.5.6 implies the discrete inf–sup condition, which in turn ensures that the problem (4.20a)–(4.20b) has a unique solution.

4.5.4 Hybridization of the mixed approximation

We will now introduce the hybridization of the mixed approximation:

Definition 4.5.8. (Hybridization of the mixed approximation) *As the hybridization of the lowest-order Raviart–Thomas mixed finite element approximation of the problem (4.15a)–(4.15b), we understand functions $\mathbf{u}_h = \mathbf{u}_{0,h} + \tilde{\mathbf{u}}$, $\mathbf{u}_{0,h} \in \mathbf{RT}_{-1}^0(\mathcal{T}_h)$, $p_h \in M_{-1}^0(\mathcal{T}_h)$, and $\lambda_h \in M_{-1}^0(\mathcal{E}_{h,D})$ satisfying*

$$\begin{aligned} & \sum_{e \in \mathcal{T}_h} \{ (\mathbf{K}^{-1}\mathbf{u}_{0,h}, \mathbf{v}_h)_{0,e} - (\nabla \cdot \mathbf{v}_h, p_h)_{0,e} + \langle \mathbf{v}_h \cdot \mathbf{n}, \lambda_h \rangle_{\partial e} \} \\ = & \sum_{e \in \mathcal{T}_h} \{ -\langle \mathbf{v}_h \cdot \mathbf{n}, p_D \rangle_{\partial e \cap \Gamma_D} + (\nabla \cdot \mathbf{v}_h, z)_{0,e} - \langle \mathbf{v}_h \cdot \mathbf{n}, z \rangle_{\partial e} - (\mathbf{K}^{-1}\tilde{\mathbf{u}}, \mathbf{v}_h)_{0,e} \} \end{aligned} \quad (4.21a)$$

$$\forall \mathbf{v}_h \in \mathbf{RT}_{-1}^0(\mathcal{T}_h),$$

$$\begin{aligned} - \sum_{e \in \mathcal{T}_h} (\nabla \cdot \mathbf{u}_{0,h}, \phi_h)_{0,e} &= - \sum_{e \in \mathcal{T}_h} \{ (q, \phi_h)_{0,e} - (\nabla \cdot \tilde{\mathbf{u}}, \phi_h)_{0,e} \} \\ & \quad \forall \phi_h \in M_{-1}^0(\mathcal{T}_h), \end{aligned} \quad (4.21b)$$

$$\sum_{e \in \mathcal{T}_h} \langle \mathbf{u}_{0,h} \cdot \mathbf{n}, \mu_h \rangle_{\partial e} = 0 \quad \forall \mu_h \in M_{-1}^0(\mathcal{E}_{h,D}). \quad (4.21c)$$

It is immediate that if $\mathbf{v}_h \in \mathbf{RT}_{-1}^0(\mathcal{T}_h)$, then $\mathbf{v}_h \in \mathbf{RT}_{0,N}^0(\mathcal{T}_h)$ if and only if

$$\sum_{e \in \mathcal{T}_h} \langle \mathbf{v}_h \cdot \mathbf{n}, \lambda_h \rangle_{\partial e} = 0 \quad \forall \lambda_h \in M_{-1}^0(\mathcal{E}_{h,D}).$$

This ensures that the triple $\mathbf{u}_{0,h}, p_h, \lambda_h$ exists and is unique and that $\mathbf{u}_{0,h}$ and p_h are at the same time the unique solutions of (4.20a)–(4.20b). We summarize the previous developments in the following theorem:

Theorem 4.5.9. (Existence and uniqueness of the mixed-hybrid approximation)

The problem (4.21a)–(4.21c) has a unique solution.

4.5.5 Error estimates

We now give two error estimates, following from the classical interpolation theory. If the solution (\mathbf{u}, p) of (4.15a)–(4.15b) is smooth enough and if $(\mathbf{u}_h, p_h, \lambda_h)$ is the solution of (4.21a)–(4.21c), we have

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{H}(\text{div}, \mathcal{S})} + \|p - p_h\|_{0,\mathcal{S}} \leq Ch(\|p\|_{1,\mathcal{S}} + \|\mathbf{u}\|_{1,\mathcal{S}} + \|q\|_{1,\mathcal{S}}),$$

where the constant C does not depend on h (see [33, Proposition IV.1.2]).

Using the piecewise linear but nonconforming approximation $\tilde{\lambda}_h \in X_0^1(\mathcal{E}_h)$ given by the values of the Lagrange multiplier λ_h at the midpoints of the edges, we have (see [33, Theorem V.3.1])

$$\|p - \tilde{\lambda}_h\|_{0,\mathcal{S}} \leq Ch^2(\|p\|_{1,\mathcal{S}} + \|\mathbf{u}\|_{1,\mathcal{S}} + \|q\|_{1,\mathcal{S}}).$$

4.6 Relation between mixed and nonconforming methods

We study in this section the relation between the hybridization of the lowest-order Raviart–Thomas mixed finite element method and the nonconforming method. We extend the results of [38] onto systems of polygons, general diffusion tensors, and general boundary conditions. This also enables us to efficiently implement the mixed finite element method in the considered case.

4.6.1 Algebraic condensation of the mixed-hybrid approximation

Let us denote, for all $e \in \mathcal{T}_h$,

$$\mathbf{u}_{0,h}|_e = \begin{pmatrix} a_e + c_e x \\ b_e + c_e y \end{pmatrix}, \quad p_h|_e = p_e$$

and similarly, for all $f \in \mathcal{E}_h$,

$$\lambda_h|_f = \lambda_f.$$

We now follow the ideas of [38]. Let $e \in \mathcal{T}_h$ be fixed. Consider in (4.21b) a test function ϕ_h equal to 1 on e and zero otherwise. This gives $c_e = q_e/2 - \tilde{\mathbf{u}}_e/2$ with

$$q_e := \frac{\int_e q \, d\mathbf{x}}{|e|}, \quad \tilde{\mathbf{u}}_e := \frac{\int_e \nabla \cdot \tilde{\mathbf{u}} \, d\mathbf{x}}{|e|}. \quad (4.22)$$

Next consider in (4.21a) two test functions, $\mathbf{v}_h = (1, 0)^t$, $\mathbf{v}_h = (0, 1)^t$ on e and zero otherwise, whose divergence is apparently zero. This gives

$$\int_e \mathbf{K}^{-1} \mathbf{u}_{0,h} \, d\mathbf{x} + \int_{\partial e} \lambda_h \mathbf{n} \, d\gamma(\mathbf{x}) = \mathbf{r}_e$$

with

$$\mathbf{r}_e := - \int_{\partial e \cap \Gamma_D} p_D \mathbf{n} \, d\gamma(\mathbf{x}) - \int_{\partial e} z \mathbf{n} \, d\gamma(\mathbf{x}) - \int_e \mathbf{K}^{-1} \tilde{\mathbf{u}} \, d\mathbf{x}.$$

Let $\tilde{\lambda}_h \in X_0^1(\mathcal{E}_{h,D})$ be given by

$$\tilde{\lambda}_h := \sum_{f \in \mathcal{E}_h} \lambda_f \varphi_f.$$

Then using (4.9), we have

$$\int_{\partial e} \lambda_h \mathbf{n} \, d\gamma(\mathbf{x}) = \sum_{f \subset \partial e} \lambda_f |f| \mathbf{n}_f = |e| \sum_{f \in \partial e} \lambda_f \nabla \varphi_f|_e = |e| \nabla \tilde{\lambda}_h|_e.$$

Next,

$$\int_e \mathbf{K}^{-1} \mathbf{u}_{0,h} \, d\mathbf{x} = \int_e \mathbf{K}^{-1} \, d\mathbf{x} \begin{pmatrix} a_e \\ b_e \end{pmatrix} + c_e \int_e \mathbf{K}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \, d\mathbf{x}.$$

Let us denote

$$\mathbf{K}_e := \left(\frac{1}{|e|} \int_e \mathbf{K}^{-1} \, d\mathbf{x} \right)^{-1} \quad e \in \mathcal{T}_h. \quad (4.23)$$

Then the above equations give

$$\begin{pmatrix} a_e \\ b_e \end{pmatrix} + c_e \frac{\mathbf{K}_e}{|e|} \int_e \mathbf{K}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \, d\mathbf{x} + \mathbf{K}_e \nabla \tilde{\lambda}_h|_e = \mathbf{K}_e \frac{\mathbf{r}_e}{|e|}$$

and consequently

$$\begin{aligned} \mathbf{u}_{0,h}|_e &= -\mathbf{K}_e \nabla \tilde{\lambda}_h|_e + \left[\frac{q_e}{2} - \frac{\tilde{\mathbf{u}}_e}{2} \right] \left[\begin{pmatrix} x \\ y \end{pmatrix} \right. \\ &\quad \left. - \frac{\mathbf{K}_e}{|e|} \int_e \mathbf{K}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \, d\mathbf{x} \right] + \mathbf{K}_e \frac{\mathbf{r}_e}{|e|}. \end{aligned} \quad (4.24)$$

We finally substitute (4.24) into (4.21c). This gives the following system of linear equations with the only unknowns the Lagrange multipliers λ_h :

$$\begin{aligned} \sum_{e \in \mathcal{T}_h} (\mathbf{K}_e \nabla \tilde{\lambda}_h, \nabla \tilde{\mu}_h)_{0,e} &= \sum_{e \in \mathcal{T}_h} \left\langle \left\{ \left[\frac{q_e}{2} - \frac{\tilde{\mathbf{u}}_e}{2} \right] \left[\begin{pmatrix} x \\ y \end{pmatrix} \right] \right. \right. \\ &\quad \left. \left. - \frac{\mathbf{K}_e}{|e|} \int_e \mathbf{K}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \, d\mathbf{x} \right\} \cdot \mathbf{n}, \mu_h \right\rangle_{\partial e} \quad \forall \mu_h \in M_{-1}^0(\mathcal{E}_{h,D}), \end{aligned} \quad (4.25)$$

where $\tilde{\mu}_h \in X_0^1(\mathcal{E}_{h,D})$ is given by

$$\tilde{\mu}_h := \sum_{f \in \mathcal{E}_h} \mu_f \varphi_f.$$

The left-hand side of (4.25) follows by

$$\langle \mathbf{K}_e \nabla \tilde{\lambda}_h|_e \cdot \mathbf{n}, \mu_h \rangle_{\partial e} = \langle \mathbf{K}_e \nabla \tilde{\lambda}_h|_e \cdot \mathbf{n}, \tilde{\mu}_h \rangle_{\partial e} = (\mathbf{K}_e \nabla \tilde{\lambda}_h, \nabla \tilde{\mu}_h)_{0,e} \quad \forall e \in \mathcal{T}_h.$$

Here, we have used the fact that $\mathbf{K}_e \nabla \tilde{\lambda}_h|_e \cdot \mathbf{n} \mu_h$ is constant over each edge and hence its integral over this edge equals to that of $\mathbf{K}_e \nabla \tilde{\lambda}_h|_e \cdot \mathbf{n} \tilde{\mu}_h$, which is a linear function with the same value at the edge midpoint by the definition of $\tilde{\mu}_h$, and finally the Green theorem in e (notice that $\mathbf{K}_e \nabla \tilde{\lambda}_h|_e$ is a constant vector in e and hence its divergence is zero).

The system given by (4.25), in the sequel called (algebraically) *condensed mixed-hybrid method*, enables a very efficient implementation of the scheme (4.21a)–(4.21c). In particular, its system matrix is symmetric and positive definite and the number of unknowns equals to the number of interior and Neumann boundary edges; remark that this number does not increase with the number of triangles sharing the given edge. Moreover, this matrix is assembled directly and one thus can avoid the inverting of local matrices, which is necessary in the traditional static condensation approach (cf. [33, Section V.1.2]). It is pointed out in [74] that the inverting of local matrices is a potential source of significant numerical errors. Finally, note that the velocity $\mathbf{u}_{0,h} \in \mathbf{RT}_{0,N}^0(\mathcal{T}_h)$ is easily obtained from the knowledge of $\tilde{\lambda}_h$ by (4.24). It is easily seen that the system (4.25) is very close to that given by the nonconforming finite element approximation (4.14). We give detailed comments on the relation between these two systems in the next section.

4.6.2 Comparison of condensed mixed-hybrid and nonconforming methods

We consider in this section the detailed relation between the condensed mixed-hybrid finite element method given by (4.25) and the nonconforming finite element method given by (4.14). We consider the matrices of the problems and the different parts of the right-hand sides separately.

System matrix

It is easily seen from (4.25), (4.14), and (4.23) that the system matrix of the condensed mixed-hybrid method is the system matrix of the nonconforming method with a piecewise constant diffusion tensor, given as the inverse of the elementwise average of the inverse of the original one. In particular, for elementwise constant diffusion tensors, these matrices coincide, as it was already shown in [38]. Simply, the mixed-hybrid method employs the harmonic average of the hydraulic conductivity tensor, whereas the nonconforming method uses instead the arithmetic average.

Sources term

Using the simple trick of replacing μ_h by $\tilde{\mu}_h$ and the Green theorem in each $e \in \mathcal{T}_h$ as at the end of Section 4.6.1, we have for the sources term of the condensed mixed-hybrid method the expression

$$\sum_{e \in \mathcal{T}_h} (q_e, \tilde{\mu}_h)_{0,e} + \sum_{e \in \mathcal{T}_h} \frac{q_e}{2} \left(\left(\begin{array}{c} x_e \\ y_e \end{array} \right) - \frac{\mathbf{K}_e}{|e|} \int_e \mathbf{K}^{-1} \left(\begin{array}{c} x \\ y \end{array} \right) d\mathbf{x}, \nabla \tilde{\mu}_h \right)_{0,e},$$

where $(x_e, y_e)^t$ are the coordinates of the barycentre^t of the triangle e . In particular, if \mathbf{K} is elementwise constant, the second term of the above expression vanishes. Hence the essential difference with the source term of the nonconforming method is the employment of the elementwise average of q given by (4.22) rather than taking q directly.

Dirichlet boundary condition term

Let p_D be smooth enough and let us consider the usual approximation $\tilde{p} \approx \sum_{f \subset \Gamma_D} p_D(Q_f) \varphi_f$. Then the Dirichlet boundary condition term in the nonconforming method becomes

$$-\sum_{e \in \mathcal{T}_h} (\mathbf{K} \nabla \tilde{p}, \nabla \tilde{\mu}_h)_{0,e} \approx -\sum_{e \in \mathcal{T}_h} \left(\mathbf{K} \sum_{f \subset \partial e \cap \Gamma_D} p_D(Q_f) \frac{|f|}{|e|} \mathbf{n}_f, \nabla \tilde{\mu}_h \right)_{0,e},$$

where $\tilde{\mu}_h \in X_0^1(\mathcal{E}_{h,D})$ and where we have employed the relation (4.9). This is obviously equivalent, up to replacing \mathbf{K} by \mathbf{K}_e , to the expression for this term from the condensed mixed-hybrid method

$$-\sum_{e \in \mathcal{T}_h} \left(\frac{\mathbf{K}_e}{|e|} \int_{\partial e \cap \Gamma_D} p_D \mathbf{n} d\gamma(\mathbf{x}), \nabla \tilde{\mu}_h \right)_{0,e} \approx -\sum_{e \in \mathcal{T}_h} \left(\frac{\mathbf{K}_e}{|e|} \sum_{f \subset \partial e \cap \Gamma_D} p_D(Q_f) |f| \mathbf{n}_f, \nabla \tilde{\mu}_h \right)_{0,e}.$$

Neumann boundary condition term

Let us for simplicity consider just one edge f where the Neumann boundary condition is prescribed, i.e. $\Gamma_N = f$. Then the Neumann boundary condition term in the nonconforming method, with the usual approximation supposing that u_N is smooth enough and for the test function φ_f , is

$$-\int_f u_N \varphi_f d\gamma(\mathbf{x}) \approx -\int_f u_N(Q_f) \varphi_f d\gamma(\mathbf{x}) = -u_N(Q_f) |f|.$$

Recall that this term equals to zero for all other test functions φ_g , $g \in \mathcal{E}_{h,D}$, $g \neq f$.

Using the same techniques as in the above paragraphs, we can express the Neumann boundary condition term in the condensed mixed-hybrid method as

$$\begin{aligned} & -\sum_{e \in \mathcal{T}_h} (\tilde{\mathbf{u}}_e, \tilde{\mu}_h)_{0,e} - \sum_{e \in \mathcal{T}_h} \frac{\tilde{\mathbf{u}}_e}{2} \left(\begin{pmatrix} x_e \\ y_e \end{pmatrix} - \frac{\mathbf{K}_e}{|e|} \int_e \mathbf{K}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} d\mathbf{x}, \nabla \tilde{\mu}_h \right)_{0,e} \\ & - \sum_{e \in \mathcal{T}_h} \left\langle \left\{ \frac{\mathbf{K}_e}{|e|} \int_e \mathbf{K}^{-1} \tilde{\mathbf{u}} d\mathbf{x} \right\} \cdot \mathbf{n}, \mu_h \right\rangle_{\partial e}. \end{aligned}$$

Let \mathbf{K} be elementwise constant; then the second term of the above expression vanishes and its third term simplifies. Let $e \in \mathcal{T}_h$ be such that $f \subset \partial e$ and let us finally consider the usual approximation $\tilde{\mathbf{u}} \approx u_N(Q_f) |f| \mathbf{v}_f^e$, where $\mathbf{v}_f^e \in \mathbf{RT}^0(e)$ is the local velocity basis function associated with the element e and its edge f . Then this term is a priori nonzero only for e and for the three test functions φ_g , $g \subset \partial e$, and has the form

$$-\frac{u_N(Q_f) |f|}{|e|} \left(\int_e \nabla \cdot \mathbf{v}_f^e d\mathbf{x}, \varphi_g \right)_{0,e} - \frac{u_N(Q_f) |f|}{|e|} \left\langle \left\{ \int_e \mathbf{v}_f^e d\mathbf{x} \right\} \cdot \mathbf{n}, \varphi_g \right\rangle_{\partial e}.$$

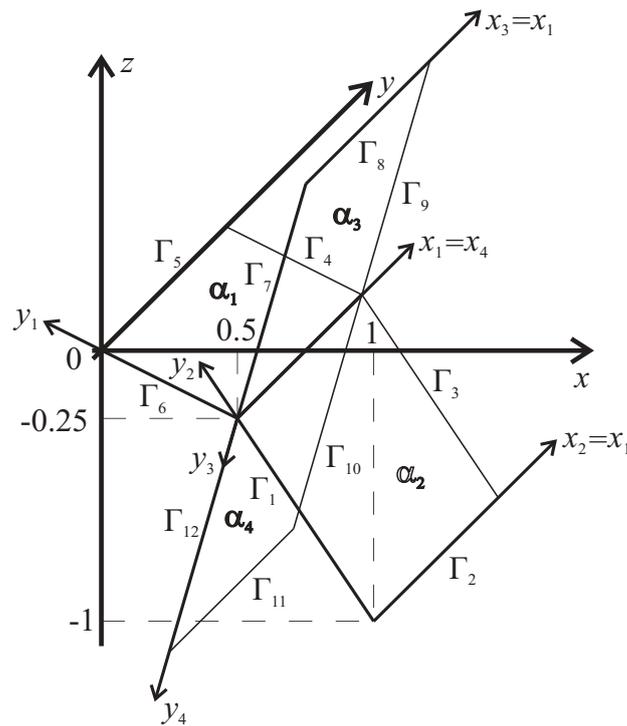
A simple computation gives

$$\int_e \nabla \cdot \mathbf{v}_f^e d\mathbf{x} = 1, \quad \int_e \mathbf{v}_f^e d\mathbf{x} = \frac{1}{2} \mathbf{w},$$

where $\mathbf{w} = (x_e, y_e)^t - (x_f, y_f)^t$ with $(x_f, y_f)^t$ being the coordinates of the vertex of e opposite to f . This finally gives for the Neumann boundary condition term in the condensed mixed-hybrid method, using simple geometrical properties of a triangle,

$$-\frac{u_N(Q_f) |f|}{3} - \frac{u_N(Q_f) |f| |g|}{2|e|} \mathbf{w} \cdot \mathbf{n}_g = -u_N(Q_f) |f| \delta_{f,g},$$

which coincides with the expression from the nonconforming method.

Figure 4.3: System \mathcal{S} for the model problems

Gravity term

Using that the gradient of z is piecewise constant, a development similar to that for the Dirichlet boundary condition gives that the expressions for the gravity term from the nonconforming and condensed mixed-hybrid methods differ just by the employment of \mathbf{K} , \mathbf{K}_e , respectively.

4.7 Numerical simulations

We present in this section the results of a numerical experiment on a model problem with a known analytical solution. We then describe the application of the proposed method to the simulation of fracture flow and compare it with other methods.

4.7.1 Model problem with a known analytical solution

We consider two simple model problems in this section. The first model problem corresponds to the system \mathcal{S} created by four rectangles as viewed in Figure 4.3. We verify on this problem the theoretical error estimates for the situation where the central edge is shared by four polygons. The second model problem is a simplification of the previous one, with just the rectangles α_1 and α_2 creating the system; there is no multiply shared edge in this case. Both model problems have the same known analytical solution in α_1 and α_2 . We consider the second model problem in order to investigate the changes of the approximation error implied by the presence of a multiply shared edge. All the computations presented in this section were done in double precision on a personal computer with machine precision being in power of 10^{-16} . The resulting systems of linear equations were solved by the preconditioned conjugate gradients method, cf. [73, 103, 111].

The first model problem is given by:

$$\begin{aligned} \mathcal{S} &= \bar{\alpha}_1 \cup \bar{\alpha}_2 \cup \bar{\alpha}_3 \cup \bar{\alpha}_4 \setminus \partial\mathcal{S}, \\ \mathbf{u} &= -(\nabla p + \nabla z) \quad \text{in } \alpha_i, i = 1, 2, 3, 4, \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \alpha_i, i = 1, 2, 3, 4, \\ p &= 0 \quad \text{on } \Gamma_1, \quad p = 0 \quad \text{on } \Gamma_2, \\ \mathbf{u} \cdot \mathbf{n} &= 0 \quad \text{on } \Gamma_3, \quad \mathbf{u} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_4, \\ p &= \sin\left(\frac{\pi x_1}{2X}\right) \sinh\left(\frac{\pi(A+B)}{2X}\right) + SA \quad \text{on } \Gamma_5, \quad p = Sy_1 \quad \text{on } \Gamma_6, \\ p &= 0 \quad \text{on } \Gamma_7, \quad p = 0 \quad \text{on } \Gamma_8, \\ \mathbf{u} \cdot \mathbf{n} &= 0 \quad \text{on } \Gamma_9, \quad \mathbf{u} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_{10}, \\ p &= \sin\left(\frac{\pi x_4}{2X}\right) \sinh\left(\frac{\pi(B+B)}{2X}\right) \quad \text{on } \Gamma_{11}, \quad p = 0 \quad \text{on } \Gamma_{12}, \end{aligned}$$

where $A = |\Gamma_4| = \sqrt{5}/4$, $X = |\Gamma_2| = 1$, $B = |\Gamma_3| = |\Gamma_9| = |\Gamma_{10}| = \sqrt{13}/4$, and $S = \partial z / \partial y_2 - \partial z / \partial y_1$. The geometry of this model problem is viewed in Figure 4.3. The exact solution can be found as

$$\begin{aligned} p|_{\alpha_1} &= \sin\left(\frac{\pi x_1}{2X}\right) \sinh\left(\frac{\pi(y_1+B)}{2X}\right) + Sy_1, \\ \mathbf{u}|_{\alpha_1} &= \left(-\frac{\pi}{2X} \cos\left(\frac{\pi x_1}{2X}\right) \sinh\left(\frac{\pi(y_1+B)}{2X}\right), \right. \\ &\quad \left. -\frac{\pi}{2X} \sin\left(\frac{\pi x_1}{2X}\right) \cosh\left(\frac{\pi(y_1+B)}{2X}\right) - S - \frac{\partial z}{\partial y_1} \right), \\ p|_{\alpha_2} &= \sin\left(\frac{\pi x_2}{2X}\right) \sinh\left(\frac{\pi y_2}{2X}\right), \\ \mathbf{u}|_{\alpha_2} &= \left(-\frac{\pi}{2X} \cos\left(\frac{\pi x_2}{2X}\right) \sinh\left(\frac{\pi y_2}{2X}\right), -\frac{\pi}{2X} \sin\left(\frac{\pi x_2}{2X}\right) \cosh\left(\frac{\pi y_2}{2X}\right) - \frac{\partial z}{\partial y_2} \right), \\ p|_{\alpha_3} &= \sin\left(\frac{\pi x_3}{2X}\right) \sinh\left(\frac{\pi y_3}{2X}\right), \\ \mathbf{u}|_{\alpha_3} &= \left(-\frac{\pi}{2X} \cos\left(\frac{\pi x_3}{2X}\right) \sinh\left(\frac{\pi y_3}{2X}\right), -\frac{\pi}{2X} \sin\left(\frac{\pi x_3}{2X}\right) \cosh\left(\frac{\pi y_3}{2X}\right) - \frac{\partial z}{\partial y_3} \right), \\ p|_{\alpha_4} &= \sin\left(\frac{\pi x_4}{2X}\right) \sinh\left(\frac{\pi(y_4+B)}{2X}\right), \\ \mathbf{u}|_{\alpha_4} &= \left(-\frac{\pi}{2X} \cos\left(\frac{\pi x_4}{2X}\right) \sinh\left(\frac{\pi(y_4+B)}{2X}\right), \right. \\ &\quad \left. -\frac{\pi}{2X} \sin\left(\frac{\pi x_4}{2X}\right) \cosh\left(\frac{\pi(y_4+B)}{2X}\right) - \frac{\partial z}{\partial y_4} \right). \end{aligned}$$

Note that the gradients of z in α_1 and α_2 are different. Hence the occurrence of the term S , which ensures the continuity of the normal trace of the velocity field. Table 4.1 gives the approximation errors in the first rectangle α_1 . The system \mathcal{S} is discretized into $4 \times 2N^2$ regular triangular elements, $h \approx 1/N$. There is the expected $O(h)$ convergence of \mathbf{u}_h , $O(h)$ convergence of the elementwise constant p_h , and $O(h^2)$ convergence of the piecewise linear but discontinuous $\tilde{\lambda}_h$.

N	Triangles	$\ p - p_h\ _{0,\mathcal{S}}$	$\ p - \tilde{\lambda}_h\ _{0,\mathcal{S}}$	$\ \mathbf{u} - \mathbf{u}_h\ _{\mathbf{H}(\text{div},\mathcal{S})}$
2	8×4	0.4445	0.1481	1.2247
4	32×4	0.2212	0.0389	0.6263
8	128×4	0.1102	0.0098	0.3150
16	512×4	0.0550	0.0025	0.1577
32	2048×4	0.0275	6.18·10 ⁻⁴	0.0789
64	8192×4	0.0138	1.54·10 ⁻⁴	0.0394
128	32768×4	0.0069	3.87·10 ⁻⁵	0.0197
256	131072×4	0.0034	9.73·10 ⁻⁶	0.0099

Table 4.1: Approximation errors in α_1 , the first model problem

N	Triangles	$\ p - p_h\ _{0,\mathcal{S}}$	$\ p - \tilde{\lambda}_h\ _{0,\mathcal{S}}$	$\ \mathbf{u} - \mathbf{u}_h\ _{\mathbf{H}(\text{div},\mathcal{S})}$
2	8×2	0.4481	0.1496	1.2236
4	32×2	0.2212	0.0393	0.6262
8	128×2	0.1102	0.0099	0.3150
16	512×2	0.0550	0.0025	0.1577
32	2048×2	0.0275	6.24·10 ⁻⁴	0.0789
64	8192×2	0.0138	1.56·10 ⁻⁴	0.0394
128	32768×2	0.0069	3.90·10 ⁻⁵	0.0197
256	131072×2	0.0034	9.76·10 ⁻⁶	0.0099

Table 4.2: Approximation errors in α_1 , the second model problem

The second model problem is given by

$$\mathcal{S} = \overline{\alpha_1} \cup \overline{\alpha_2} \setminus \partial\mathcal{S},$$

$$\begin{aligned} \mathbf{u} &= -(\nabla p + \nabla z) \quad \text{in } \alpha_i, i = 1, 2, \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \alpha_i, i = 1, 2. \end{aligned}$$

The boundary conditions on $\Gamma_1\text{--}\Gamma_6$ are given as in the previous case. Also the exact solution in α_1 and α_2 stays unchanged. Table 4.2 gives the approximation errors in the first rectangle α_1 for this model problem. As the exact solution in α_1 coincides with that of the first model problem, we can compare these results with that of Table 4.1. The difference in approximation error is very small even for rough triangulations and disappears for increasing N . Hence a confirmation of the conclusions outlined by the theory: the presence of multiply shared interpolygon boundaries does not influence the approximation properties of the lowest-order Raviart–Thomas mixed finite element method.

4.7.2 Real problem

We give an example of fracture flow around the explorational drill hole Ptp-3 in the granitoid massif of Potůčky, Western Bohemia in this section. There exists a large variety of approaches to modeling the flow through a network of polygonal disks representing the rock fractures. In [34, 50, 87] the networks of polygonal disks are replaced by networks of one-dimensional pipes. This allows for fast calculations with large networks, but the precision is compromised.

The models proposed in [9, 21, 54, 85, 113] discretize the polygonal networks into triangular or quadrilateral meshes. Because of a very complex geometry, the number of mesh elements is often sizably increased. Finite difference, finite volume, finite element, or boundary element methods are used for the discretization. We refer e.g. to [27] for a more detailed survey.

Our intention was twofold. First, we have constructed a very accurate mesh of the fracture network, which had at the same time as few elements as possible. Second, we have used the mixed finite element method studied in this chapter for the discretization of the fracture flow problem. We have approximated the original three-dimensional fractures by planar polygonal disks whose frequency, size, orientation, assigned aperture, wall roughness, and filling were statistically described from field measurements (core-log evaluation, acoustic camera scanning, ...), given in [89]. We have next computed the intersections of the polygons. In order to simplify the system of intersections in each polygon, these were slightly moved and stretched in the polygon planes. This allows a significant decrease of the number of triangular elements necessary to discretize each polygon and an improvement of their shapes. The triangular mesh has to respect the system of intersections in each polygon, but the interpolygon geometrical correspondence vanishes. This was replaced with an element edges correspondence, sufficient for the mixed finite element method. Briefly, the corresponding edges do not necessarily match geometrically—only what is the outflow from one triangular element through a given edge has to be the inflow into the neighboring ones through the edges that are associated with the given one. Finally, based on the assigned aperture, fracture wall roughness, and filling, the hydraulic permeability of each element was set. The classical parallel plate model was thus avoided.

The optimized triangulation of the fracture network and the model allowing for variable permeability inside the fractures together with the mixed finite element method ensuring the mass balance in each element even for meshes with no real geometrical correspondence have proved a good correspondence between observed phenomena and the numerical approximation. The model gave an accurate velocity field within fracture planes and thus in the whole simulated network. Namely, the channeling effect was successfully simulated both in fracture planes and in the entire network. This effect is given by the fact that the natural three-dimensional fractures have varying apertures and consequently the flow is not evenly distributed within the fracture planes. An example of the distribution of the piezometric head in a fracture network is given in Figure 4.4. These results are summarized in a paper written in collaboration with J. Maryška and O. Severýn which has been published in *Computational Geosciences*.

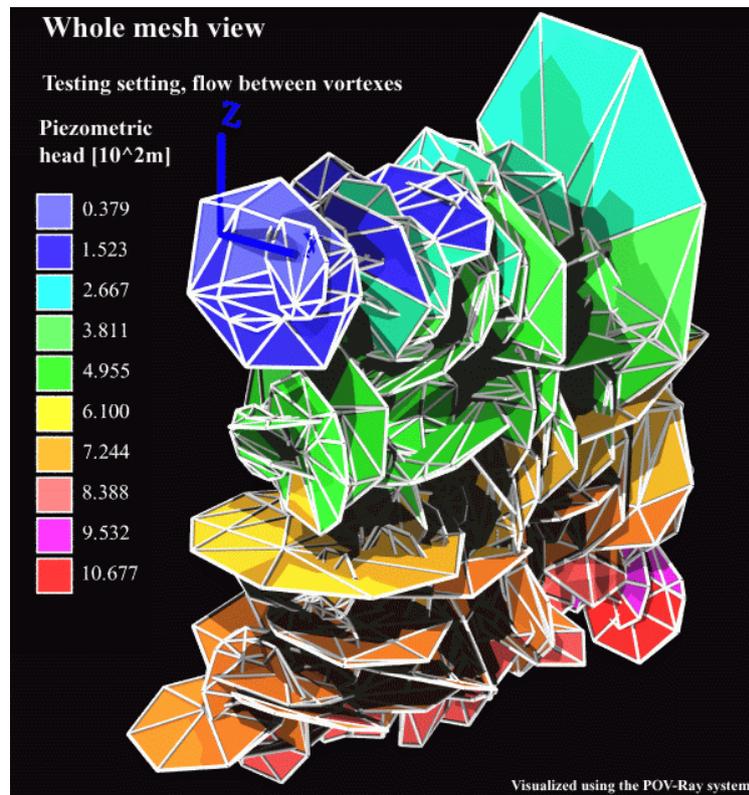


Figure 4.4: Distribution of the piezometric head in a fracture network

Bibliographie

- [1] AAVATSMARK I., BARKVE T., BØE Ø., MANNSETH T., Discretization on unstructured grids for inhomogeneous, anisotropic media. Part I: Derivation of the methods, *SIAM J. Sci. Comput.* **19** (1998), 1700–1716.
- [2] AAVATSMARK I., BARKVE T., BØE Ø., MANNSETH T., Discretization on unstructured grids for inhomogeneous, anisotropic media. Part II: Discussion and numerical results, *SIAM J. Sci. Comput.* **19** (1998), 1717–1736.
- [3] ACHDOU Y., JAPHET C., MADAY Y., NATAF F., A new cement to glue non-conforming grids with Robin interface conditions: The finite volume case, *Numer. Math.* **92** (2002), 593–620.
- [4] ADAMS R.A., *Sobolev Spaces*, Academic Press, New York, 1975.
- [5] ADLER P.M., THOVERT J.-F., *Fractures and Fracture Networks*, Kluwer, Dordrecht, 1999.
- [6] AFIF M., AMAZIANE B., Convergence of finite volume schemes for a degenerate convection–diffusion equation arising in flow in porous media, *Comput. Methods Appl. Mech. Engrg.* **191** (2002), 5265–5286.
- [7] AFTOSMIS M., GAITONDE D., SEAN TAVARES T., On the accuracy, stability and monotonicity of various reconstruction algorithms for unstructured meshes, *AIAA* (1994), paper No. 94-0415.
- [8] AGOUZAL A., BARANGER J., MAÎTRE J.-F., OUDIN F., Connection between finite volume and mixed finite element methods for a diffusion problem with nonconstant coefficients. Application to a convection diffusion problem, *East-West J. Numer. Math.* **3** (1995), 237–254.
- [9] ANDERSSON J., DVERSTORP B., Conditional simulations of fluid flow in three-dimensional networks of discrete fractures, *Water Resour. Res.* **23** (1987), 1876–1886.
- [10] ANGOT A., DOLEJŠÍ V., FEISTAUER M., FELCMAN J., Analysis of a combined barcentric finite volume–nonconforming finite element method for nonlinear convection–diffusion problems, *Appl. Math.* **43** (1998), 263–310.
- [11] ARBOGAST T., COWSAR L.C., WHEELER M.F., YOTOV I., Mixed finite element methods on nonmatching multiblock grids, *SIAM J. Numer. Anal.* **37** (2000), 1295–1315.
- [12] ARBOGAST T., DAWSON C.N., KEENAN P.T., WHEELER M.F., YOTOV I., Enhanced cell-centered finite differences for elliptic equations on general geometry, *SIAM J. Sci. Comput.* **19** (1998), 404–425.
- [13] ARBOGAST T., WHEELER M.F., YOTOV I., Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences, *SIAM J. Numer. Anal.* **34** (1997), 828–852.

- [14] ARBOGAST T., WHEELER M.F., ZHANG N., A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media, *SIAM J. Numer. Anal.* **33** (1996), 1669–1687.
- [15] ARNOLD D.N., BREZZI F., Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates, *RAIRO Modél. Math. Anal. Numér.* **19** (1985), 7–32.
- [16] BABUŠKA I., Error-bounds for finite element method, *Numer. Math.* **16** (1971), 322–333.
- [17] BANK R.E., ROSE D.J., Some error estimates for the box method, *SIAM J. Numer. Anal.* **24** (1987), 777–787.
- [18] BARANGER J., MAÎTRE J.-F., OUDIN F., Connection between finite volume and mixed finite element methods, *M2AN Math. Model. Numer. Anal.* **30** (1996), 445–465.
- [19] BARRETT J.W., KNABNER P., Finite element approximation of transport of reactive solutes in porous media. Part I: Error estimates for non-equilibrium adsorption processes, *SIAM J. Numer. Anal.* **34** (1997), 201–227.
- [20] BARRETT J.W., KNABNER P., Finite element approximation of transport of reactive solutes in porous media. Part II: Error estimates for the equilibrium adsorption processes, *SIAM J. Numer. Anal.* **34** (1997), 455–479.
- [21] BASTIAN P., CHEN Z., EWING R.E., HELMIG R., JAKOBS H., REICHENBERGER V., Numerical simulation of multiphase flow in fractured porous media, in Chen Z., Ewing R.E., Shi Z.C. eds., *Numerical Treatment of Multiphase Flows in Porous Media*, pp. 50–68, Springer-Verlag, Berlin, 2000.
- [22] BAUGHMAN L.A., WALKINGTON N.J., Co-volume methods for degenerate parabolic problems, *Numer. Math.* **64** (1993), 45–67.
- [23] BEAR J., *Dynamics of Fluids in Porous Media*, American Elsevier, New York, 1972.
- [24] BEAR J., Modeling flow and contaminant transport in fractured rocks, in Bear J., Tsang C.F., de Marsily G. eds., *Flow and Contaminant Transport in Fractured Rock*, pp. 1–38, USA: Academic Press, 1993.
- [25] BEAR J., VERRUIJT A., *Modelling Groundwater Flow and Pollution*, Reidel, Dordrecht, 1987.
- [26] BERNARDI C., MADAY Y., PATERA A., A new nonconforming approach to domain decomposition: The mortar element method, in Brezis H., Lions J.L. eds., *Nonlinear Partial Differential Equations and their Applications*, pp. 13–51, Pitman, London, 1994.
- [27] BOGDANOV I.I., MOURZENKO V.V., THOVERT J.-F., ADLER P.M., Effective permeability of fractured porous media in insteady state flow, *Water Resour. Res.* **39** (2003), 1023, doi:10.1029/2001WR000756.
- [28] BRENNER S.C., Poincaré–Friedrichs inequalities for piecewise H^1 functions, *SIAM J. Numer. Anal.* **41** (2003), 306–324.
- [29] BREZIS H., *Analyse Fonctionnelle, Théorie et Applications*, Masson, Paris, 1983.
- [30] BREZZI F., On the existence, uniqueness and approximation of saddle point problems arising from Lagrangian multipliers, *RAIRO Modél. Math. Anal. Numér.* **8** (1974), 129–151.
- [31] BREZZI F., DOUGLAS J. JR., DURAN R., FORTIN M., Mixed finite elements for second order elliptic problems in three variables, *Numer. Math.* **51** (1987), 237–250.
- [32] BREZZI F., DOUGLAS J. JR., MARINI L.D., Two families of mixed finite elements for second order elliptic problems, *Numer. Math.* **47** (1985), 217–235.

- [33] BREZZI F., FORTIN M., *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [34] CACAS M.C., LEDOUX E., DE MARSILY G., TILLIE B., BARBREAU A., DURAND E., FEUGA B., PEAUDE CERF P., Modeling fracture flow with a stochastic discrete fracture network: Calibration and validation 1. The flow model, *Water Resour. Res.* **26** (1990), 479–489.
- [35] CAI Z., MANDEL J., MCCORMICK S., The finite volume element method for diffusion equations on general triangulations, *SIAM J. Numer. Anal.* **28** (1991), 392–402.
- [36] CAUTRÈS R., HERBIN R., HUBERT F., The Lions domain decomposition algorithm on non matching cell-centered finite volume meshes, to appear in *IMA J. Numer. Anal.*
- [37] CHAVENT G., YOUNÈS A., ACKERER PH., On the finite volume reformulation of the mixed finite element method for elliptic and parabolic PDE on triangles, *Comput. Methods Appl. Mech. Engrg.* **192** (2003), 655–682.
- [38] CHEN Z., Equivalence between and multigrid algorithms for nonconforming and mixed methods for second-order elliptic problems, *East-West J. Numer. Math.* **4** (1996), 1–33.
- [39] CHEN Z., EWING R.E., Degenerate two-phase incompressible flow III: Sharp error estimates, *Numer. Math.* **90** (2001), 215–240.
- [40] CHEN Z., EWING R.E., JIANG Q., SPAGNUOLO A.M., Error analysis for characteristic-based methods for degenerate parabolic problems, *SIAM J. Numer. Anal.* **40** (2002), 1491–1515.
- [41] CIARLET P.G., Basic error estimates for elliptic problems, in Ciarlet P.G., Lions J.L. eds., *Handbook of Numerical Analysis*, vol. 2, pp. 17–351, Elsevier Science B.V., Amsterdam, 1991.
- [42] CIARLET P.G., *The Finite Element Method for Elliptic Problems*, North-Holland Publishing Company, Amsterdam, 1978.
- [43] CLÉMENT PH., Approximation by finite element functions using local regularization, *RAIRO Modél. Math. Anal. Numér.* **9** (1975), 77–84.
- [44] COUDIÈRE Y., VILA J.-P., VILLEDIEU PH., Convergence rate of a finite volume scheme for a two dimensional convection–diffusion problem, *M2AN Math. Model. Numer. Anal.* **33** (1999), 493–516.
- [45] CROUZIEX M., RAVIART P.-A., Conforming and nonconforming methods for solving the stationary Stokes equations I, *RAIRO Modél. Math. Anal. Numér.* **7** (1973), 33–76.
- [46] DAWSON C., Analysis of an upwind-mixed finite element method for nonlinear contaminant transport equations, *SIAM J. Numer. Anal.* **35** (1998), 1709–1724.
- [47] DAWSON C., AIZINGER V., Upwind-mixed methods for transport equations, *Comput. Geosci.* **3** (1999), 93–110.
- [48] DEBIEZ C., DERVIEUX A., MER K., NKONGA B., Computation of unsteady flows with mixed finite-volume/finite-element upwind methods, *Internat. J. Numer. Methods Fluids* **27** (1998), 193–206.
- [49] DEIMLING K., *Nonlinear Functional Analysis*, Springer-Verlag, Berlin-Heidelberg, 1985.
- [50] DERSHOWITZ W.S., FIDELIBUS C., Derivation of equivalent pipe network analogues for three-dimensional discrete fracture networks by the boundary element method, *Water Resour. Res.* **35** (1999), 2685–2691.
- [51] DOLEJŠÍ V., FEISTAUER M., FELCMAN J., On the discrete Friedrichs inequality for nonconforming finite elements, *Numer. Funct. Anal. Optim.* **20** (1999), 437–447.

- [52] DOUGLAS J. JR., ROBERTS J.E., Global estimates for mixed methods for second order elliptic equations, *Math. Comp.* **44** (1985), 39–52.
- [53] EBMEYER C., Error estimates for a class of degenerate parabolic equations, *SIAM J. Numer. Anal.* **35** (1998), 1095–1112.
- [54] ELSWORTH D., A hybrid boundary element-finite element analysis procedure for fluid flow simulation in fractured rock masses, *Int. J. Numer. Anal. Methods Geomech.* **10** (1986), 569–584.
- [55] EWING R.E., LAZAROV R.D., LIN T., LIN Y., Mortar finite volume element approximations of second-order elliptic problems, *East-West J. Numer. Math.* **8** (2000), 93–110.
- [56] EWING R.E., LIN T., LIN Y., On the accuracy of the finite volume element method based on piecewise linear polynomials, *SIAM J. Numer. Anal.* **39** (2002), 1865–1888.
- [57] EYMARD R., GALLOUËT T., Convergence d'un schéma de type éléments finis - volumes finis pour un système couplé elliptique - hyperbolique, *RAIRO Modél. Math. Anal. Numér.* **27** (1993), 843–861.
- [58] EYMARD R., GALLOUËT T., HERBIN R., A finite volume scheme for anisotropic diffusion problems, submitted to *C. R. Acad. Sci. Paris., Ser. I*, 2004.
- [59] EYMARD R., GALLOUËT T., HERBIN R., Convergence of finite volume schemes for semilinear convection diffusion equations, *Numer. Math.* **82** (1999), 91–116.
- [60] EYMARD R., GALLOUËT T., HERBIN R., Finite volume approximation of elliptic problems and convergence of an approximate gradient, *Appl. Numer. Math.* **37** (2001), 31–53.
- [61] EYMARD R., GALLOUËT T., HERBIN R., Finite volume methods, in Ciarlet P.G., Lions J.L. eds., *Handbook of Numerical Analysis*, vol. 7, pp. 713–1020, Elsevier Science B.V., Amsterdam, 2000.
- [62] EYMARD R., GALLOUËT T., HERBIN R., MICHEL A., Convergence of a finite volume scheme for nonlinear degenerate parabolic equations, *Numer. Math.* **92** (2002), 41–82.
- [63] EYMARD R., GALLOUËT T., HILHORST D., NAÏT SLIMANE Y., Finite volumes and nonlinear diffusion equations, *M2AN Math. Model. Numer. Anal.* **32** (1998), 747–761.
- [64] EYMARD R., GUTNIC M., HILHORST D., The finite volume method for the Richards equation, *Comput. Geosci.* **3** (2000), 259–294.
- [65] FAILLE I., A control volume method to solve an elliptic equation on a two-dimensional irregular mesh, *Comput. Methods Appl. Mech. Engrg.* **100** (1992), 275–290.
- [66] FAILLE I., NATAF F., SAAS L., WILLIEN F., Finite volume methods on non-matching grids with arbitrary interface conditions and highly heterogeneous media, in Kornhuber R., Hoppe R., Périaux J., Pironneau O., Widlund O., Xu J. eds., *Domain Decomposition Methods in Science and Engineering*, Lect. Notes Comput. Sci. Eng. 40, pp. 50–68, Springer-Verlag, Berlin, 2000.
- [67] FEISTAUER M., FELCMAN J., MEDVIĐOVÁ-LUKÁČOVÁ M., On the convergence of a combined finite volume–finite element method for nonlinear convection–diffusion problems, *Numer. Methods Partial Differential Equations* **13** (1997), 1–28.
- [68] FORSYTH P.A., A control volume finite element approach to NAPL groundwater contamination, *SIAM J. Sci. Stat. Comput.* **12** (1991), 1029–1057.
- [69] FRAEJIS DE VEUBEKE B., Displacement and equilibrium models in the finite element method, in Zienkiewicz O.C., Holister G. eds., *Stress Analysis*, Wiley, New York, 1965.
- [70] GILBERT J.R., MOLER C., SCHREIBER R., Sparse matrices in MATLAB: Design and implementation, *SIAM J. Matrix Anal. Appl.* **13** (1992), 333–356.

- [71] HACKBUSCH W., On first and second order box schemes, *Computing* **41** (1989), 277–296.
- [72] HERBIN R., An error estimate for a finite volume scheme for a diffusion–convection problem on a triangular mesh, *Numer. Methods Partial Differential Equations* **11** (1995), 165–173.
- [73] HESTENES M.R., STIEFEL E., Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Stand.* **49** (1952), 409–436.
- [74] HOTEIT H., ERHEL J., MOSÉ R., PHILIPPE B., ACKERER PH., Numerical reliability for mixed methods applied to flow problems in porous media, *Comput. Geosci.* **6** (2002), 161–194.
- [75] HUGHES T.J.R., ENGEL G., MAZZEI L., LARSON M.G., The continuous Galerkin method is locally conservative, *J. Comput. Phys.* **163** (2000), 467–488.
- [76] IDELSOHN S., ONATE E., Finite volumes and finite elements: Two “good friends”, *Internat. J. Numer. Methods Engrg.* **37** (1994), 3323–3341.
- [77] JAFFRÉ J., Éléments finis mixtes et décentrage pour les équations de diffusion–convection, *Calcolo* **23** (1984), 171–197.
- [78] JÄGER W., KAČUR J., Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes, *M2AN Math. Model. Numer. Anal.* **29** (1995), 605–627.
- [79] JAVANDEL I., DOUGHTY C., TSANG C.F., *Groundwater Transport: Handbook of Mathematical Models*, American Geophysical Union Water Resources Monograph Series, Volume 10, 1984.
- [80] JOE B., Delaunay triangular meshes in convex polygons, *SIAM J. Sci. Stat. Comput.* **7** (1986), 514–539.
- [81] KAČUR J., Solution of degenerate convection–diffusion problems by the method of characteristics, *SIAM J. Numer. Anal.* **39** (2001), 858–879.
- [82] KARLSEN K.H., RISEBRO N.H., TOWERS J.D., Upwind difference approximations for degenerate parabolic convection–diffusion equations with a discontinuous coefficient, *IMA J. Numer. Anal.* **22** (2002), 623–664.
- [83] KNABNER P., OTTO F., Solute transport in porous media with equilibrium and nonequilibrium multiple-site adsorption: Uniqueness of weak solutions, *Nonlinear Anal.* **42** (2000), 381–403.
- [84] KNOBLOCH P., Uniform validity of discrete Friedrich’s inequality for general nonconforming finite element spaces, *Numer. Funct. Anal. Optim.* **22** (2001), 107–126.
- [85] KOUDINA N., GONZALEZ GARCIA R., THOVERT J.-F., ADLER P.M., Permeability of three-dimensional fracture networks, *Phys. Rev. E* **57** (1998), 4466–4479.
- [86] LETNIEWSKI F.W., Three-dimensional Delaunay triangulations for finite element approximations to a second-order diffusion operator, *SIAM J. Sci. Stat. Comput.* **13** (1992), 765–770.
- [87] LONG J.C.S., GILMOUR P., WITHERSPOON P.A., A model for steady state flow in random three dimensional networks of disc-shaped fractures, *Water Resour. Res.* **21** (1985), 1105–1115.
- [88] MARINI L.D., An inexpensive method for the evaluation of the solution of the lowest order Raviart–Thomas mixed method, *SIAM J. Numer. Anal.* **22** (1985), 493–496.
- [89] MAROS G., PALOTÁS K., KOROKNAI B., SALLAY E., SZONGOTH G., KASZA Z., ZILÁHI-SEBESS L., Core log evaluation of borehole Ptp-3 in the Krušné hory mts, MS Geological Institute of Hungary, Budapest, 2001.

- [90] MEDVIĐOVÁ-LUKÁČOVÁ M., Combined finite element–finite volume method (convergence analysis), *Comment. Math. Univ. Carolinae* **38** (1997), 717–741.
- [91] NEČAS J., *Les Méthodes Directes en Théorie des Equations Elliptiques*, Masson, Paris, 1967.
- [92] NÉDÉLEC J.C., Mixed finite elements in \mathbb{R}^3 , *Numer. Math.* **35** (1980), 315–341.
- [93] NOCHETTO R.H., SCHMIDT A., VERDI C., A posteriori error estimation and adaptivity for degenerate parabolic problems, *Math. Comp.* **69** (1999), 1–24.
- [94] NOCHETTO R.H., VERDI C., Approximation of degenerate parabolic problems using numerical integration, *SIAM J. Numer. Anal.* **25** (1988), 784–814.
- [95] OHLBERGER M., A posteriori error estimates and adaptive methods for convection dominated transport processes, Ph.D. thesis, Albert–Ludwigs–Universität Freiburg, Germany, 2001.
- [96] OHLBERGER M., A posteriori error estimates for vertex centered finite volume approximations of convection–diffusion–reaction equations, *M2AN Math. Model. Numer. Anal.* **35** (2001), 355–387.
- [97] PASCAL F., Sur des méthodes d’approximation effectives et d’analyse numérique pour les équations de la mécanique des fluides, Habilitation thesis, Université de Paris-Sud, Orsay, France, 2002.
- [98] POP I.S., Regularization methods in the numerical analysis of some degenerate parabolic equations, Ph.D. thesis, Faculty of Mathematics and Informatics, University Cluj Napoca, Romania, 1998.
- [99] POP I.S., YONG W.-A., A numerical approach to degenerate parabolic equations, *Numer. Math.* **92** (2002), 357–381.
- [100] PUTTI M., CORDES C., Finite element approximation of the diffusion operator on tetrahedra, *SIAM J. Sci. Comput.* **19** (1998), 1154–1168.
- [101] QUARTERONI A., SACCO R., SALERI F., *Numerical Mathematics*, Springer-Verlag, New York, 2000.
- [102] QUARTERONI A., VALLI A., *Domain Decomposition Methods for Partial Differential Equations*, Oxford University Press, New York, 1999.
- [103] QUARTERONI A., VALLI A., *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1994.
- [104] RAMAROSY N., TALISMAN, guide d’utilisation, société HydroExpert, 2001. HydroExpert, 53 rue Charles Frérot, 94 250 Gentilly, France, www.hydroexpert.com.
- [105] RAVIART P.-A., THOMAS J.-M., A mixed finite element method for 2-nd order elliptic problems, in Galligani I., Magenes E. eds., *Mathematical Aspects of Finite Element Methods*, Lecture Notes in Math. 606, pp. 292–315, Springer, Berlin, 1977.
- [106] REDDY J.N., ODEN J.T., Mathematical theory of mixed finite element approximations, *Q. Appl. Math.* **33** (1975), 255–280.
- [107] REKTORYS K., *Variational Methods in Mathematics, Science, and Engineering*, Kluwer, Dordrecht, 1982.
- [108] ROBERTS J.E., THOMAS J.-M., Mixed and hybrid methods, in Ciarlet P.G., Lions J.L. eds., *Handbook of Numerical Analysis*, vol. 2, pp. 523–639, Elsevier Science B.V., Amsterdam, 1991.
- [109] RULLA J., WALKINGTON N.J., Optimal rates of convergence for degenerate parabolic problems in two dimensions, *SIAM J. Numer. Anal.* **33** (1996), 56–67.

- [110] RUSSELL T.F., WHEELER M.F., Finite element and finite difference methods for continuous flows in porous media, in Ewing R.E. ed., *The Mathematics of Reservoir Simulation*, pp. 35–106, SIAM, Philadelphia, 1983.
- [111] SAAD Y., *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, 1996.
- [112] SELMIN V., The node-centered finite volume approach: Bridge between finite differences and finite elements, *Comput. Methods Appl. Mech. Engrg.* **102** (1993), 107–138.
- [113] SLOUGH K.J., SUDICKY E.A., FORSYTH P.A., Numerical simulations of multiphase flow and phase partitioning in discretely fractured geological media, *J. Contam. Hydrol.* **40** (1999), 107–136.
- [114] SONIER F., EYMARD R., Mathematical and numerical properties of control-volume finite element scheme for reservoir simulation, *SPE Reservoir Engineering* (November 1994), 283–289.
- [115] STRANG G., FIX G.J., *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, 1973.
- [116] TEMAM R., *Navier–Stokes Equations*, North-Holland, Amsterdam, 1979.
- [117] VAN DER VORST H.A., Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems, *SIAM J. Sci. Stat. Comput.* **13** (1992), 631–644.
- [118] VAN GENUCHTEN M., A closed form for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Amer. J.* **44** (1980), 892–898.
- [119] WANFANG Z., WHEATER H.S., JOHNSTON P.M., State of the art of modelling two-phase flow in fractured rock, *Envir. Geol.* **31** (1997), 157–166.
- [120] YOTOV I., Mixed finite element methods for flow in porous media, Ph.D. thesis, Rice University, Houston, Texas, 1996.
- [121] YOUNÈS A., ACKERER PH., CHAVENT G., From mixed finite elements to finite volumes for elliptic PDEs in two and three dimensions, *Internat. J. Numer. Methods Engrg.* **59** (2004), 365–388.
- [122] YOUNÈS A., MOSE R., ACKERER PH., CHAVENT G., A new formulation of the mixed finite element method for solving elliptic and parabolic PDE with triangular elements, *J. Comput. Phys.* **149** (1999), 148–167.
- [123] ZHENG C., BENNETT G.D., *Applied Contaminant Transport Modeling*, Van Nostrand Reinhold, New York, 1995.
- [124] ZIENKIEWICZ O.C., *The Finite Element Method*, McGraw-Hill, London, 1977.