Université de Lille

Université de Lille

École doctorale **EDSPI**

Unité de recherche **Laboratoire Paul Painlevé**

Thèse présentée par **Benoît Gaudeul**

Soutenue le **30 août 2021**

En vue de l'obtention du grade de docteur de l'Université de Lille

Discipline **Mathématique**

Approximation numérique entropique pour des systèmes de diffusion croisée issus de la physique

Thèse dirigée par Clément Cancès
Claire Chainais-Hillairet

Composition du jury

Rapporteurs	Jan-Frederik Pietschmann	professeur	
	Francis Filbet	professeur	
Examinateurs	Boris Andreianov	professeur	président du jury
	Virginie Ehrlacher	MCF HDR	
	Hélène Hivert	MCF	
Directeurs de thèse	Clément Cancès	directeur de recherche	
	Claire Chainais-Hillairet	professeure	

Colophon

Mémoire de thèse intitulé « Approximation numérique entropique pour des systèmes de diffusion croisée issus de la physique », écrit par Benoît Gaudeul, achevé le 24 août 2021, composé au moyen du système de préparation de document LaTeX et de la classe yathesis dédiée aux thèses préparées en France.

Université de Lille

Université de Lille

École doctorale **EDSPI**

Unité de recherche **Laboratoire Paul Painlevé**

Thèse présentée par **Benoît Gaudeul**

Soutenue le **30 août 2021**

En vue de l'obtention du grade de docteur de l'Université de Lille

Discipline **Mathématique**

Approximation numérique entropique pour des systèmes de diffusion croisée issus de la physique

Thèse dirigée par Clément Cancès
Claire Chainais-Hillairet

Composition du jury

Rapporteurs	Jan-Frederik Pietschmann	professeur	
	Francis Filbet	professeur	
Examinateurs	Boris Andreianov	professeur	président du jury
	Virginie Ehrlacher	MCF HDR	
	Hélène Hivert	MCF	
Directeurs de thèse	Clément Cancès	directeur de recherche	
	Claire Chainais-Hillairet	professeure	

Université de Lille

Université de Lille

Doctoral School **EDSPI**

University Department **Laboratoire Paul Painlevé**

Thesis defended by **Benoît Gaudeul**

Defended on **August 30, 2021**

In order to become Doctor from Université de Lille

Academic Field **Mathematics**

Entropic numerical approximations for cross-diffusion systems arising in physics

Thesis supervised by Clément Cancès
Claire Chainais-Hillairet

Committee members

Referees	Jan-Frederik Pietschmann	Professor	
	Francis Filbet	Professor	
Examiners	Boris Andreianov	Professor	Committee President
	Virginie Ehrlacher	HDR Associate Professor	
	Hélène Hivert	Associate Professor	
Supervisors	Clément Cancès	Senior Researcher	
	Claire Chainais-Hillairet	Professor	

Remerciements

En trois ans de thèse, j'ai pu bénéficier du soutien mathématique, amical, et spirituel de nombreuses personnes sans lesquelles ce travail n'aurait jamais commencé ou aboutis.

J'aimerais ainsi remercier en premier lieu Claire et Clément qui ont accepté de me prendre en thèse et m'ont guidé dans ma découverte des approximations numériques entropiques. Nos discussions ont considérablement enrichi les perspectives de recherches que j'avais, m'évitant redites et détours [1], et leurs encouragements m'ont convaincu de laisser voir les résultats obtenus en dehors du laboratoire à travers publications et exposés. Là encore, il est peu de mots pour exprimer ma gratitude pour les heures qu'ils ont passées à émonder mes nombreux brouillons. D'aucuns auraient dit de cette tâche "this quest is however hopeless".

J'aimerais aussi remercier chaleureusement Francis Filbet et Jan-Frederik Pietschmann d'avoir accepté de rapporter cette thèse. Le temps qu'ils y ont consacré m'honore, et je leur en suis infiniment reconnaissant. Ils ont aussi accepté avec Boris Andreianov, Virginie Ehrlacher, et Hélène Hivert de venir à quelques jours de la rentrée faire partie de mon jury de thèse, et je remercie chacun d'eux.

Merci à Jürgen de m'avoir permis de travailler sur des systèmes physiquement réalistes, et ainsi de porter ma pierre de matheux à la transition écologique.

Au laboratoire, j'ai eu l'occasion d'user de nombreux voisins de bureau. S'ils ont eu la patience de ne jamais se plaindre de mes ronchonnements sur la sécheresse et la précision nécessaire pour écrire un article, ou de la colonisation du tableau par mes gribouillages, Claire, Antoine, Julien, Federica, Tarlan, Florent, Igor et David n'ont jusqu'à présent pas reçu le crédit qu'ils méritaient. J'ai beaucoup apprécié leur compagnie et me réjouis de retrouver l'un d'eux à Lyon. Nos échanges ont contribué pour une bonne partie aux idées présentes dans cette thèse, et leur soutien amical son achèvement. Je les en remercie du fond du cœur.

Cette thèse a aussi été inspirée par les rencontres avec les membres des équipes Rapsody et Paradyse, du laboratoire Paul Painlevé, et les doctorants des différentes formations. Ma faible mémoire des noms et des visages a mené à quelques situations cocasses, au RU ou ailleurs, et me prévient de tenter une liste. Je leur en suis d'autant plus reconnaissant qu'ils me le pardonneront.

À la maison j'ai su compter sur le soutien quotidien de Chloé, son attention et l'entraide que nous avons vécu ces trois ans tant pour ses mémoires que pour cette thèse fut un trésor que je chéris et souhaite voir grandir. En fine psychologue elle a su attribuer mes sautes d'humeur aux défauts de compacité des schémas

1. À titre d'exemple, le chapitre sur la modélisation variationnelle aurait pu disparaître au profit d'un essai de catégorisation des co-cluster graphs

présentés ici, sans savoir précisément ce que cela signifiait. Merci d'avoir su surmonter ses craintes pour accepter d'accueillir Miguel dans notre foyer malgré ses faiblesses. Merci aussi d'avoir su accepter les horaires improbables que l'université me permettait, et d'avoir insisté -en dépit des résultats- pour que je dorme, même la veille d'une date limite auto-imposée.

J'aimerais aussi remercier ses parents, frères et sœurs, qui ont accepté de nous voir en permanence les temps des confinements, et les miens qui ont accepté de ne pas avoir de nouvelles quand je tentais de finir la rédaction des pages que vous lisez. Leur compréhension et leur soutien me furent très précieux. Merci à Pauline, toujours prête à prendre le temps d'un coup de fil ou à relire mon anglais, et à Martin qui semble s'engager dans une aventure scientifique comparable. Avec Claire, ils se sont donné bien du mal pour que la présentation de ce manuscrit soit à la hauteur d'un mariage avec la science.

Ces années de thèse auraient été d'une grande tristesse sans la présence amicale de Jean-Cyrille, Amaury, Joséphine, François, Gregory, Gaëlle, Yannick, Sébastien, Karine, François, Thérèse, Roxanne... Sans nos déjeuners et jeux de société, la vie à Lille aurait été aussi claire qu'un ciel d'octobre. Merci à Laurent, Pierre, Samuel, Nicolas pour les joyeux coups de fil que nous avons pu échanger, merci à Joachim pour son amitié et ses conseils sur la survie de La Plante (dont l'espèce précise reste inconnue). Outre leur soutien amical, nombre d'entre eux m'ont aussi porté par la prière, et je sais devoir au moins l'annexe de cette thèse à leur intercession.

Table des matières

Remerciements vii

Table des matières **ix**

1 Introduction **1**
1.1 Modèles étudiés dans la thèse 2
1.1.1 Nernst-Planck-Poisson 3
1.1.2 Au-delà de Nernst-Planck-Poisson 4
1.1.3 Détournement du modèle d'électrolyte 8
1.1.4 Vers d'autres modèles d'électrolytes 8
1.1.5 Fabrication de panneaux solaires 13
1.2 Schémas 15
1.2.1 Maillage 16
1.2.2 Quelques schémas pour Nernst-Planck-Poisson 18
1.2.3 Pistes d'analyse de ces schémas 20
1.3 Organisation et résultats principaux 25
1.3.1 Idées clefs et résultats principaux du chapitre 2 25
1.3.2 Idées clefs et résultats principaux du chapitre 3 28
1.3.3 Idées clefs, résultats principaux et perspectives du chapitre 4 29
1.3.4 Idées clefs et résultats principaux du chapitre 5 29

2 Finite volume schemes for unipolar degenerated drift-diffusion **33**
2.1 Introduction 34
2.1.1 Motivation 34
2.1.2 A simplified unipolar degenerate drift-diffusion model . . . 36
2.1.3 Entropy structure and weak solutions 37
2.2 Finite Volume approximations 40
2.2.1 Discretization of $(0,T)\times\Omega$ 40
2.2.2 A common basis for the Finite Volume schemes 42
2.2.3 Numerical fluxes for the conservation of the chemical species 43
2.2.4 Main results and organisation of the paper 45
2.3 Numerical analysis for fixed meshes 47

2.3.1 Face concentration and face dissipation 47
2.3.2 Uniform a priori estimates 50
2.3.3 Existence of a solution to the schemes 54
2.4 About the convergence towards a weak solution 56
2.4.1 Reconstruction operators 57
2.4.2 Compactness properties for the approximate concentration 58
2.4.3 Convergence towards a weak solution 63
2.5 Numerical comparison of the schemes 68
2.5.1 1D time evolution and convergence test 68
2.5.2 1D stationary convergence test 70
2.5.3 2D Unipolar Field Effect Transistor 71
2.6 Conclusion 73
2.A L^∞ bound on the TPFA FV approximate Poisson equation . . . 76
2.B Proof of Lemma 2.3.2 78
2.C Comparison of face concentration functionals 81
2.D Some notations 84

3 Schemes for Nernst-Planck-Poisson with volume constraints 87
3.1 Introduction 88
3.1.1 The Nernst-Planck-Poisson system with finite ionic volumes 89
3.1.2 Key properties of the continuous system 91
3.1.3 Positioning and outline 94
3.2 Discretization and main Theorems 94
3.2.1 Discretization of $(0,T)\times\Omega$ 94
3.2.2 A common setting for the Finite Volume schemes 96
3.2.3 Numerical fluxes for the conservation equations 97
3.2.4 Main theorems 99
3.3 Fixed Mesh analysis 100
3.3.1 Analysis of numerical flux based functions 100
3.3.2 *A priori* estimates 102
3.3.3 Existence of solutions 106
3.4 Convergence 108
3.4.1 Reconstruction operators 108
3.4.2 Compactness 109
3.4.3 Identification 122
3.5 Numerical Examples 126
3.5.1 Species redistribution in a one-dimensional cell filled with binary electrolyte 126
3.5.2 1D stationary convergence test 127
3.5.3 An electrolytic diode 127
3.A Chemical free energy density and chemical potentials 133

3.B Proof of Lemma 3.3.2 . . . 134
3.B.1 Limit of $\Psi_{\delta,\epsilon,M,i}$ . . . 135
3.B.2 Limit of $\Upsilon_{\delta,M}$ . . . 136
3.C Study of a numerical scheme for $h_i = \log(c_i) - \alpha \log(c_0)$ . . . 137
3.D A simple convergence lemma . . . 139

4 Variationnal modeling **141**
4.1 Introduction . . . 142
4.1.1 The recipe of Variational modelling for generalized gradient flows . . . 143
4.2 Two preliminary corrections to the Poisson Nernst Planck system 145
4.2.1 Common settings . . . 146
4.2.2 Two ways of enforcing volume conservation . . . 147
4.3 Several possible choices for Ψ . . . 151
4.3.1 Fick-like dissipation . . . 151
4.3.2 Solvent-free dissipation . . . 152
4.3.3 Variation of the solvent-free dissipation . . . 153
4.3.4 Dissipation with macroscopic velocity . . . 154
4.3.5 Stefan-Maxwell dissipation . . . 156
4.4 Construction of schemes . . . 158
4.4.1 Discretisation of $\Omega \times [0,T]$ . . . 158
4.4.2 Formulation of the schemes . . . 160
4.5 Numerical comparison of the models . . . 164

5 A convergent scheme for a cross-diffusion system **173**
5.1 Introduction . . . 174
5.1.1 The system under study . . . 174
5.1.2 Formal structure . . . 175
5.1.3 Objectives . . . 179
5.2 Finite Volume approximation . . . 181
5.2.1 Discretization of $(0,T) \times \Omega$ . . . 181
5.2.2 Numerical scheme . . . 183
5.2.3 Main results and organization . . . 184
5.3 Numerical analysis on a fixed mesh . . . 186
5.3.1 A priori estimates . . . 186
5.3.2 Existence of solutions . . . 188
5.3.3 Entropy dissipation . . . 189
5.4 Convergence analysis . . . 192
5.4.1 Reconstruction operators . . . 192
5.4.2 Compactness properties . . . 193
5.4.3 Convergence towards a weak solution . . . 197

5.5 Numerical results 199
5.5.1 Convergence under grid refinement 199
5.5.2 On the influence of the parameter $a^\star$ 201
5.5.3 A 2D test case with reaction 201
5.6 Conclusion 203

A About the inverse activity coefficients **207**

Bibliography **211**

Abstract **223**

Chapitre 1

Introduction

L'objet de cette thèse est de proposer des outils pour simuler des phénomènes physiques complexes. Les systèmes décris dans les chapitres 3 et 4 servent à simuler le mouvement des ions dans des solutions fortement concentrées. C'est à dire, à représenter dans un ordinateur le fonctionnement d'une batterie, le comportement de la saumure d'une piqûre de corrosion, ou certains mécanismes de piles à combustibles. Dans le chapitre 5, le modèle a pour motivation la production de panneaux solaires par dépôt de vapeur. Le chapitre 2 lui a un double intérêt : d'une part il prépare mathématiquement le travail effectué dans le chapitre 3, d'une autre, il permet de simuler le fonctionnement de certains semi-conducteurs, matériaux clefs de l'électronique.

En dépit de ces applications possibles dans la transition écologique, les chapitres de cette thèse emploient un jargon de mathématicien plutôt que celui propre à chacun de ces domaines. Cette introduction a l'ambition, outre de présenter les idées et résultats principaux, d'introduire une partie de ce jargon. Ainsi, avant d'entrer dans le détail de la structure de cette introduction, on s'intéresse à décrypter le sens du titre.

Nous l'avons vu, les applications possibles de cette thèse sont variées, mais les modèles mathématiques sous-jacents ont un point commun : il s'agit de modèles de diffusion croisée. Ces modèles servent en physique et en biologie pour des systèmes où la vitesse et la direction de diffusion d'une espèce dépends non seulement de sa répartition (des zones où elle est très présente vers les zones où elle y est moins), mais aussi de celle des autres espèces [1]. Un petit peu comme sur une autoroute où les vitesses des voitures, motos et camions sont liées. Évidemment sur une autoroute, les conducteurs ont d'autres objectifs que d'éviter les collisions ce qui conduit à des phénomènes (et donc des modèles) différents [127].

1. Ce cadre est plus général que celui des matrices d'Onsager non diagonales, mais si vous êtes familier avec celles-ci c'est une bonne première approximation.

Les modèles considérés ont un autre point commun : on peut calculer une quantité qui décroît au cours du temps. En mathématiques, on appelle ça une fonctionnelle de Lyapunov, ou une entropie sous certaines conditions de convexité. Cette entropie ne correspond pas à celle de la thermodynamique, et ici sera souvent assimilée à une énergie libre. Dans la Section 1.1 on verra plus en détail l'origine physique des systèmes étudiés dans les différents chapitres de la thèse.

Outre la proximité mathématique des modèles, on cherche à créer des méthodes utilisables sur un ordinateur pour en calculer des solutions approchées ou approximations numériques. On s'intéressera exclusivement à des méthodes de volumes finis à deux points et on cherchera des schémas (méthodes de calcul d'approximations) qui préservent la décroissance de l'entropie. Ces méthodes sont introduites en Section 1.2 et on y illustre sur un modèle simple dérivé précédemment les idées d'analyse des deux premiers chapitres.

Enfin, en Section 1.3, on présentera les résultats principaux de chaque chapitre en s'appuyant sur les acquis des sections précédentes.

1.1 Modèles étudiés dans la thèse

Le principe de base de l'électrochimie est de faire réagir des ions avec des électrodes. Cela peut servir à produire de l'électricité (comme la pile de Volta), à en stocker, à déposer de la matière sur un objet qui servira d'électrode, à produire ou détruire des molécules et ions comme lors de la fabrication de l'eau de Javel. L'électrochimie sert aussi à éviter certains phénomènes naturels comme la corrosion, et peut alors chercher à ce que les ions et les électrodes ne réagissent pas.

Dans un grand nombre de cas, les ions sont dissous dans un fluide, souvent de l'eau, appelé électrolyte. Un même objet peut contenir différents électrolytes et des interfaces pour séparer les espèces. Il y a de nombreux phénomènes à l'œuvre, entre autres : des phénomènes électriques, des réactions chimiques au niveau des électrodes, mais aussi dans le mélange, des mouvements de particules. Les modèles étudiés dans cette thèse ne traitent pas de toutes ces difficultés physiques, mais sont centrés sur le mouvement de particules par diffusion et sous l'effet du champ electrique, donc sur la compréhension de l'électrolyte.

Au vu des échelles caractéristiques des phénomènes physiques observés, la plupart des modèles se font à une échelle macroscopique : on ne compte pas les ions un par un, mais on regarde la concentration de chaque espèce chimique. On dispose ainsi de N espèces dans notre mélange et on notera c_i la concentration de la i-ème espèce. Une question naturelle est : le solvant fait-il partie de ces N espèces ? Si oui, est-il légitime de lui donner un rôle spécifique ? Après un détour par le modèle de Nernst-Planck-Poisson en section 1.1.1 où le solvant n'est pas modélisé on présentera un modèle inspiré de [60, 59, 70] en section 1.1.2, qui lui le prends en

compte et lui donne un rôle particulier.

1.1.1 Nernst-Planck-Poisson

Les premiers modèles d'électrolyte remontent au 19$^{\text{eme}}$ siècle. Ils se basent sur deux équations : le principe de Lavoisier [2], garantis que les espèces chimiques se conservent, autrement dit, du fait de notre choix de ne pas regarder les réactions :

$$\partial_t c_i + \operatorname{div}(J_i) = 0,$$

où J_i représente le flux de matière de la i-eme espèce et reste à déterminer. L'idée est de considérer que du fait des collisions avec le solvant, J_i se décompose en une partie inertielle (transport par un champ te vitesse u) et une partie proportionnelle aux forces s'exerçant sur chaque particule. Par souci de simplicité, on ne considère que deux forces : la force de diffusion qui suit la loi de Fick et la force électrique. Somme toute :

$$J_i = uc_i - D_i \left(\nabla c_i - \frac{z_i c_i e}{k_B T} E \right)$$

avec e la charge de l'élémentaire (charge du proton, l'opposé de la charge de l'électron), z_i le nombre de charges de la i-eme espèce, E le champ électrique k_B la constante de Boltzmann, et T la température supposée constante. Ce modèle néglige les forces magnétiques, et est couramment utilisé sous une forme "stationnaire" avec $u = 0$ et une approximation électrostatique $E = -\nabla\Phi$ calculé par l'équation de Maxwell-Gauss. Somme toute :

$$\begin{aligned} \partial_t c_i + \operatorname{div}(J_i) &= 0, \\ J_i = -D_i \left(\nabla c_i + \frac{z_i c_i e}{k_B T} \nabla \Phi \right), \\ -\epsilon_0 \epsilon_r \, \Delta \, \Phi &= \sum_{i=1}^{N} N_A z_i c_i e. \end{aligned}$$

On préfère travailler sous une forme adimensionnée $c \to c/c_{\text{ref}}$, $J_i \to J_i/c_{\text{ref}}$, $\Phi \to \frac{e}{k_B T}\Phi$:

$$\partial_t c_i + \operatorname{div}(J_i) = 0, \qquad \forall i \in [\![1, N]\!], \tag{NPP.a}$$

$$J_i = -D_i \left(\nabla c_i + z_i c_i \nabla \Phi \right), \qquad \forall i \in [\![1, N]\!], \tag{NPP.b}$$

$$-\lambda^2 \, \Delta \, \Phi = \sum_{i=1}^{N} z_i c_i, \tag{NPP.c}$$

2. « Rien ne se perd, rien ne se crée »

où $\lambda = \left(\frac{\epsilon_0 \epsilon_r k_B T}{N_A e^2 c_{\text{ref}}}\right)^{\frac{1}{2}}$ est la longueur de Debye. Le flux (NPP.b) peut se réécrire :

$$J_i = -D_i c_i \nabla \left(\log(c_i) + z_i \Phi\right), \qquad \forall i \in [\![1, N]\!]. \tag{1.1.1}$$

On appelle ainsi $\log(c_i) + z_i\Phi$ le potentiel électrochimique, et on parlera d'équilibre thermique lorsque ce potentiel est constant.

Ce modèle dispose d'un fonctionnelle de Lyapunov. Avec des conditions de bord adéquates :

$$H = \int \sum_{i=1}^{N} c_i \log c_i + \frac{\lambda^2}{2} |\nabla \Phi|^2, \tag{1.1.2}$$

décroît le long des trajectoire, et on a :

$$\frac{dH}{dt} = -\int_\Omega D_i c_i \nabla \left(\log(c_i) + z_i \Phi\right)^2. \tag{1.1.3}$$

Dans l'expression de H, le terme $\frac{\lambda^2}{2}|\nabla\Phi|^2$ correspond aussi à $\frac{1}{2}\sum z_i c_i \Phi$, et s'interprète comme une énergie électrique. En revanche, le terme $\sum c_i \log c_i$ n'a pas une telle interprétation énergétique. Il peut être assimilé à une approximation de l'entropie de mélange : $-\sum_{i=0}^{N} c_i \log x_i$ avec $x_i = \frac{c_i}{\sum_{i=0}^{N} c_i}$ la fraction molaire de l'espèce i et c_0 la concentration de solvant. Cette approximation est valide dans le cadre fortement dilué $c_i << c_0$. Ce modèle est très utilisé en pratique du fait de sa simplicité et de l'existence de schémas numériques performants (voir Section 1.2). La limite pratique principale est l'absence de couplage mécanique qui permet des concentrations arbitrairement grandes, et en pratique démesurément grandes pour des différences de potentiel relativement faibles.

1.1.2 Au-delà de Nernst-Planck-Poisson

Pour ces raisons, et d'autres considérations physiques, Wolfgang Dreyer, Clemens Guhlke et Rüdiger Müller ont proposé en 2013 un couplage de la diffusion, des forces électriques, et de la mécanique [60] basé sur la thermodynamique hors équilibre [55]. En 2014 Wolfgang Dreyer, Clemens Guhlke et Manuel Landstorfer ont étendu ce modèle à des espèces de volumes molaires différents [59]. Par souci de simplicité on ajoutera les hypothèses faites par Jürgen Fuhrmann pour la simulation numérique dans [70], on négligera aussi les effets de polarisation et la contribution de l'état de référence à l'énergie libre. Posons pour chaque espèce M_i sa masse molaire, v_i son volume molaire supposé constant, et z_i son nombre de charges (à un facteur $N_A e$ près). L'équation de conservation de chaque espèce et

la préservation des moments (en supposant le mélange non visqueux) donnent :

$$\partial_t M_i c_i + \operatorname{div}(M_i c_i u + J_i) = 0, \qquad \forall i \in [\![0, N]\!], \tag{1.1.4}$$

$$\partial_t \rho u + \operatorname{div}(\rho u \otimes u - \sigma) = zE, \tag{1.1.5}$$

$$\epsilon_0 \operatorname{div}(E) = z, \qquad E = -\nabla\Phi, \tag{1.1.6}$$

où u est la vitesse macroscopique (ici la vitesse barycentrique), σ est le tenseur des contraintes, $z = \sum_{i=0}^{N} z_i c_i$ la charge, E le champ électrique, Φ le potentiel électrique, $\rho = \sum_{i=0}^{N} M_i c_i$ la densité du fluide, et J_i le flux de diffusion massique. Pour que l'équation (1.1.5) traduise la conservation des moments, on a scindé le flux de chaque espèce en une part convective $M_i c_i u$ et une part diffusive J_i qui doit satisfaire

$$\sum_{i=0}^{N} J_i = 0. \tag{1.1.7}$$

Un des flux étant défini par cette équation, on calculera le flux de diffusion du solvant à l'aide des N autres flux qui restent à définir.

Ce système dispose d'une énergie libre H composée de trois termes : une énergie cinétique $\frac{\rho u^2}{2}$, une énergie électrique $\frac{1}{2} z\Phi$, et une énergie microscopique $H_{\text{micro}}(c_1, ..., c_N)$. Dans ce cadre, le second principe de la thermodynamique se traduit de la façon suivante :

$$0 \leq \xi := -\frac{1}{T} \sum_{i=0}^{N} J_i \cdot \left(-\frac{z_i}{M_i} E + \frac{1}{M_i} \nabla \mu_i \right), \tag{1.1.8}$$

où μ_i, le potentiel chimique est calculé à partir de l'énergie libre $\Psi(c_0, c_1, \ldots, c_N)$ via :

$$\mu_i = \frac{\partial H_{\text{micro}}}{\partial c_i}.$$

En éliminant le flux du solvant par $J_0 = -\sum_{i=1}^{N} J_i$, on a :

$$\xi := -\frac{1}{T} \sum_{i=0}^{N} J_i \cdot \left((\frac{z_0}{M_0} - \frac{z_i}{M_i}) E + \frac{1}{M_i} \nabla \mu_i - \frac{1}{M_0} \nabla \mu_0 \right),$$

où D_i est le coefficient de diffusion de la loi de Fick.

On fait le choix de flux suivant qui est bien compatible avec la positivité de ξ :

$$J_i = -M_i D_i c_i \nabla \left((\mu_i - \frac{M_i}{M_0} \mu_0) + (z_i - \frac{M_i}{M_0} z_0) \Phi \right). \tag{1.1.9}$$

Le stress σ est également fixé par les lois de la thermodynamique. Grâce à

l'absence de polarisation, ce tenseur est scalaire (proportionnel à l'identité) et on a :

$$\sigma = (\Psi - \sum_{i=0}^{N} c_i \mu_i) I,$$

d'où :

$$\operatorname{div} \sigma = -\sum_{i=0}^{N} c_i \nabla \mu_i$$

Il nous suffit de déterminer l'énergie libre microscopique pour clore le système. Pour cela, on pose :

$$H_{\text{micro}} = H_{\text{mélange}} + H_{\text{mécanique}}$$

avec :

$$\Psi_{\text{mélange}} = k_B T \sum_{i=0}^{N} c_i \log \frac{c_i}{\overline{c}}, \qquad \overline{c} = \sum_{i=0}^{N} c_i.$$

Le terme $\Psi_{\text{mécanique}}$ est plus complexe. On pose

$$r = \sum_{i=0}^{N} c_i v_i, \quad p = K(r-1), \quad \Psi_{\text{mécanique}} = Kr\log(r) - p,$$

où r est le taux de compression, p la pression, et $K \geq 0$ le module supposé indépendant de la composition du mélange. On peut remarquer que $\Psi_{\text{mécanique}}$ vue comme une fonction de r est convexe et nulle en 1. On obtient ainsi :

$$\mu_i = k_B T \log \frac{c_i}{\overline{c}} + v_i K \log(r).$$

Dans la pratique, les électrolytes peuvent souvent être considérés comme incompressibles, autrement dit, K est très grand. Dans cette limite, si à l'état initial on a $r = 1$, alors l'énergie libre initiale H est finie. Elle reste donc finie, c'est-à-dire qu'à tout instant on a $r = 1$. On peut alors développer $log(r)$ au premier ordre et on obtient :

$$\mu_i = k_B T \log \frac{c_i}{\overline{c}} + v_i p.$$

Finalement, sous ces hypothèses et en introduisant $N_i = J_i/M_i$ le flux molaire, le système considéré se résume ainsi :

$$\partial_t c_i + \operatorname{div}(c_i u + N_i) = 0, \qquad \forall i \in [\![0, N]\!],$$

$$N_i = -D_i c_i \nabla \left((\mu_i - \frac{M_i}{M_0}\mu_0) + (z_i - \frac{M_i}{M_0} z_0)\Phi \right) \qquad \forall i \in [\![1, N]\!],$$

$$\mu_i = k_B T \log \frac{c_i}{\overline{c}} + v_i p, \qquad \forall i \in [\![0, N]\!],$$

$$\sum_{i=0}^{N} c_i v_i = 1, \qquad \sum_{i=0}^{N} J_i = 0, \qquad -\epsilon_0 \, \Delta \, \Phi = \sum_{i=0}^{N} z_i c_i,$$

$$\partial_t \rho u + \operatorname{div} (\rho u \otimes u) = - \sum_{i=0}^{N} c_i \nabla (\mu_i + z_i \Phi).$$

Le traitement de termes du type $\operatorname{div} u \otimes u$ est très difficile mathématiquement, aussi on aimerait faire l'hypothèse $u = 0$. Cela pose un premier problème physique : l'hypothèse d'incompressibilité indique que la vitesse volumique $u_v = u + \sum_{i=0}^{N} v_i N_i$ est à divergence nulle. En dimension un, sous ces deux hypothèses un mélange à deux espèces ($N = 1$) sera donc au repos si les masses et volumes molaires ne sont pas proportionnels.

On pourrait s'affranchir de ce problème en remplaçant la vitesse barycentrique dans (1.1.5) par la vitesse volumique, cela remplacerait les ratios de masses molaires par des ratios de masses volumiques. On aurait ainsi :

$$N_i = -D_i c_i \nabla \left(k_B T (\log \frac{c_i}{\overline{c}} - \frac{v_i}{v_0} \log \frac{c_0}{\overline{c}}) + (z_i - \frac{v_i}{v_0} z_0) \Phi \right), \qquad \forall i \in [\![1, N]\!]$$

par une élimination immédiate de la pression. Cette idée est proche du modèle proposé en section 4.3.2 du chapitre 4, mais revient à abandonner le principe physique de conservation des moments. Le découplage d'avec la pression amenant à étudier le couplage électro-diffusif séparément des aspects mécaniques, tout en bénéficiant de l'équation $\sum_{i=0}^{N} c_i v_i = 1$, nous nous intéresserons tout de même au modèle adimensionné suivant dans le chapitre 3 :

$$\partial_t c_i + \operatorname{div} (N_i) = 0, \quad N_i = -D_i c_i \nabla \left(h_i + (z_i - \frac{v_i}{v_0} z_0) \Phi \right), \quad 1 \leq i \leq N, \tag{1.1.10a}$$

$$h_i = \log \frac{c_i}{\overline{c}} - \frac{v_i}{v_0} \log \frac{c_0}{\overline{c}}, \qquad \forall i \in [\![1, N]\!], \tag{1.1.10b}$$

$$\sum_{i=0}^{N} c_i v_i = 1, \tag{1.1.10c}$$

$$- \lambda^2 \, \Delta \, \Phi = \sum_{i=1}^{N} z_i c_i. \tag{1.1.10d}$$

Ce nouveau modèle revient à modifier le modèle de Nernst-Planck-Poisson en remplaçant $\log(c_i)$ de (1.1.1) par h_i, et à ajouter le solvant à la modélisation en lui

faisant remplir les vides.

1.1.3 Détournement du modèle d'électrolyte

Le remplacement de $\log c_i$ par h_i implique deux difficultés mathématiques. La première est l'apparition d'une non-linéarité, la seconde est l'apparition de croisements entre les concentrations. Pour éviter ces croisements, on considère un électrolyte contenant deux espèces, donc $N = 1$. On supposera également une égalité entre les volumes molaires :

$$v_1 = v_0 \qquad \text{d'où} \qquad h_1 = \log \frac{c_1}{\frac{1}{v_0} - c_1}.$$

Quitte à changer d'unités, on pose $v_0 = 1$. Pour que le couplage ait un intérêt, on suppose $0 \neq z_{\text{eff}} := z_1 - z_0$. Le cas particulier $z_1 = -z_0$ correspond au modèle de liquide ionique proposé en [71] à la section 4.3. Dans le chapitre 2, on ajoute un terme source à l'équation de Poisson. On obtient ainsi :

$$\begin{aligned} \partial_t c - \operatorname{div} Dc\nabla(\log \frac{c}{1-c} + z\Phi) &= 0, \\ -\lambda^2 \Delta \Phi &= zc + c^{\text{dop}}. \end{aligned}$$

Cette addition permet à peu de frais de considérer ce modèle comme un système de dérive-diffusion unipolaire, c'est-à-dire un réseau de charges fixes au milieu desquelles évoluent des charges mobiles. Cette situation est notamment présente dans les semi-conducteurs et certains électrolytes solides. En pratique, il semble qu'il existe bien des matériaux pour lesquels la relation entre la concentration du porteur mobile c et son potentiel chimique h est proche de $h = \log \frac{c}{c_{\text{ref}} - c}$ [124, 1]. Ces matériaux sont donc couverts par l'étude menée au chapitre 2.

1.1.4 Vers d'autres modèles d'électrolytes

Dans le chapitre 4 on cherche à construire des modèles proches de celui défini en section 1.1.2 en explicitant le mécanisme de dissipation à l'œuvre. Plus formellement, l'idée physique est de prescrire ξ, la dissipation d'énergie libre, dans l'équation (1.1.8) pour en déduire l'expression des flux. Dans la section 1.1.2, on postule une expression des flux pour constater ensuite qu'ils satisfont au second principe. Ici on postulera une valeur de ξ pour en déduire l'expression des flux.

La technique utilisée est celle de la modélisation variationnelle [113]. Ce cadre est adapté dès que les effets inertiels sont négligeables. Cela revient à remplacer l'équation de conservation des moments (1.1.5) par une condition d'optimalité sur les flux. On présente ici deux dérivations de l'équation de la chaleur (diffusion

suivant la loi de Fick) pour exposer la méthode. Une deuxième présentation de la méthode est disponible au chapitre 4 avec une dérivation du modèle de Nernst-Planck-Poisson différente de celle présentée en section 1.1.1. Ces deux dérivations sont inspirées de [110]. Enfin en Section 1.1.4.c on montrera le lien entre les grandeurs issues de la modélisation variationnelle et la décroissance de l'énergie libre notée ici E.

1.1.4.a Une première dérivation de la loi de Fick

La modélisation variationnelle repose sur choix de quatre éléments :

— Un espace d'états admissible : $\mathcal{Z}$ représentant les inconnues modélisées. Dans ce premier exemple, nous prendrons

$$\mathcal{Z} = L^2(\Omega, \mathbb{R}^N),$$

où Ω est le domaine modélisé et N le nombre d'espèces. On peut remarquer que nous n'imposons pas la positivité des concentrations.

— Une énergie : $E : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ qui sera une fonctionnelle de Lyapunov. Ici nous prendrons

$$E(c_1, ..., c_N) = \int_\Omega \sum_{i=1}^{N} \frac{c_i^2}{2}.$$

— Un mécanisme de transformations possibles. Bien que le cadre théorique soit plus large, nous ne considérerons que des transformations qui garantissent la conservation de la masse :

$$\partial_t c_i + \operatorname{div} J_i = 0. \tag{1.1.11}$$

Il reste possible de restreindre les flux admissibles $\mathcal{J}$, ainsi pour des conditions de bords de Neumann homogènes on posera :

$$\mathcal{J} = \left\{ (J_1, \dots, J_N) \in H_{\text{div}}(\Omega, \mathbb{R}^d)^N \mid \forall i, J_i \cdot n = 0 \text{ sur } \partial\Omega \right\}.$$

— Une fonction de coût de transformation $\Psi : \mathcal{J} \to \overline{\mathbb{R}}$. Ici :

$$\Psi(J) = \int_\Omega \sum_{i=1}^{N} \frac{|J_i|^2}{2D_i}.$$

On demande souvent que Ψ soit minimale en zéro et convexe. Bien que ce ne soit pas le cas dans cet exemple introductif, Ψ peut dépendre de l'état du système. Elle est parfois appelée potentiel de dissipation du fait des liens

avec la décroissance de E mis en exergue en section 1.1.4.c. Par un léger abus de langage, on l'appellera plus simplement "dissipation" bien qu'il ne s'agisse pas directement de ξ tel que définie dans l'équation (1.1.8).

L'idée de la modélisation variationnelle est d'ajouter au mécanisme de transformation (1.1.11) une condition d'optimalité des flux :

$$J_c \in \underset{J\in\mathcal{J}}{\operatorname{argmin}}\, \Psi(J) + \langle \frac{\delta E}{\delta c}, -\operatorname{div} J\rangle, \tag{1.1.12}$$

où $\frac{\delta E}{\delta c}$ est la dérivée de E.

Pour avoir finitude et unicité de cet argmin, il suffit d'ajouter une hypothèse de coercivité sur Ψ. La minimalité de Ψ en zéro permet de garantir la décroissance de E, comme on le verra en section 1.1.4.c.

Sous la forme actuelle, il n'est pas évident que les flux suivent la loi de Fick :

$$J_i = -D_i \nabla c_i.$$

Pour expliciter cela, on cherche les points critiques de $\Psi(J) + \langle \frac{\delta E}{\delta c}, -\operatorname{div} J\rangle$, c'est-à-dire les points où la dérivée en J est nulle. Pour la dissipation, on a pour tout H :

$$\frac{\delta \Psi}{\delta J} \cdot H = \int_\Omega \sum_{i=1}^{N} \frac{J_i \cdot H_i}{D_i}.$$

En intégrant par partie, on a :

$$\frac{\delta}{\delta J}\left(\Psi(J) + \frac{\delta E}{\delta c}, -\operatorname{div} J\rangle\right) = \left(\frac{J_1}{D_1} + \nabla c_1, ..., \frac{J_N}{D_N} + \nabla c_N\right).$$

On a donc un unique point critique, et il suit la loi de Fick ! Du fait de la régularité et de la coercivité de la fonctionnelle considérée, ce point est un minimum.

L'énergie considérée, ici la norme L^2 , est plus adaptée à des vitesses qu'à des concentrations, et on lui préfère souvent pour les modèles de diffusion l'énergie de mélange : $c_i \log \frac{c_i}{c_{\text{ref}}}$. En effet introduire le couplage électrique à partir de cette "énergie de mélange" et ce mécanisme de dissipation donne le flux :

$$D_i \left(\nabla c_i + z_i \nabla \Phi\right),$$

autrement dit un terme de dérive électrique indépendant de la concentration, ce qui a peu de sens physique. Il est aussi possible de retrouver la loi de Fick en prenant l'approximation usuelle de l'énergie de mélange $c_i \log \frac{c_i}{c_{\text{ref}}}$, avec c_{ref} constant. C'est le sujet de la prochaine section.

1.1.4.b Une autre dérivation de la loi de Fick

On considère désormais :

$$\mathcal{Z} = \Omega \to [0, +\infty)^N, \qquad E : c \in \mathcal{Z} \mapsto \int_\Omega \sum_{i=1}^N c_i \log(c_i), \tag{1.1.13}$$

$$\mathcal{J} = \left\{ (J_1, \dots, J_N) \in H_{\text{div}}(\Omega, \mathbb{R}^d)^N \mid \forall i, J_i \cdot n = 0 \text{ sur } \partial\Omega \right\}, \tag{1.1.14}$$

$$\Psi : J \in \mathcal{J} \mapsto \int_\Omega \sum_{i=1}^N \frac{|J_i|^2}{2D_i c_i}. \tag{1.1.15}$$

Ce choix de modélisation correspond à l'exemple donné dans [113] section 2.3. Il y a deux façons de tenir compte du bord de $\mathcal{Z}$: une première façons est de ne rien faire et de dire si on en sort le modèle n'est plus valide. Une autre manière de faire est de préciser le sens de $\frac{|J_i|^2}{c_i}$ pour J_i et c_i nul. En posant zéro dans ce cas et $+\infty$ dans les autres cas où c_i est négatif ou nul, on retrouve la fonction de Benamou-Brenier [13]. Comme nous savons que l'équation de la chaleur vérifie le principe du maximum [3], nous choisirons la première option. Dans le chapitre 4 en section 4.2.2 on propose une façon d'intégrer des contraintes qui sans régler cette ambiguïté garantis la positivité et s'intègre mieux dans la formulation du problème.

On procède donc comme précédemment. La dérivée de Ψ par rapport à J est $\frac{J_1}{D_1 c_1}, \dots, \frac{J_N}{D_N c_N}$, celle de $\langle \frac{\delta E}{\delta c}, -\operatorname{div} J \rangle$ est $\nabla \log(c)$, on a donc :

$$J_i = -D_i c_i \nabla \log c_i.$$

Il suffit désormais d'utiliser le théorème de dérivation des fonctions composées (chain rule en anglais) pour obtenir la loi de Fick désirée. Dans le Chapitre 4, on étends cette énergie et ce potentiel de dissipation à un couplage électrique qui nous donne le système de Nernst-Planck-Poisson.

1.1.4.c Quelques généralités

Dans les deux exemples précédents, on peut remarquer que :

$$\frac{dE(c)}{dt} = -2\Psi(J_c) \leq 0.$$

Ici, $-\frac{dE(c)}{dt}$ correspond à ξ de l'équation (1.1.8). On s'intéresse dans cette section à généraliser cette proximité entre potentiel de dissipation et dissipation réelle.

3. i.e préserve les bornes de la condition initiale.

Pour ce faire, on considère l'ensemble $\mathcal{Z}$, la fonction E, et l'espace vectoriel $\mathcal{J}$ assez généraux. Par souci de simplicité, on ne précisera pas les hypothèses précises de régularité et de compatibilité de ces choix, on procédera donc très formellement. Pour Ψ, on suppose qu'elle est convexe et coercive pour garantir l'existence et l'unicité d'une solution au problème de minimisation (1.1.12). On suppose aussi qu'elle est minimale en zéro. Sous ces hypothèses, E est une fonctionnelle de Lyapunov, autrement dit $t \mapsto E(c(t))$ est décroissante. On a en effet, au moins formellement :

$$\frac{dE(c)}{dt} = \langle \frac{\delta E}{\delta c}, -\operatorname{div} J_c \rangle.$$

Par minimalité de Ψ en zéro, puis en comparant la valeur de $\Psi(J) + \langle \frac{\delta E}{\delta c}, -\operatorname{div} J \rangle$ en J_c et en 0, on a :

$$\frac{dE(c)}{dt} \leq \Psi(J_c) - \Psi(0) + \langle \frac{\delta E}{\delta c}, -\operatorname{div} J_c \rangle \leq 0.$$

Il est possible d'aller plus loin et de relier effectivement Ψ à la décroissance de l'énergie. Pour cela on introduit μ la dérivée de E, et on constate comme précédemment que :

$$\langle \frac{\delta E}{\delta c}, -\operatorname{div} J_c \rangle = \langle \nabla \mu, J_c \rangle.$$

On a ainsi en ajoutant et retranchant $\Psi(J_c)$:

$$\frac{dE(c)}{dt} = -\Psi(J_c) + \min_{J \in \mathcal{J}} \Psi(J) - \langle -\nabla \mu, J \rangle.$$

En introduisant la transformée de Legendre Ψ^*, on a :

$$\frac{dE(c)}{dt} = -\Psi(J_c) - \Psi^*(-\nabla \mu),$$

où $\nabla \mu$ est $(\nabla \mu_1, \nabla \mu_2, .. \nabla \mu_N)$. Ce résultat est extrêmement général du fait de la faiblesse des hypothèses sur Ψ.

On dispose d'une expression plus simple dans le cas -fréquent- où :

$$\Psi(J) = \int_\Omega \frac{J^T A(c) J}{2}.$$

On peut remarquer qu'on ne dispose pas d'unicité pour A, en effet pour toute matrice anti-symétrique ϵ, $J^T \epsilon J = 0$, d'où $J^T(A(c)+\epsilon)J = J^T A(c) J$. On peut donc sans perdre en généralité supposer A symétrique. Ce cas est appelé cas d'Onsager car le flux et les forces entropiques sont positivement liées [107, 108]. En effet, l'

hypothèse de coercivité sur Ψ indique que A est définie positive, donc inversible. En notant B son inverse, on remarque que :

$$J_c = B\nabla\mu, \qquad \Psi^*(F) = \int_\Omega F^T \frac{B(c)}{2} F,$$

d'où :

$$\frac{dE(c)}{dt} = -\int_\Omega J_c^T A(c) J_c = -2\Psi(c).$$

1.1.5 Fabrication de panneaux solaires

Il existe de nombreux types de panneaux solaires. Sans entrer dans le détail de fonctionnement de chaque type de panneau solaire, le principe est que le rayonnement solaire apporte assez d'énergie à des électrons pour les faire passer de la bande de valence à la bande de conductivité, autrement dit à les séparer de leur noyau d'origine. En empilant une couche dopée positivement, c'est-à-dire avec quelques atomes ayant un électron excédentaire dans leur couche externe et une couche dopée négativement (où les impuretés sont plutôt des atomes auxquels il faudrait ajouter un électron), les électrons seront plus facilement arrachés dans la zone dopée positivement et auront tendance à se lier plus facilement aux éléments de dopage de la couche négative qu'au semi-conducteur ambiant. De même, le déficit d'électron, ou trou, se transmettra de proche en proche et restera davantage sur les atomes ayant une couche externe à neuf électrons [4].

Des couches intermédiaires peuvent être ajoutées pour limiter davantage les recombinaisons, collecter ou injecter les électrons (afin de fermer le circuit de production de courant), protéger des agressions extérieures, etc.

La largeur de chaque couche contrôle non seulement la distance à parcourir pour un électron et un trou nouvellement créés, mais aussi la quantité de lumière captée par chaque couche. Une couche trop épaisse et les électrons et les trous se recombinent dans la couche positive. Une couche trop fine et le rayonnement atteint la couche négative où les électrons sont plus difficiles à arracher. Dans ces deux cas, l'efficacité du panneau solaire est limitée. Il est donc crucial de bien maîtriser l'épaisseur de chaque couche lors de la production du panneau.

Il existe de nombreuses techniques pour s'assurer de l'épaisseur de chaque couche. L'une d'entre elles est le dépôt de vapeur. L'idée est de placer le support sur lequel déposer les couches dans un milieu saturé en vapeurs qui se condenseront sur ce support. Ce processus s'effectue à très basse pression si bien que la vapeur se

4. En comptant les électrons partagés par des liaisons chimiques, dans le cas où le semi-conducteur est le silicone, l'atome aura 5 électrons "en propre", ce pourrait être de l'arsenic ou du phosphore.

transforme immédiatement en solide. Du fait des conditions de température dans le solide, les atomes peuvent échanger de place avec une probabilité non négligeable, ce qui permet de produire des structures cristallines complexes telles que la chalcopyrite des panneaux CIGS [5] ou la kesterite des cellules éponymes [58]. Cela peut aussi amener les différentes couches à se mélanger. Le modèle de mélange proposé ici et étudié numériquement dans le chapitre 5 a été proposé par Athmane Bakhta et Virginie Ehrlacher dans l'annexe A.1 de leur article [11].

Par souci de simplicité, nous faisons la dérivation en dimension deux plutôt qu'en dimension d. Il aurait également été possible de faire la dérivation en dimension 1, ce choix est fait dans [11], mais cela rend plus difficile le passage de la dérivée scalaire au gradient, à la divergence ou à un autre opérateur.

On décompose l'espace $\mathbb{R}^2$ en cellules carrées de côté δx qu'on peut se représenter comme les mailles d'un réseau cristallin. Le temps sera représenté par une succession d'instants successifs distants de δt. Dans chaque cellule se trouvent jusqu'à N espèces différentes et on note $u_i^{k,l,n}$ le nombre moyen d'atomes se trouvant dans la cellule $[k\delta x,(k+1)\delta x[\times[l\delta x,(l+1)\delta x[$ à l'instant $n\delta t$. Entre deux instants, on permet aux atomes d'échanger avec un atome d'une cellule qui partage un côté avec sa cellule actuelle. On note $p_{i,j}\geq 0$ la probabilité pour un atome de l'espèce i d'échanger de place avec un atome de l'espèce j. Nécessairement, $p_{i,j}=p_{j,i}$. On suppose que ces probabilités sont indépendantes de l'état des cellules. En moyenne, pour deux cellules voisines (k,l) et (k',l'), il y a (en développant à l'ordre 1 en $p_{i,j}$) $p_{i,j}u_i^{k,l,n}u_j^{k',l',n}$ échanges entre un élément de l'espèce i de la cellule (k,l) et un élément de l'espèce j de la cellule voisine. Au total, on a donc :

$$u_i^{k,l,n+1}=u_i^{k,l,n}+\sum_{j=1}^{N}p_{i,j}u_j^{k,l,n}(u_i^{k-1,l,n}+u_i^{k+1,l,n}+u_i^{k,l-1,n}+u_i^{k,l+1,n})$$
$$-\sum_{j=1}^{N}p_{i,j}u_i^{k,l,n}(u_j^{k-1,l,n}+u_j^{k+1,l,n}+u_j^{k,l-1,n}+u_j^{k,l+1,n}).$$

On peut noter que même si l'état initial était parfaitement connu avec u entier, ce n'est plus forcément le cas par la suite. On remarque aussi qu'on peut retrancher dans chaque somme le terme $4u_i^{k,l,n}u_j^{k,l,n}$. On peut enfin ignorer les échanges entre

5. On trouvera aussi parfois les noms de Gallite et Roquésite

termes d'une même espèce. Après quelques calculs, on obtient :

$$\frac{u_i^{k,l,n+1} - u_i^{k,l,n}}{\delta t} = \frac{\delta x^2}{\delta t} \sum_{j \neq i} p_{i,j} u_j^{k,l,n} \left(\frac{u_i^{k-1,l,n} - 2u_i^{k,l,n} + u_i^{k+1,l,n}}{\delta x^2} + \frac{u_i^{k,l-1,n} - 2u_i^{k,l,n} + u_i^{k,l+1,n}}{2\delta x^2} \right) - \frac{\delta x^2}{\delta t} \sum_{j \neq i} p_{i,j} u_i^{k,l,n} \left(\frac{u_j^{k-1,l,n} - 2u_j^{k,l,n} + u_j^{k+1,l,n}}{\delta x^2} + \frac{u_j^{k,l-1,n} - 2u_j^{k,l,n} + u_j^{k,l+1,n}}{2\delta x^2} \right).$$

Sous cette forme, on remarque que les fractions présentes approchent ce qu'on aimerait appeler la dérivée de u. Pour pouvoir considérer cela on fait tendre δx vers zéro, non en réduisant la taille des mailles de notre réseau cristallin (ce qui ferait décroître u_i) mais simplement en changeant d'unité de longueur. Pour conserver une limite non triviale, on considère aussi des temps de plus en plus grands de sorte que dans notre nouvelle unité de temps δt soit lui aussi zéro et qu'on aie $\frac{\delta x^2}{\delta t} \to \alpha$. Dans ces changements d'unité, $p_{i,j}$ lui reste inchangé. On a alors :

$$\partial_t u_i = \sum_{j \neq i} a_{i,j} \big(u_j (\partial_{xx} u_i + \partial_{yy} u_i) - u_i (\partial_{xx} u_j + \partial_{yy} u_j) \big),$$

où $a_{i,j}$ est égal à $\alpha p_{i,j}$. En remarquant de $\operatorname{div} u_i \nabla u_j = \nabla u_i \cdot \nabla u_j + u_i \, \Delta u_j$, on a :

$$\partial_t u_i + \operatorname{div} \left(\sum_{j \neq i} a_{i,j} \big(u_i \nabla u_j - u_j \nabla u_i \big) \right) = 0.$$

Cette équation est étudiée dans le chapitre 4. Le fait que $p_{i,j}$ ne dépende pas de la face considérée est une hypothèse très forte dans le cas d'un matériau cristallin. La prise en compte d'une anisotropie est incompatible avec les méthodes numériques étudiées dans cette thèse et que nous présentons maintenant.

1.2 Schémas

Pour simuler les modèles présentés dans la section précédente, nous utilisons des schémas volumes finis à deux points en espace et la méthode d'Euler implicite en temps. Le principe de la méthode des volumes finis est d'approcher des intégrales de l'équation sous forme forte. Pour une équation de conservation de la masse, on cherche donc sur chaque volume de contrôle à approcher la quantité inclue dans le

volume (ou intégrale de la solution) et l'intégrale des flux sur le bord du domaine [6]. La méthode d'approximation à deux points se fixe comme limite pour approcher l'intégrale d'un flux entre deux cellules de contrôle en utilisant uniquement les valeurs de la solution dans ces deux cellules. Cette méthode, largement utilisée dans l'industrie, a l'intérêt de préserver les propriétés physiques des solutions telles que la conservation de la masse et la positivité des concentrations [67]. Elle nécessite cependant de choisir un maillage adapté. La définition de tels maillages est donnée en section 1.2.1 ainsi que des notations usuelles. Dans la section suivante 1.2.2, on propose deux schémas pour l'équation de Nernst-Planck-Poisson présentée en section 1.1.1 dans le cas à une inconnue de concentration ($N = 1, D = 1, z = 1, \lambda = 1$). Dans ce cadre là elle est parfois aussi appelée équation de dérive-diffusion. Enfin en section 1.2.3 on propose des pistes pour l'analyse numérique de ces schémas en suivant des idées clefs des chapitres 2 et 3.

Pour ce qui suit on considère $\Omega \subset \mathbb{R}^d$ un ouvert polytopal connexe non vide dans lequel on considère l'équation d'évolution :

$$\partial_t c + \operatorname{div} F = 0, \tag{1.2.1}$$
$$F = -c\nabla\left(\log c + \phi\right) \tag{1.2.2}$$
$$-\Delta\,\Phi = c. \tag{1.2.3}$$

assortie d'une condition initiale $c^0 > 0$, de conditions de bords de Neumann homogènes pour $c : F \cdot n = 0$ sur $\partial\Omega$ et pour $\Phi : \nabla\Phi \cdot n = 0$ sur $\Gamma_N \subset \partial\Omega$. Sur le reste de la frontière $\Gamma_D = \partial\Omega \setminus \Gamma_N$, on impose une condition de Dirichlet inhomogène : $\Phi = \Phi^D$, où $\Phi^D \in H^1(\Omega)$ ne dépends pas du temps. On se donne également un intervalle de temps de simulation fini $[0, T]$.

1.2.1 Maillage

Les schémas numériques à deux points (TPFA) nécessitent une condition d'orthogonalité sur le maillage [89, 68]. On précise cette condition dans la définition suivante, compatible en tout point avec celles des section 2.2.1, 3.2.1, 5.2.1 et 4.4. De même les notations de l'introduction correspondent avec celles des chapitres suivants.

Definition 1. *Un maillage admissible de Ω est un triplet cellules, faces, centres $\left(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}}\right)$ satisfaisant les conditions suivantes.*

(i) L'ensemble $\mathcal{T}$ est fini et chaque cellule $K \in \mathcal{T}$ est un polytope convexe ouvert et non vide. On suppose de plus les cellules disjointes et que leur adhérence

6. C'est un cas d'application du théoreme de Green-Ostrogradski

recouvre Ω *:*

$$K \cap L = \emptyset \quad si\ K, L \in \mathcal{T}\ avec\ K \neq L, \qquad et \quad \bigcup_{K\in\mathcal{T}} \overline{K} = \overline{\Omega}.$$

(ii) *Chaque face* $\sigma \in \mathcal{E}$ *est fermée, et incluse dans un hyperplan de* $\mathbb{R}^d$ *et de mesure non nulle* $m_\sigma = \mathcal{H}^{d-1}(\sigma) > 0$*. On suppose enfin que les faces n'ont qu'au plus des arrêtes communes, autrement dit* $\mathcal{H}^{d-1}(\sigma \cap \sigma') = 0$ *pour tout* $\sigma, \sigma' \in \mathcal{E}$ *tels que* $\sigma' \neq \sigma$*.*

(iii) *Pour chaque cellule* $K \in \mathcal{T}$ *on dispose d'un ensemble de faces* $\mathcal{E}_K$ *inclus dans* $\mathcal{E}$ *tel que l'union de ces faces soit la frontière de la cellule :* $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \sigma$*.*

(iv) *Chaque face* $\sigma \in \mathcal{E}$ *est la face d'au moins une cellule :* $\exists K \in \mathcal{T}, \sigma \in \mathcal{E}_K$*. Autrement dit :* $\mathcal{E} = \bigcup_{K\in\mathcal{T}} \mathcal{E}_K$*.*

(v) *Enfin, on suppose que les faces ne sont pas inutilement fractionnées : pour chaque cellules* $K, L \in \mathcal{T}$ *distinctes, l'intersection de leurs adhérences est donc soit une face* $\sigma = K|L \in \mathcal{E}$*, soit un ensemble de mesure* $(d-1)$ *nulle. Dans le premier cas, on parle de cellules voisines.*

(vi) *les centres des cellules sont à l'interieur de leur cellule* $(x_K \in K, \forall K \in \mathcal{T})$ *et satisfont à une condition d'orthogonalité : si* K *et* L *sont des cellules voisines, alors le vecteur* $x_K - x_L$ *est orthogonal à la face* $K|L$*.*

(vii) *cette condition s'étend aux faces qui sont à la frontière de* Ω*. Pour celles-ci on suppose qu'on dispose de* $x_\sigma \in \sigma$ *un centre de face tel que* $x_K - x_\sigma$ *soit orthogonal à* σ

(viii) *on suppose enfin que notre maillage est adapté aux conditions de bords. En notant* $\mathcal{E}_{int} = \bigcup_{K\in\mathcal{T}} \bigcup_{L\in\mathcal{T}\setminus\{K\}} \mathcal{E}_K \cap \mathcal{E}_L$ *l'ensemble des faces internes*[7]*,* $\mathcal{E}_{ext} = \mathcal{E} \setminus \mathcal{E}_{int}$*, l'ensemble des faces externes, on dispose de* $\mathcal{E}_N, \mathcal{E}_D$ *une partition de* $\mathcal{E}_{ext}$ *telle que tout* $\sigma \in \mathcal{E}_D$ *soit dans l'adhérence de* Γ_D *et tout* $\sigma \in \mathcal{E}_N$ *dans celle de* Γ_N*.*

Pour des maillages triangulaires, les centres des cellules doivent être les centres des cercles circonscrits, autrement dit, notre maillage ne doit pas avoir d'angle droit ou obtus. Cette notion est plus restrictive que les maillages de Delaunay. Pour de tels maillages le cercle circonscrit ne doit pas contenir de sommets d'autres triangles, mais certains triangles peuvent avoir un angle obtus. Cette définition peut aussi s'interpréter en $x_K - x_L$ est orienté de K vers L. Par souci de simplicité, on gardera l'hypothèse (vi) inchangée. Une première question assez naturelle est l'existence de maillage admissibles. On peut en construire assez facilement en définissant d'abord les centres x_K puis les cellules comme étant les points de Ω les

7. Du fait des hypothèses précédentes, $\mathcal{E}_K \cap \mathcal{E}_L$ est au plus un singleton et l'union $\bigcup_{L\in\mathcal{T}\setminus\{K\}} \mathcal{E}_K \cap \mathcal{E}_L$ est disjointe.

plus proches de leur centre. Pour l'adaptation aux conditions de bord, on perturbe légèrement le premier choix de centre, et on obtiens un maillage admissible. Un tel maillage est appelé maillage de Voronoï. Le maillage dual d'un maillage de Voronoï correspond au maillage obtenu en considérant les sommets de se maillage comme des centres de mailles et les segments $[x_K, x_L]$ comme des arrêtes (au sens de sous-face de dimension 1). En dimension deux ce maillage dual est un maillage de Delaunay dans le sens plus général mentionné précédemment (qui autorise donc les angles droits). Ces maillages ont de nombreux intérêts [80] dont celui de permettre une reconstruction P_1 sur le maillage dual d'une solution connue sur chaque centre de maille.

Par souci de simplicité, on mentionnera parfois l'ensemble du triplet de maillage par le seul $\mathcal{T}$. Pour les besoin des différents calculs, on introduit m_K la mesure de la maille K,

$$d_\sigma = \begin{cases} |x_K - x_L| & \text{si } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ |x_K - x_\sigma| & \text{si } \sigma \in \mathcal{E}_{\text{ext}}, \end{cases} \qquad \text{and} \quad \tau_\sigma = \frac{m_\sigma}{d_\sigma}.$$

On introduit enfin la taille $h\mathcal{T}$ et la régularité $\zeta_\mathcal{T}$ du maillage :

$$h_\mathcal{T} = \max_{K\in\mathcal{T}} \operatorname{diam}(K), \qquad \zeta_\mathcal{T} = \min_{K\in\mathcal{T}} \min_{\sigma\in\mathcal{E}_K} \frac{\mathrm{d}(x_K, \sigma)}{d_\sigma}.$$

Pour la discrétisation en temps, on découpe simplement $[0, T]$ en N_T intervalles de longueur $\Delta t = T/N_T$ et on introduit $t_n = n\Delta t$.

1.2.2 Quelques schémas pour Nernst-Planck-Poisson

Il s'agit désormais d'exploiter ce maillage pour approcher les valeurs moyennes des solutions. En effet, en intégrant par partie les équations [8] (1.2.1), (1.2.3) sur une cellule K on obtient :

$$\begin{aligned} \partial_t \int_K c + \sum_{\sigma\in\mathcal{E}_K} \int_\sigma F \cdot n = 0, \\ -\sum_{\sigma\in\mathcal{E}_K} \int_\sigma \nabla\Phi \cdot n = \int_K c, \end{aligned}$$

où n est la normale sortante de la maille K. L'idée du schéma deux points est d'approcher l'intégrale sur chaque face interne $\sigma = K|L$ en utilisant uniquement les valeur -approchée elles aussi- des intégrales sur chaque cellule : $\int_K c$, $\int_K \Phi$, $\int_L c$,

8. ici le Théoreme de Green-Ostrogradski

$\int_L \Phi$. On note ainsi c_K^n la valeur approchée de $⨍_K c(t_n)$ la valeur moyenne sur la maille K, de même $\Phi_K^n \simeq ⨍_K \Phi(t_n)$. On approche assez naturellement

$$\int_{K|L} \nabla\Phi(t_n)\cdot n \simeq m_\sigma \frac{\Phi_L^n - \Phi_K^n}{d_\sigma} = \tau_\sigma\left(\Phi_L^n - \Phi_K^n\right).$$

Afin de pouvoir traiter les conditions de bord avec une formule compacte, on note :

$$c_{K\sigma}^n = \begin{cases} c_L^n & \text{si } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ c_K^n & \text{si } \sigma \in \mathcal{E}_{\text{ext}}, \end{cases} \qquad \Phi_{K\sigma}^n = \begin{cases} \Phi_L^n & \text{si } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ \Phi_K^n & \text{si } \sigma \in \mathcal{E}_N, \\ \Phi_\sigma^D = ⨍_\sigma \Phi^D & \text{si } \sigma \in \mathcal{E}_D. \end{cases}$$

On introduit aussi $\boldsymbol{c}^n$ le vecteur de $\mathbb{R}^{\mathcal{T}}$ composé des c_K^n, $\boldsymbol{c}$ le vecteur des $\boldsymbol{c}^n$, et de même $\boldsymbol{\Phi}^n$ et $\boldsymbol{\Phi}$ représentent notre discrétisation du potentiel. On notera en gras les vecteurs dépendant du maillage. Dans certains chapitres, on notera en majuscule les vecteurs en le nombre de concentrations, et on ne distingue pas les vecteurs de $\mathbb{R}^d$ par une notation particulière. Dans ce même objectif de simplification des expressions, on pose :

$$D_{K\sigma}\boldsymbol{u} = u_{K\sigma} - u_K, \quad D_\sigma \boldsymbol{u} = |D_{K\sigma}\boldsymbol{u}|, \qquad \forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_K,$$

ou $\boldsymbol{u}$ est une fonction de $\boldsymbol{c}^n$ ou $\boldsymbol{\Phi}^n$. Avec ces notations, et en étendant l'approximation naturelle au bord Dirichlet, on approche l'équation (1.2.3) pour le potentiel par :

$$-\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D_{K\sigma}\boldsymbol{\Phi}^n = m_K c_K^n. \tag{1.2.4}$$

Pour la conservation de la masse (1.2.1), on utilise le schéma d'Euler implicite en temps, on a donc :

$$\frac{c_K^n - c_K^{n-1}}{\Delta t} m_K + \sum_{\sigma\in\mathcal{E}_{K,\text{int}}} F_{K\sigma}^n = 0. \tag{1.2.5}$$

On muni cette équation de la condition initiale : $c_K^0 = ⨍_K c^0$, et il suffit de relier $F_{K\sigma}^n$ à $c_K^n, c_{K\sigma}^n, \Phi_K^n, \Phi_{K\sigma}^n$ pour obtenir un schéma numérique. On propose maintenant deux formules pour le "flux numérique" $F_{K\sigma}$. Par souci de simplicité on omettra l'exposant ^{n} indiquant le pas de temps.

Une première idée très naturelle est de poser :

$$F_{K\sigma} = c_\sigma D_{K\sigma}\left(\log(\boldsymbol{c}) + \boldsymbol{\Phi}\right).$$

C'est sur cette formulation que se base le flux centré, avec c_σ la moyenne arithmé-

tique :

$$F_{K\sigma} = \frac{c_K + c_{K\sigma}}{2} D_{K\sigma} \left(\log(\boldsymbol{c}) + \boldsymbol{\Phi}\right). \tag{c}$$

Le second schéma est issu de [116]. L'idée est de poser $F_{K\sigma}$ tel que le problème :

$$c' + c\Phi' = F_{K\sigma}, \qquad c(0) = c_K, c(d_\sigma) = c_L, \Phi' = \frac{\Phi_L - \Phi_K}{d_\sigma}$$

ait une solution. Cela implique :

$$\frac{F_{K\sigma}}{m_\sigma \Phi'} = \frac{c_K e^{\Phi_K} - c_{K\sigma} e^{\Phi_{K\sigma}}}{e^{\Phi_K} - e^{\Phi_{K\sigma}}}$$

En simplifiant :

$$F_{K\sigma} = \tau_\sigma \Big(B\big(\Phi_{K\sigma} - \Phi_K\big) c_K - B\big(\Phi_K - \Phi_{K\sigma}\big) c_{K\sigma} \Big), \tag{sg}$$

où $B = \frac{x}{e^x - 1}$ est la fonction de Bernoulli, prolongée par 1 par continuité en zéro.

Dans les deux cas, nos flux se mettent sous la forme :

$$F_{K\sigma} = \tau_\sigma \mathcal{F}(c_K, c_{K\sigma}, \Phi_K, \Phi_{K\sigma}),$$

où $\mathcal{F}$ a la propriété d'anti-symétrie :

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = -\mathcal{F}(c_L, c_K, \Phi_L, \Phi_K).$$

1.2.3 Pistes d'analyse de ces schémas

On essaiera de garder une certaine similarité entre le schéma de preuve présenté ici et celui des chapitres 2 et 3, quitte à utiliser les résultats qui y seront démontrés.

1.2.3.a Estimations *a priori* ou Propriétés physiques des solutions

Afin d'uniformiser l'étude de nos schémas, on introduit $\mathcal{C}$ la concentration d'interface telle que :

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = \mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) \left(\log(c_K) - \log(c_L) + \Phi_K - \Phi_L\right).$$

Une telle fonction $\mathcal{C}$ est bien définie, continue, et vérifie pour chacun des deux schémas (c) et (sg) :

$$\min(c_K, c_L) \leq \mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) \leq \max(c_K, c_L).$$

La preuve de ce résultat est similaire (quoique plus simple) à celle du Lemme 3.3.1 page 101. On introduit désormais :

$$\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L) = \mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) \left(\log(c_K) - \log(c_L) + \Phi_K - \Phi_L\right)^2 .$$

un équivalent discret de (1.1.3). Une deuxième idée clef des chapitres 2 et 3 est d'utiliser cette fonction de dissipation pour obtenir une borne inférieure à nos solutions discrètes. Pour cela on s'appuie sur la définition de coercivité très fine suivante :

$$\lim_{c \to 0} \inf_{\substack{c_K \in [\delta, +\infty[c_L \in]0, c[, \\ (\Phi_K, \Phi_L) \in [-M, M]^2}} \mathcal{D}(c_K, c_L, \Phi_K, \Phi_L) = +\infty \qquad \forall \delta > 0, M > 0.$$

L'idée pour le schéma centré est de minorer la concentration d'interface $\mathcal{C}$, pour le schéma de Scharfetter-Gummel il faut être plus fin et estimer $\mathcal{F}$. Cette propriété de coercivité permet de propager une borne inférieure sur la concentration sous réserve de :

(H1) disposer sur au moins une cellule de $\epsilon > 0$ tel que $\epsilon \leq c$

(H2) disposer d'une estimation L^∞ sur Φ.

(H3) disposer d'une borne sur $\mathcal{D}(c_K, c_{K\sigma}, \Phi_K, \Phi_{K\sigma})$

Pour exploiter ces points, on suppose disposer de $\boldsymbol{c} \in]0, +\infty[^{\mathcal{T} \times [\![0,n]\!]}$ solution de notre système. Le point (H1) découle de la conservation de la masse. De fait de la propriété de symétrie des flux $F_{K|L} = -F_{L|K}$, on a :

$$\sum_{K \in \mathcal{T}} c_K^n m_K = \int_\Omega c^0, \qquad \forall n \in [\![0, N_T]\!].$$

Du fait de notre hypothèse de positivité, on dispose donc d'au moins une cellule K_{init} telle que :

$$c_{K_{\text{init}}}^n \geq \fint_\Omega c^0$$

Pour le point (H2), on aimerait utiliser la proposition 2.A.1 page 77. Pour cela il nous faut un borne supérieure sur nos concentrations. Celle-ci dérive elle aussi de la conservation de la masse :

$$m_K c_K^n \leq \int_\Omega c^0, \quad \forall K \in \mathcal{T}.$$

Contrairement aux Chapitres 2 et 3, cette estimation sur $\boldsymbol{\Phi}$ dépend du maillage.

Le point (H3) vient lui d'une propriété intéressante en elle-même de nos schémas : la décroissance de l'énergie libre 1.1.2. Pour ce faire on introduit son équi-

valent discret :

$$E_{\mathcal{T}}(\boldsymbol{c}^n, \boldsymbol{\Phi}^n) = \sum_{K\in\mathcal{T}} m_K c_K \log c_K + \frac{1}{2}\sum_{\sigma\in\mathcal{E}} \tau_\sigma \left(D_\sigma \boldsymbol{\Phi}^n\right)^2 - \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D\cap\mathcal{E}_K} \tau_\sigma \Phi_\sigma^D D_{K\sigma}\boldsymbol{\Phi}^n.$$

Ce troisième terme est nécessaire pour tenir compte des apports d'énergie externes liés à des condition de Dirichlet inhomogènes. Par souci de simplicité, on choisi dès lors de prendre $\Phi^D = 0$. Ce choix limite certes l'intérêt physique du modèle étudié -et n'est donc pas fait dans le reste de la thèse- mais le traitement des conditions de bord apporte peu à la démarche mathématique présentée dans cette section. On a ainsi :

$$\frac{E_{\mathcal{T}}(\boldsymbol{c}^t, \boldsymbol{\Phi}^t) - E_{\mathcal{T}}(\boldsymbol{c}^{t-1}, \boldsymbol{\Phi}^{t-1})}{\Delta t} \leq -\sum_{\sigma=K|L\in\mathcal{E}_{\text{int}}} \tau_\sigma \mathcal{D}(c_K^t, c_L^t, \Phi_K^t, \Phi_L^t) \leq 0, \quad \forall 1 \leq t \leq n. \tag{1.2.6}$$

L'inégalité difficile à démontrer est la première. D'abord la convexité de $E_{\mathcal{T}}$ donne une majoration du terme de droite :

$$E_{\mathcal{T}}(\boldsymbol{c}^t, \boldsymbol{\Phi}^t) - E_{\mathcal{T}}(\boldsymbol{c}^{t-1}, \boldsymbol{\Phi}^{t-1}) \leq \sum_{K\in\mathcal{T}} m_K (c_K^t - c_K^{t-1}) \log(c_K^t) + \sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\boldsymbol{\Phi}^t D_{K\sigma}(\boldsymbol{\Phi}^t - \boldsymbol{\Phi}^{t-1}),$$

où le terme constant dans la dérivée de $c\log c$ est éliminé grâce à la conservation de la masse. L'étape de calcul suivante est appelée intégration par partie discrète. Le principe repose sur l'idée suivante :

$$\sum_{\sigma\in\mathcal{E}_{\text{int}}} \tau_\sigma D_{K\sigma}\boldsymbol{u} D_{K\sigma}\boldsymbol{v} = \sum_{K\in\mathcal{T}} u_K \sum_{\sigma\in\mathcal{E}_{K,\text{int}}} D_{K\sigma}\boldsymbol{v}$$

Pour une arrête interne $\sigma = K|L$, $D_{K\sigma}\boldsymbol{v}$ et $D_{L\sigma}\boldsymbol{v}$ sont de signes opposés, donc dans la somme la différence $v_K - v_L$ apparaîtra une fois avec le facteur u_K et une fois avec le facteur $-u_L$. Dans le cas présent, on a :

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\boldsymbol{\Phi}^n D_{K\sigma}(\boldsymbol{\Phi}^t - \boldsymbol{\Phi}^{t-1}) = \sum_{K\in\mathcal{T}} \Phi_K \sum_{\sigma\in\mathcal{E}_K} D_{K\sigma}(\boldsymbol{\Phi}^t - \boldsymbol{\Phi}^{t-1}).$$

En appliquant l'équation (1.2.4), on a :

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\boldsymbol{\Phi}^t D_{K\sigma}(\boldsymbol{\Phi}^t - \boldsymbol{\Phi}^{t-1}) = \sum_{K\in\mathcal{T}} \Phi_K \left(c_K^t - c_K^{t-1}\right) m_K.$$

Somme toute :

$$E_{\mathcal{T}}(\boldsymbol{c}^t, \boldsymbol{\Phi}^t) - E_{\mathcal{T}}(\boldsymbol{c}^{t-1}, \boldsymbol{\Phi}^{t-1}) \leq \sum_{K\in\mathcal{T}} m_K(c_K^t - c_K^{t-1}) \left(\log(c_K^t) + \Phi_K^t\right).$$

On remarque que le terme de droite correspond -à un facteur Δt près- à l'approximation de la dérivée en temps multipliée par $\log(c_K^t) + \Phi_K^t$ et sommée sur le domaine d'où

$$\sum_{K\in\mathcal{T}} m_K(c_K^t - c_K^{t-1}) \left(\log(c_K^t) + \Phi_K^t\right) = \Delta t \sum_{K\in\mathcal{T}} \left(\log(c_K^t) + \Phi_K^t\right) \sum_{\sigma\in\mathcal{E}_{K,\text{int}}} F_{K\sigma}^t.$$

Grâce à notre propriété de symétrie des flux, on peut à nouveau intégrer par partie :

$$\sum_{K\in\mathcal{T}} m_K(c_K^t - c_K^{t-1}) \left(\log(c_K^t) + \Phi_K^t\right) = \Delta t \sum_{\sigma\in\mathcal{E}_{\text{int}}} F_{K\sigma}^t D_{K\sigma} \left(\log(\boldsymbol{c}^t) + \boldsymbol{\Phi}^t\right).$$

En récapitulant et en utilisant la fonction $\mathcal{D}$, on obtiens bien l'équation (1.2.6). Cette inégalité sur la décroissance de l'énergie libre ne nous donne pas directement (H3). On a certes par positivité de $\mathcal{D}$:

$$\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L) \leq \frac{E_{\mathcal{T}}(c^0, \Phi^0) - E_{\mathcal{T}}(c^n, \Phi^n)}{\Delta t \min \tau_\sigma}, \qquad \forall K|L \in \mathcal{E}_{\text{int}},$$

mais rien ne garantis *a priori* que le terme de droite soit borné. En remarquant que les fonctions $c \mapsto c\log c$ et $x \mapsto x^2$ sont minorées, on obtiens une borne inférieure sur $E_{\mathcal{T}}(c^n, \Phi^n)$, donc une borne supérieure sur $\mathcal{D}$.

Nos trois hypothèses étant vérifiées, on dispose de $\epsilon > 0$ tel que :

$$\epsilon < c_K^t, \forall K \in \mathcal{T}, t \in [\![1, N]\!]$$

Quitte à prendre une borne inférieure plus petite, on peut cacher notre borne supérieure sur c dans cette inégalité :

$$\epsilon < c_K^t < \frac{1}{\epsilon}, \forall K \in \mathcal{T}, t \in [\![1, N]\!]. \tag{1.2.7}$$

1.2.3.b Existence de solution

Dans la section précédente, nous étions partis de l'hypothèse qu'on disposait de solutions positives à nos schémas numériques pour en donner des propriétés qualitatives. Cependant nos schémas forment des systèmes non linéaires et il n'est *a priori* pas toujours possible d'en trouver une solution. On s'attache donc désormais à démontrer que nous ne rencontrons pas ce problème en utilisant la méthode du

degrès topologique [98, 56]. En clair, on déformera notre systeme non linéaire continuement jusqu'à le transformer en un systeme linéaire dont on connaît des solutions, ici l'équation de la chaleur et l'équation de Poisson. Ce processus se déroule en deux étape, la première se fait à potentiel nul et déforme le schéma habituel pour l'équation de la chaleur en un schéma non linéaire.

Plus concrètement, pour $\alpha \in [0,1]$, on pose pour tout $K \in \mathcal{T}$:

$$\frac{c_K^\star - c_K^{n-1}}{\Delta t} m_K + \sum_{\sigma \in \mathcal{E}_{K,\text{int}}} \tau_\sigma \left((1-\alpha) D_{K\sigma} \boldsymbol{c}^\star + \alpha F_{K\sigma} \right) = 0 \tag{1.2.8}$$

$$F_{K\sigma} = \tau_\sigma \mathcal{F}(c_K^\star, c_{K\sigma}^\star, 0, 0) \tag{1.2.9}$$

$$\sum_{\sigma \in \mathcal{E}_K} D_{K\sigma} \boldsymbol{\Phi}^\star = 0 \tag{1.2.10}$$

En remarquant que cette transformation ne modifie pas les solutions du schéma de Scharfetter-Gummel, on garde la borne de l'équation (1.2.7). Ensuite, on active le potentiel, en posant pour α entre 1 et 2 :

$$\frac{c_K^\star - c_K^{n-1}}{\Delta t} m_k + \sum_{\sigma \in \mathcal{E}_{K,\text{int}}} \tau_\sigma \left(F_{K\sigma} \right) = 0 \tag{1.2.11}$$

$$F_{K\sigma} = \tau_\sigma \mathcal{F}(c_K^\star, c_{K\sigma}^\star, \alpha \Phi_K^\star, \alpha \Phi_{K\sigma}^\star) \tag{1.2.12}$$

$$\sum_{\sigma \in \mathcal{E}_K} D_{K\sigma} \boldsymbol{\Phi}^\star = \alpha c_K^\star. \tag{1.2.13}$$

Il est nécessaire de procéder en deux étapes pour préserver l'inégalité d'entropie dissipation (1.2.6). De même, dans la deuxième étape, pour préserver cette inégalité, il est nécessaire d'introduire le potentiel électrique à la même vitesse dans les deux équations, α faisant office de charge. Pour les conditions de bord inhomogènes sur le potentiel, elles peuvent être introduites au cours de n'importe laquelle des deux étapes. Ayant pu conserver cette inégalité d'entropie-dissipation, nous conservons a nouveau l'équation (1.2.7). On a ainsi créé une fonction continue $\mathcal{H} : [0,2] \times \mathbb{R}^\mathcal{T} \times \mathbb{R}^\mathcal{T} \to \mathbb{R}^\mathcal{T} \times \mathbb{R}^\mathcal{T}$ telle que $\mathcal{H}(0, \star, \star) = 0$ aie un degré topologique égal à un et que $\mathcal{H}(2, \star, \star) = 0$ corresponde à notre schéma. On a en outre montré qu'au moins une trajectoire de solution à $\mathcal{H}(\alpha, \star, \star) = 0$ restaient dans un compact, on dispose donc d'une solution à $\mathcal{H}(2, \star, \star) = 0$ dans ce compact, c'est à dire d'une solution à notre schéma. □

Les chapitres 2, 3 et 5 comportent également une partie sur la convergence des schémas et leur qualités numériques. Il s'agit dans les deux cas d'une preuve en trois temps temps. D'abord on introduit une reconstruction de nos solutions discrètes sur Ω en suivant la méthodologie de [64]. On utilise ensuite un résultat de

compacité qui donne l'existence d'une limite à ces reconstructions lorsque le $h_{\mathcal{T}}, \Delta t$ tendent vers zéros avec $\zeta_{\mathcal{T}} > \zeta^\star$. Pour cette deuxième étape, on utilise des résultats d'inégalités fonctionelles discrètes tels que [16, Theorem 5], [86, Theorem 2.1] puis de compacité [9, Theorem 3.9],[78, Lemma 3.4],[31, Lemma 9]. Dans le chapitre 3 la propriété de compacité est particulièrement difficile a obtenir. Enfin, on identifie la limite obtenue à une solution faible de notre problème de départ. Dans cet objectif, on utilise dans chacun de ces chapitres une deuxième reconstruction des gradients [63] (dans le cadre de [52]) et le Lemme 3.D.2 page 140.

1.3 Organisation et résultats principaux

1.3.1 Idées clefs et résultats principaux du chapitre 2

Dans le chapitre 2, on s'intéresse au modèle construit en section 1.1.3. Ce chapitre est issu d'un article publié dans IMAJNA (IMA Journal of Numerical Analysis) [29] et résumé dans une contribution à la conférence FVCA IX (Finite volumes for complex applications) [30].

Comme en section 1.2, on cherche à discrétiser le flux de dérive-diffusion qui s'écrit :

$$F = c\nabla\left(h(c) + \Phi\right) \qquad h(c) = \log\frac{c}{1-c}. \tag{F-C}$$

Si il est possible de discrétiser ce flux directement, nous considérons trois manière différentes de considérer la non linéarité $-\log 1 - c$. La première manière est de considérer le potentiel chimique excédentaire [95] : $\nu(c) = h(c) - \log(c)$, sous cette forme, on a :

$$F = \nabla c + c\nabla\left(\Phi + \nu(c)\right). \tag{F-S}$$

Une deuxième option est de considérer une augmentation de la diffusion [125] : r tel que $r'(c) = ch'(c)$, on a alors :

$$F = r'(c)\nabla c + c\nabla\Phi \tag{F-BCH}$$

Enfin, on peut considérer les coefficients d'activité et coefficients inverse d'activité [103] $a(c) = \exp h(c), \beta(c) = \frac{c}{a(c)}$ ce qui donne :

$$F = \beta(c)\left(\nabla a(c) + a(c)\nabla\Phi\right). \tag{F-AB}$$

Pour chacune de ces formulations, on propose un flux numérique. La formulation initiale (F-C) donne :

$$F_{K\sigma} = \tau_\sigma \frac{c_K + c_{K\sigma}}{2} D_{K\sigma} h(\boldsymbol{c}) + \boldsymbol{\Phi} \tag{C}$$

C'est la traduction directe du schéma centré de la section précédente, les flux numériques associés aux trois autres expressions s'appuient sur le schéma de Scharfetter-Gummel vu dans la même section. La formulation en potentiel chimique excédentaire (F-S) donne

$$F_{K\sigma} = \tau_\sigma \Big(B\big(-D_{K\sigma}\mathbf{\Phi} + \nu(\boldsymbol{c})\big)c_K - B\big(D_{K\sigma}\mathbf{\Phi} + \nu(\boldsymbol{c})\big)c_{K\sigma} \Big). \tag{S}$$

Ce schéma est appelé schéma de Sedan car le code de simulation SEDAN III [126] est la trace la plus ancienne que nous ayons trouvée de cette idée. La formulation en diffusion augmentée (F-BCH) donne le schéma de Bessemoulin Chatard :

$$F_{K\sigma} = \tau_\sigma \mathfrak{d}r(c_K, c_{K\sigma}) \Big(B\big(\frac{\Phi_{K\sigma} - \Phi_K}{\mathfrak{d}r(c_K, c_{K\sigma})}\big)c_K - B\big(\frac{\Phi_K - \Phi_{K\sigma}}{\mathfrak{d}r(c_K, c_{K\sigma})}\big)c_{K\sigma} \Big) \tag{BCH}$$

$$\mathfrak{d}r(c_K, c_L) = \begin{cases} \frac{h(c_K) - h(c_L)}{\log(c_K) - \log(c_L)} & \text{si } c_K \neq c_L, \\ r'(c_K) & \text{si } c_K = c_L. \end{cases}$$

Il vient d'une idée de Marianne Bessemoulin-Chatard [15] et pose une approximation logarithmique de la dérivée de r pour préserver l'équilibre thermique. Enfin, l'expression en activités (F-AB) donne :

$$F_{K\sigma} = \tau_\sigma \frac{\beta(c_K) + \beta(c_{K\sigma})}{2} \Big(B\big(\Phi_{K\sigma} - \Phi_K\big)a(c_K) - B\big(\Phi_K - \Phi_{K\sigma}\big)a(c_{K\sigma}) \Big) \tag{AB}$$

Bien que le dernier ait été proposé par Jürgen Fuhrmann en [70, 74], il faisait partie des auteurs des articles utilisant ce schéma et nous sommes donc restés factuels en l'appelant "activity based" ou schéma en activités. Le choix de la moyenne arithmétique pour β quoique cohérent avec la littérature existante n'est probablement pas optimal. En annexe de cette thèse, on propose un autre choix de moyenne au travers de la preuve d'une conjecture de J. Fuhrmann.

Après une analyse numérique très similaire à celle de la section précédente, nous sommes arrivés à deux résultats assez généraux. Le théorème 2.2.1 page 45 concerne l'existence de solution à nos schémas numériques :

Theorem 1.3.1. *Soit $(\mathcal{T}, \mathcal{E}, (\boldsymbol{x}_K)_{K\in\mathcal{T}})$ un maillage admissilbe et $\boldsymbol{c}^0 \in \mathbb{R}^{\mathcal{T}}$ une donnée initiale positive non uniformément nulle. Pour tout $n \in [\![1, N_T]\!]$, et quelque soit le flux numérique choisi, notre schéma a une solution $(\boldsymbol{c}^n, \mathbf{\Phi}^n) \in [0,1]^{\mathcal{T}} \times \mathbb{R}^{\mathcal{T}}$. De plus cette solution satisfait pour tout $1 \leq n \leq N_T$,*

$$E_{\mathcal{T}}(\boldsymbol{c}^n, \mathbf{\Phi}^n) \leq E_{\mathcal{T}}(\boldsymbol{c}^{n-1}, \mathbf{\Phi}^{n-1}) \text{ et } 0 < c_K^n < 1, \quad \forall K \in \mathcal{T}.$$

où

$$E_{\mathcal{T}}(\boldsymbol{c}^n, \boldsymbol{\Phi}^n) = \sum_{K\in\mathcal{T}} m_K H(c_K^n) + \frac{1}{2}\sum_{\sigma\in\mathcal{E}} \tau_\sigma \left(D_\sigma \boldsymbol{\Phi}^n\right)^2 - \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D\cap\mathcal{E}_K} \tau_\sigma \Phi_\sigma^D D_{K\sigma}\boldsymbol{\Phi}^n.$$

Notre second résultat porte sur la convergence des solutions approchées vers des solutions faibles. La définition 3 page 39 indique ainsi :

Definition 2. *Un couple (c, Φ) est une solution faible de notre système d'EDP si :*

- $c \in L^\infty(Q_T; [0,1])$ *avec* $r(c) \in L^2((0,T); H^1(\Omega))$ *;*
- $\Phi - \Phi^D \in L^\infty((0,T), \mathcal{H}_{\Gamma^D})$ *;*
- *pour tout* $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$,

$$\iint_{Q_T} c\partial_t\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega c^0\varphi(0,\cdot)\mathrm{d}\boldsymbol{x} - \iint_{Q_T} (\nabla r(c) + c\nabla\Phi)\cdot\nabla\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0;$$

- *pour tout* $\psi \in \mathcal{H}_{\Gamma^D}$ *et presque tout* $t \in (0,T)$,

$$\int_\Omega \nabla\Phi(t,\boldsymbol{x})\cdot\nabla\psi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega (c(t,\boldsymbol{x}) + c^{\mathrm{dop}}(\boldsymbol{x}))\psi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

où $\mathcal{H}_{\Gamma^D}$ est l'espace des fonctions H^1 nulles sur Γ_D. Pour la reconstruction, on pose simplement $\pi_{\mathcal{T},\boldsymbol{\Delta t}}$ l'opérateur qui à $\boldsymbol{u} \in \mathbb{R}^{(\mathcal{T}\times[\![1,N_T]\!])}$ associe la fonction $L^1([0,T]\times\Omega)$ définie presque partout par :

$$\pi_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}(t,\boldsymbol{x}) = u_K^n \quad \text{si } (t,\boldsymbol{x}) \in](n-1)\Delta t, n\Delta t]\times K.$$

On peut désormais donner le théorème 2.2.2 page 46 correspondant :

Theorem 1.3.2. *Pour le schéma centré et le schéma de Sedan, une suite de solutions approchées $(\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{c}_m, \pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\Phi}_m)_{m\geq 1}$ telle que $h_{\mathcal{T}_m}, \Delta t_m$ tendent vers zéros alors que la régularité du maillage ne dégénère pas vérifient, à une sous-suite près :*

$$c_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} c \quad p.p.\ dans\ Q_T, \qquad \Phi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \Phi \quad dans\ L^2(Q_T),$$

où (c,Φ) est une solution faible de notre système d'EDP.

Pour le schéma en activités, le problème vient du fait que la concentration d'interface est le produit d'une moyenne des coefficients d'activités et d'une moyenne des coefficients inverses d'activité qui ne sont pas compatibles. En effet de façon générale, le produit de moyennes n'est pas une moyenne du produit. Pour le schéma de Bessemoulin-Chatard, il y a un très léger défaut de compacité. Dans les deux cas cela ne signifie pas que ces schémas ne convergent pas, juste que notre analyse ne les couvre pas.

1.3.2 Idées clefs et résultats principaux du chapitre 3

Dans le chapitre 3, on s'intéresse au modèle construit en section 1.1.2 et sous la forme de (1.1.10) page 7. Ce chapitre est issu d'un article soumis pour publication [81].

Le cas du chapitre précédent étant un cas très particulier de ce cadre, on sait d'expérience que les schéma centrés et de Sedan -si on peut les étendre- auront une analyse numérique plus simple. Il est également possible d'étendre le schéma en activités. Pour le schéma Bessemoulin-Chatard, son extension n'est pas naturelle. En effet, même si on peut toujours appliquer la formule du flux correspondant :

$$F_{K\sigma,i} = \tau_\sigma \mathfrak{d}r_i(c_K, c_{K\sigma})\Big(B\big(\frac{\Phi_{K\sigma}-\Phi_K}{\mathfrak{d}r_i(c_K,c_{K\sigma})}\big)c_K - B\big(\frac{\Phi_K-\Phi_{K\sigma}}{\mathfrak{d}r_i(c_K,c_{K\sigma})}\big)c_{K\sigma}\Big)$$
$$\mathfrak{d}r_i(cK, iK, c_{L,i}) = \frac{h_i(c_{K,i}) - h_i(c_{L,i})}{\log(c_{K,i}) - \log(c_{L,i})}$$

cela ne correspond pas à la jacobienne de $h = (h_i)_{i\in[\![1,N]\!]}$, et n'est donc pas consistant avec le flux continu.

On utilise ainsi les formulations suivantes des flux continus :

$$\begin{aligned} F_i &= -D_i c_i \nabla\left(h_i(C) + z_i\Phi\right) \quad \forall i \in [\![1,N]\!] \\ &= -D_i\left(\nabla c_i + c_i\nabla\left(\nu_i(C) + z_i\Phi\right)\right) \quad \forall i \in [\![1,N]\!] \end{aligned}$$

où $h_i(C) = \log\frac{c_i}{\bar{c}} - k_i \log\frac{c_0}{\bar{c}}$ est le potentiel électrochimique effectif, $C = (c_0, ...c_N)$ est le vecteur des concentrations, et $\nu_i(C) = h_i(C) - \log c_i$ le potentiel chimique excédentaire.

A ces flux nous associons les flux numériques :

$$F_{K\sigma,i} = \tau_\sigma \frac{c_{K,i} + c_{K\sigma,i}}{2} D_{K\sigma} h(\boldsymbol{C}) + \boldsymbol{\Phi} \tag{C'}$$

et

$$F_{K\sigma,i} = \tau_\sigma\Big(B\big(-D_{K\sigma}\boldsymbol{\Phi} + \nu(\boldsymbol{c})\big)c_K - B\big(D_{K\sigma}\boldsymbol{\Phi} + \nu(\boldsymbol{c})\big)c_{K\sigma}\Big). \tag{S'}$$

Pour le solvant c_0, on utilise l'incompressibilité $\sum_{i=0}^N c_i v_i = 0$ pour calculer sa concentration.

Le croisement des termes dans le potentiel complique fortement l'analyse numérique de ce schéma. Ainsi, nous devons d'abord minorer la concentration de solvant avant de minorer celle des autres espèces, mais aussi prendre en compte de nombreuses sortes de dégénérescences possibles dans l'estimation de compacité. Malgré ces difficultés techniques, nous avons pu démontrer l'existence de solutions à notre schéma dans le théoreme 3.2.1 page 99. Pour la convergence des solutions

approchées vers une solution faible, nous n'arrivons à identifier la limite comme solution faible que dans l'hypothèse $\inf_{\Omega\times[0,T]} c_0 > 0$. Cette hypothèse permet d'obtenir une borne sur le gradient de $\log(c_0)$.

1.3.3 Idées clefs, résultats principaux et perspectives du chapitre 4

Dans ce chapitre, on ne cherche pas à analyser des schémas numériques mais à proposer des modèles justifiant le choix de matrice d'Onsager par des considérations sur la physique sous-jacente. La démarche, basée sur la modélisation variationnelle est présentée à la section 1.1.4. On y propose deux manières de prendre en compte l'incompressibilité :

— en ajoutant le terme $\chi(\sum_{i=1}^N c_i v_i = 1)$ -ou χ vaut 0 en 0 et $+\infty$ ailleurs- à l'énergie. C'est équivalent au fait d'imposer dans la dissipation une somme de flux volumiques à divergence nulle. On parlera alors de contrainte faible.

— en forçant dans la dissipation les flux volumiques à être de somme nulle : $\sum_{i=1}^N v_i J_i$. On nommera cette contrainte d'incompressibilité forte

A ces deux manières de prendre en compte l'incompressibilité, on ajoute différentes dissipations "de base". Une correspond à la diffusion Fickienne (voir section 4.3.1), une donne l'équation étudiée au chapitre 3 (cf section 4.3.2), une autre corrige cette dissipation pour obtenir un résultat symétrique dans le modèle réduit étudié au chapitre 2 (en section 4.3.3) cela correspond aussi à [26]. En Section 4.3.4, on considère une dissipation inspirée de Maxwell-Stefan où la dissipation se produit par la différence de vitesse entre le fluide et celle de chaque espèce (avec une pénalisation de la vitesse globale pour la corecivité). Enfin en section 4.3.5, on s'intéresse au système de Maxwell-Stefan. Malheureusement, ce système manque de coercivité et ne se prête pas -en général- à une modélisation variationnelle, néanmoins une contrainte d'incompressibilité forte suffit à rendre le système coercif, et donne ainsi un sens "physique" au procédé utilisé dans [40].

Des résultats numériques illustrent ces différents modèles. Ce chapitre n'est pas encore mûr pour publication et si il donne lieu à un article, ce dernier inclura idéalement une comparaison à des données expérimentales pour constater d'une différence d'adéquation des modèles au réel et un couplage avec la solvatation pour prendre en compte le fait que les ions sont très solvatés dans le mélange et moins à proximité des électrodes.

1.3.4 Idées clefs et résultats principaux du chapitre 5

Ce chapitre est issu d'un article publié dans SINUM (SIAM Journal on Numerical Analysis) [34] et résumé dans une contribution à la conférence FVCA IX

(Finite volumes for complex applications) [42]. On y étudie le modèle présenté en section 1.1.5. Cette fois ci, l'analyse est différente, on remarque d'abord que pour $a_{i,j} = a^\star$, on a :

$$\sum_{j\neq i} a_{i,j}\big(u_i\nabla u_j - u_j\nabla u_i\big) = a^\star u_i \nabla \sum_{j=1}^N u_j - a^\star \sum_{j=1}^N u_j \nabla u_i.$$

En supposant que le nombre d'atomes par maille cristalline est constant, $\nabla \sum_{i=1}^N u_j$ est nul. Quitte à changer d'unités, on peut supposer cette somme égale à un. On remarque que par construction du modèle, $\sum_{i=1}^N u_j$ est constant en temps, il suffit donc qu'il soit constant en espace initialement. Sous cette hypothèse que l'on fait désormais, on a :

$$\sum_{j\neq i} a_{i,j}\big(u_i\nabla u_j - u_j\nabla u_i\big) = -a^\star \nabla u_i.$$

En d'autres termes, on s'intéresse à une perturbation de l'équation de la chaleur :

$$\partial_t u_i - a^\star \,\Delta\, u_i + \operatorname{div}\Big((a_{i,j} - a^\star)\sum_{i\neq j} u_i u_j \nabla \log\frac{u_j}{u_i}\Big) = 0. \tag{1.3.1}$$

Cette structure a été utilisée dans [14] avec une hypothèse de proximité entre $a_{i,j}$ et $a^\star$ pour montrer l'existence et l'unicité de solutions fortes. Ce modèle dispose d'une fonctionnelle de Lyapunov : $\int_\Omega \sum_{i=1}^N u_i \log u_i$ qui correspond à l'énergie libre de mélange. Afin de préserver au niveau discret cette propriété, on cherche à préserver au niveau discret l'égalité $\nabla u_j = u_j \nabla \log u_j$. Pour ce fait, on introduit à nouveau une valeur d'interface $u_{K|L,i}$ comme étant la moyenne logarithmique de $u_{K,i}$ et $u_{L,i}$:

$$u_{K|L,i} = \begin{cases} 0 & \text{si } \min(u_{K,i}, u_{L,i}) \leq 0, \\ u_{K,i} & \text{sinon, si } u_{K,i} = u_{L,i}, \\ \frac{u_{K,i} - u_{L,i}}{\log u_{K,i} - \log u_{L,i}} & \text{sinon.} \end{cases}$$

Avec cet outil, on définit les flux :

$$F_{K\sigma,i} = a^\star D_{K\sigma}\boldsymbol{u}_i + \sum_{j\neq i}(a_{i,j} - a^\star)\Big(u_{\sigma,i} D_{K\sigma}\boldsymbol{u}_j - u_{\sigma,j} D_{K\sigma}\boldsymbol{u}_i\Big)$$

Il reste à définir $a^\star$. Le choix optimal de ce paramètre reste une question ouverte, néanmoins, avec $a^\star > 0$ et des conditions très peu restrictives sur $a_{i,j}$ nous avons pu démontrer l'existence de solution positives à notre schéma. De plus elles ont un

équivalent discret de notre fonctionnelle de Lyapunov :

$$E_{\mathcal{T}} = \sum_{K\in\mathcal{T}} m_k \sum_{i=1}^{N} u_{K,i} \log u_{K,i}.$$

On donne aussi un résultat de convergence sous l'hypothèse plus restrictive $a_{i,j} > 0$. En effet dans ce cas, si $a^\star > \min a_{i,j} > 0$, on a :

$$E_{\mathcal{T}}(\boldsymbol{U}^n) - E_{\mathcal{T}}(\boldsymbol{U}^{n-1}) + \Delta t \min_{1\leq i,j\leq N} a_{i,j} \sum_{\sigma\in\mathcal{E}} \sum_{i=1}^{N} \tau_\sigma u_{i,\sigma}^n (D_{K\sigma} \log(\boldsymbol{u}_i^n))^2 \leq 0,$$

ce qui nous permet d'obtenir sans trop de soucis une estimation sur la norme H^1 discrète de $\boldsymbol{u}_i$. Dans le cas où le graphe des zéros de $a_{i,j}$ est un "cluster graph" c'est à dire si "$a_{i,j} = 0$ et $a_{j,k} = 0$ implique $a_{i,k} = 0$", on dispose d'un résultat plus fin. En notant $S_1, ..., S_n$ l'ensemble des espèces de chacun des n clusters, et S_0 l'ensemble des espèces qui ont un coefficient a strictement positif, on a :

$$E_{\mathcal{T}}(\boldsymbol{U}^n) - E_{\mathcal{T}}(\boldsymbol{U}^{n-1}) + \Delta t \sum_{\sigma\in\mathcal{E}} \tau_\sigma D_\sigma \leq 0$$

$$D_\sigma = a^\star \sum_{i\in S_0}^{N} u_{i,\sigma}^n (D_{K\sigma} \log(\boldsymbol{u}_i^n))^2 + a^\star \sum_{k=1}^{n} D_\sigma^k$$

$$D_\sigma^k = \sum_{i\in S_k} u_{\sigma,i} \left(1 - \sum_{j\in S_k} u_{\sigma,j}\right) \left(D_{K\sigma} \log \boldsymbol{u}_i\right)^2 + \left(\sum_{i\in S_k} D_{K\sigma} \boldsymbol{u}_i\right)^2.$$

Ce résultat ne permet pas de traiter le cadre général mais est un outil très pratique pour nombre de cas particuliers de "cluster graph" tel que les graphes étoilés ($n = 1$) comme c'est le cas dans [83]. Ces cas ne sont pas traités dans le chapitre 5 du fait de leur nombre et par souci de simplicité. En revanche on y propose un mécanisme supplémentaire de réactions chimiques.

Chapter 2

A numerical analysis focused comparison of several finite volume schemes for a unipolar degenerate drift-diffusion model

Ce chapitre est un travail en collaboration avec Clément Cances, Claire Chainais, et Jürgen Führmann. Il a été publié dans IMAJNA [29].

In this paper, we consider a unipolar degenerate drift-diffusion system where the relation between the concentration of the charged species c and the chemical potential h is $h(c) = \log \frac{c}{1-c}$. We design four different finite volume schemes based on four different formulations of the fluxes. We provide a stability analysis and existence results for the four schemes. The convergence proof with respect to the discretization parameters is established for two of them. Numerical experiments illustrate the behaviour of the different schemes.

Outline of the current chapter

2.1 Introduction **34**
2.1.1 Motivation 34
2.1.2 A simplified unipolar degenerate drift-diffusion model 36
2.1.3 Entropy structure and weak solutions 37
2.2 Finite Volume approximations **40**
2.2.1 Discretization of $(0,T)\times\Omega$ 40
2.2.2 A common basis for the Finite Volume schemes . . . 42
2.2.3 Numerical fluxes for the conservation of the chemical species 43
2.2.4 Main results and organisation of the paper 45
2.3 Numerical analysis for fixed meshes **47**
2.3.1 Face concentration and face dissipation 47
2.3.2 Uniform a priori estimates 50
2.3.3 Existence of a solution to the schemes 54
2.4 About the convergence towards a weak solution **56**
2.4.1 Reconstruction operators 57
2.4.2 Compactness properties for the approximate concentration 58
2.4.3 Convergence towards a weak solution 63
2.5 Numerical comparison of the schemes **68**
2.5.1 1D time evolution and convergence test 68
2.5.2 1D stationary convergence test 70
2.5.3 2D Unipolar Field Effect Transistor 71
2.6 Conclusion **73**
2.A L^∞ bound on the TPFA FV approximate Poisson equation **76**
2.B Proof of Lemma 2.3.2 **78**
2.C Comparison of face concentration functionals **81**
2.D Some notations **84**

2.1 Introduction

2.1.1 Motivation

Unipolar drift-diffusion models describe the transport of a charged species in the presence of a fixed or moving countercharge. They consist of the coupling of a drift-diffusion equation on the density of the charged species c with a Poisson

equation on the electric potential Φ. They can be written under a general form as

$$\begin{cases} \partial_t c + \operatorname{div}(\mathbf{J}) = 0, \quad \mathbf{J} = -\eta(c)\nabla(h(c) + \Phi), \\ -\lambda^2 \Delta \Phi = c + c^{\mathrm{dop}}, \end{cases}$$

where h is the chemical potential, η the mobility coefficient, λ the scaled Debye length coming from the nondimensionalisation of the physical model and c^{dop} describes the doping profile of the media.

Such models occur in many interesting application cases. Charge carriers in most classical semiconductors exhibit a relationship $c = \mathcal{F}(h)$, where $\mathcal{F}$ is the Fermi integral of index $\frac{1}{2}$ which can be approximated in the range $-\infty < h \lessapprox 1.3$ by the function $\mathcal{F}(h) = \frac{1}{\gamma + \exp(-h)}$ with $\gamma = 0.27$ [18]. For $\gamma = 1$, this relationship corresponds to the Fermi integral of index -1 and implies $h = \log \frac{c}{1-c}$. It is the limit for vanishing disorder of the Gauss-Fermi integral [96, 112] which is used to describe organic semiconductors [51]. A similar relationship is valid for the oxygen ion concentration in a solid oxide electrolyte [124] and a simple model of an ionic liquid [71].

While the relationship between chemical potential and concentration is sufficient to describe the thermodynamic equilibrium, the description of charge transport driven by the sum of the gradients of the chemical potential and the electrostatic potential Φ needs an additional specification of the mobility coefficient η. Setting this coefficient proportional to the concentration c is common in the case of semiconductors [117]. A similar ansatz describes the limit of large lattice mass density in solid oxide electrolytes. It also follows from a formal reduction of a generalized Nernst-Planck model [60, 59] to the case of a mixture of two charged species including an infinitely mobile and charged solvent – ionic liquids – as performed in [71]. We hint that more general and fully consistent models for both solid oxide electrolytes and ionic liquids consider mobility coefficients of the type $c(1-c)$ [124, 25, 82].

In this paper, we consider that the mobility coefficient is $\eta(c) = c$ and the chemical potential $h(c) = \log \frac{c}{1-c}$ (corresponding to $\mathcal{F}(h) = \frac{1}{1+\exp(-h)}$). Strong degeneration described by a bounded dependency of the concentration c on the chemical potential h leads to some structural mathematical challenges in the corresponding drift-diffusion models. These need to be addressed properly in numerical schemes. The consideration of this simplified model is a starting point for the study of generalized Nernst-Planck models for multiple ionic species in electroneutral solvents [60, 59, 70, 71]. Moreover, the design of discretization methods for the case where $\eta(c) = c(1-c)$ is also a possible topic of further investigation following the present paper.

2.1.2 A simplified unipolar degenerate drift-diffusion model

Let us now define the framework of the study. We consider the evolution of the concentration c of a charged species in a connected bounded open domain Ω of $\mathbb{R}^d$ ($d \leq 3$) with polyhedral and Lipschitz continuous boundary $\partial\Omega$ during a finite but arbitrary time $T > 0$. After nondimensionalisation with appropriate scaling, we regard the following system of partial differential equations (PDEs). The concentration c satisfies the conservation law

$$\partial_t c + \operatorname{div}(\mathbf{J}) = 0 \quad \text{in } (0,T)\times\Omega. \tag{2.1.1}$$

The flux $\mathbf{J}$ is negatively proportional to the gradient of the electrochemical potential as expressed by the expression

$$\mathbf{J} = -c\nabla\left(h(c) + \Phi\right) \quad \text{in } (0,T)\times\Omega, \tag{2.1.2}$$

where $h(c) = \log\left(\frac{c}{1-c}\right)$ is the chemical potential. In what follows, we consider that the electrostatic potential Φ is related to space charge density thanks to the Poisson equation

$$-\Delta\Phi = c + c^{\mathrm{dop}} \quad \text{in } (0,T)\times\Omega, \tag{2.1.3}$$

which means that the Debye length is set to 1. Extension to general Debye length is straightforward. The doping profile c^{dop} is assumed to be constant w.r.t. time and to be bounded, i.e., $c^{\mathrm{dop}} \in L^\infty(\Omega)$.

One interpretation of c is the concentration of majority carriers (holes) in a p-type organic semiconductor with constant in time doping. Another interpretation of c is the cation concentration in an ionic liquid following the formal approach introduced in [71].

The system is supplemented with the prescription of the initial concentration

$$c_{|t=0} = c^0 \in L^\infty(\Omega) \quad \text{with} \quad 0 \leq c^0 \leq 1 \quad \text{and} \quad 0 < \overline{c} = \fint_\Omega c^0 \mathrm{d}\boldsymbol{x} < 1, \tag{2.1.4}$$

and of boundary conditions. The choice of the boundary conditions may depend on the targeted application: organic semiconductor or ionic liquid. For the analysis purpose, we consider boundary conditions which are well adapted to the ionic liquid model. Other boundary conditions will also be considered in the numerical simulations in Section 2.5. They are no-flux boundary conditions for the concentration:

$$\mathbf{J}\cdot\boldsymbol{n} = 0 \quad \text{on } (0,T)\times\partial\Omega. \tag{2.1.5}$$

The Poisson equation (2.1.3) is supplemented by inhomogeneous Dirichlet boundary conditions on a part Γ_D of $\partial\Omega$ with positive measure, and by homogeneous

Neumann boundary condition on the remaining part $\Gamma_N = \partial\Omega \setminus \Gamma_D$ of the boundary:

$$\Phi = \Phi^D \quad \text{on } (0,T)\times\Gamma_D, \qquad \nabla\Phi\cdot\boldsymbol{n} = 0 \quad \text{on } (0,T)\times\Gamma_N. \tag{2.1.6}$$

Throughout the paper, we assume that Φ^D is defined on the whole domain Ω and does not depend on time, with $\Phi^D \in H^1(\Omega)\cap L^\infty(\Omega)$.

The goal of this paper is to study and compare several different Finite Volume schemes for the system (2.1.1)–(2.1.6). They are based on various reformulations of the flux $\mathbf{J}$. Indeed, we may introduce either the so-called excess chemical potential [95] $\nu(c) = h(c) - \log(c) = -\log(1-c)$, or the activity and the inverse of the activity coefficient [103] respectively defined by $a(c) = e^{h(c)} = \frac{c}{1-c}$, and $\beta(c) = \frac{c}{a(c)} = 1-c$, or the diffusion enhancement [125] $r(c) = -\log(1-c)$ satisfying $r'(c) = ch'(c)$. Even though ν and r happen to be the same function for the nonlinearity considered in this paper, we keep different notations to emphasize their different physical meaning and expression in more complex systems. The different notations used throughout the paper are collected in Appendix 2.D. Then the flux $\mathbf{J}$, initially defined by (2.1.2), satisfies

$$\mathbf{J} = -\nabla c - c\nabla\left(\Phi + \nu(c)\right), \tag{2.1.7}$$
$$= -\beta(c)(\nabla a(c) + a(c)\nabla\Phi), \tag{2.1.8}$$
$$= -r'(c)\nabla c - c\nabla\Phi. \tag{2.1.9}$$

These formulations (2.1.2), (2.1.7), (2.1.8) and (2.1.9) lead to different schemes that we aim to compare from a numerical analysis point of view. We may notice that the flux $\mathbf{J}$ can also be expressed as

$$\mathbf{J} = -\nabla r(c) - c\nabla\Phi. \tag{2.1.10}$$

This last formulation will be used to define the weak solution to (2.1.1)–(2.1.6).

Before going to the discretization of the problem, let us highlight the entropy structure of system (2.1.1)–(2.1.6), which plays a central role in what follows.

2.1.3 Entropy structure and weak solutions

The goal of this section is to shortly depict the gradient flow structure of the system (2.1.1)–(2.1.6). We stay here at a formal level and remain sloppy about regularity issues. The solutions (c,Φ) to (2.1.1)–(2.1.6) are supposed to be regular enough so that the following calculations are justified. Define the mixing entropy density

$$H(c) = c\log(c) + (1-c)\log(1-c),$$

which is an antiderivative of h, then the electrochemical energy is given by

$$E(c,\Phi) = \int_\Omega \left\{ H(c) + \frac{1}{2}|\nabla\Phi|^2 \right\} \mathrm{d}\boldsymbol{x} - \int_{\Gamma_D} \Phi^D \nabla\Phi \cdot \boldsymbol{n} \mathrm{d}\boldsymbol{\gamma}. \tag{2.1.11}$$

The next proposition shows that the electrochemical energy is a Lyapunov functional. Moreover, the dissipation rate for the energy is explicitly given.

Proposition 2.1.1. *Let (c,Φ) be a smooth solution to (2.1.1)–(2.1.6), with c bounded away from 0 and 1, then*

$$\frac{\mathrm{d}}{\mathrm{d}t} E(c,\Phi) + \int_\Omega c\,|\nabla(h(c)+\Phi)|^2 \,\mathrm{d}\boldsymbol{x} = 0.$$

Proof. We notice first that since Φ^D does not depend on time,

$$\frac{\mathrm{d}}{\mathrm{d}t} E(c,\Phi) = \int_\Omega (h(c)\partial_t c + \nabla\Phi \cdot \partial_t \nabla\Phi) \mathrm{d}\boldsymbol{x} - \int_{\Gamma_D} \Phi^D \partial_t \nabla\Phi \cdot \boldsymbol{n} \mathrm{d}\boldsymbol{\gamma}.$$

Then we apply the Gauss theorem, and we use the Poisson equation (2.1.3) with a constant doping profile, to get

$$\frac{\mathrm{d}}{\mathrm{d}t} E(c,\Phi) = \int_\Omega (h(c)+\Phi)\partial_t c.$$

Multiplying the conservation law (2.1.1) by $h(c)+\Phi$ and integrating over the domain Ω yields

$$\int_\Omega \partial_t c(h(c)+\Phi) = -\int_\Omega c\,|\nabla(h(c)+\Phi)|^2 \,\mathrm{d}\boldsymbol{x},$$

thanks to the no-flux boundary condition (2.1.5). This concludes the proof of Proposition 2.1.1. □

Let $c \in L^\infty(\Omega;[0,1])$. We denote by $\Phi[c]$ the unique solution to (2.1.3). One can easily check that the energy functional $c \mapsto E(c,\Phi[c])$ is bounded on $L^\infty(\Omega;[0,1])$. Indeed, H takes values in $[-\log 2, 0]$ and the bounds on the electrical energy can be obtained by multiplying the Poisson equation by $\Phi - \Phi^D$ and Φ and integrating over Ω. Therefore, $E(c(t),\Phi(t))$ is finite for all $t>0$, whence a $L^\infty((0,T);H^1(\Omega))$ estimate on Φ. We also deduce from Proposition 2.1.1 that the total energy dissipation is bounded, i.e.

$$\int_0^T \int_\Omega c\,|\nabla(h(c)+\Phi)|^2 \,\mathrm{d}\boldsymbol{x}\mathrm{d}t \le C \tag{2.1.12}$$

for some C uniform with respect to the final time horizon T. Using again that $0 \leq c \leq 1$, we deduce from (2.1.12) that

$$\int_0^T \int_\Omega |\nabla r(c)|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq \int_0^T \int_\Omega c\,|\nabla h(c)|^2\,\mathrm{d}\boldsymbol{x}\mathrm{d}t \leq C. \tag{2.1.13}$$

The aforementioned $L^\infty((0,T);H^1(\Omega))$ estimate on the potential Φ and Estimate (2.1.13) on $r(c)$ suggest a notion of weak solution which is based on the expression (2.1.10) of the flux $\mathbf{J}$. In what follows, we denote the vector spaces:

$$\mathcal{H}_{\Gamma^D} = \{f \in H^1(\Omega), f_{|\Gamma_D} = 0\} \qquad \text{and} \qquad Q_T = (0,T)\times\Omega.$$

Definition 3. *A couple (c,Φ) is a weak solution of (2.1.1)–(2.1.6) if*

- $c \in L^\infty((Q_T;[0,1])$ *with* $r(c) \in L^2((0,T);H^1(\Omega))$
- $\Phi - \Phi^D \in L^\infty((0,T),\mathcal{H}_{\Gamma^D})$;
- *for all* $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$,

$$\iint_{Q_T} c\partial_t\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega c^0\varphi(0,\cdot)\mathrm{d}\boldsymbol{x} - \iint_{Q_T} (\nabla r(c) + c\nabla\Phi)\cdot\nabla\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0; \tag{2.1.14}$$

- *for all* $\psi \in \mathcal{H}_{\Gamma^D}$ *and almost all* $t \in (0,T)$,

$$\int_\Omega \nabla\Phi(t,\boldsymbol{x})\cdot\nabla\psi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega (c(t,\boldsymbol{x}) + c^{\mathrm{dop}}(\boldsymbol{x}))\psi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{2.1.15}$$

The goal of this paper is to compare from a numerical analysis point of view several different numerical schemes to approximate the solutions to (2.1.1)–(2.1.6). We pay particular attention to the preservation at the discrete level of the key properties of the continuous model, in particular concerning the preservation of the physical bounds $0 \leq c \leq 1$ and the energy/energy dissipation relation highlighted in Proposition 2.1.1. The definition of the Finite Volume approximation is detailed in the next section.

Existence of weak solutions to (2.1.1)–(2.1.6) is a by-product of Theorem 2.2.2 which states the convergence of some finite volume approximations towards weak solutions. As far as we know, there is no uniqueness result covering the model in its full generality. It seems to us that the closest uniqueness result is due to Gajewski [75] in the framework of bipolar drift-diffusion system. This proof requires an $L^\infty(Q_T)$ bound on the chemical potential $h(c)$, which has not been yet established for our system.

2.2 Finite Volume approximations

This section is organized as follows. First, in Section 2.2.1, we state the requirements on the mesh and fix some notations. Then in Section 2.2.2, we describe the common basis for the different schemes to be studied in this paper. All the methods presented in this paper rely on so-called two-point flux approximations, but four different schemes are introduced in Section 2.2.3 based on the formulations (2.1.2), (2.1.7), (2.1.8) and (2.1.9) of the flux $\mathbf{J}$. Then in Section 2.2.4, we state our two main results. The first one, namely Theorem 2.2.1, focuses on the case of a fixed mesh. We are interested in the existence of a solution to the nonlinear system corresponding to the schemes, and the dissipation of the energy at the discrete level. More precisely, one establishes that all the studied schemes satisfy a discrete counterpart to Proposition 2.1.1. Our second main result, namely Theorem 2.2.2, is devoted to the convergence of the scheme as the time step and the mesh size tend to 0.

2.2.1 Discretization of $(0,T)\times\Omega$

In this paper, we perform a parallel study of four numerical schemes based on two-point flux approximation (TPFA) finite volume schemes. As explained in [61, 67], this approach appears to be very efficient as soon as the continuous problems to be solved numerically are isotropic and one has the freedom to choose a suitable mesh fulfilling the so-called orthogonality condition [89, 68]. We recall here the definition of such a mesh, which is illustrated in Figure 2.1.

Definition 4. *An admissible mesh of Ω is a triplet $\left(\mathcal{T},\mathcal{E},(\boldsymbol{x}_K)_{K\in\mathcal{T}}\right)$ such that the following conditions are fulfilled.*

(i) Each control volume (or cell) $K\in\mathcal{T}$ is non-empty, open, polyhedral and convex. We assume that

$$K\cap L=\emptyset \quad \text{if } K,L\in\mathcal{T} \text{ with } K\neq L, \qquad \text{while} \quad \bigcup_{K\in\mathcal{T}}\overline{K}=\overline{\Omega}.$$

(ii) Each face $\sigma\in\mathcal{E}$ is closed and is contained in a hyperplane of $\mathbb{R}^d$, with positive $(d-1)$-dimensional Hausdorff (or Lebesgue) measure denoted by $m_\sigma=\mathcal{H}^{d-1}(\sigma)>0$. We assume that $\mathcal{H}^{d-1}(\sigma\cap\sigma')=0$ for $\sigma,\sigma'\in\mathcal{E}$ unless $\sigma'=\sigma$. For all $K\in\mathcal{T}$, we assume that there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K=\bigcup_{\sigma\in\mathcal{E}_K}\sigma$. Moreover, we suppose that $\bigcup_{K\in\mathcal{T}}\mathcal{E}_K=\mathcal{E}$. Given two distinct control volumes $K,L\in\mathcal{T}$, the intersection $\overline{K}\cap\overline{L}$ either reduces to a single face $\sigma\in\mathcal{E}$ denoted by $K|L$, or its $(d-1)$-dimensional Hausdorff measure is 0.

(iii) *The cell centers* $(\boldsymbol{x}_K)_{K\in\mathcal{T}}$ *belong to their cell:* $\boldsymbol{x}_K \in K$*, and are such that, if* $K, L \in \mathcal{T}$ *share a face* $K|L$*, then the vector* $\boldsymbol{x}_L - \boldsymbol{x}_K$ *is orthogonal to* $K|L$*.*

(iv) *For the boundary faces* $\sigma \subset \partial\Omega$*, we assume that either* $\sigma \subset \Gamma_D$ *or* $\sigma \subset \overline{\Gamma}_N$*. For* $\sigma \subset \partial\Omega$ *with* $\sigma \in \mathcal{E}_K$ *for some* $K \in \mathcal{T}$*, we assume additionally that there exists* $\boldsymbol{x}_\sigma \in \sigma$ *such that* $\boldsymbol{x}_\sigma - \boldsymbol{x}_K$ *is orthogonal to* σ*.*

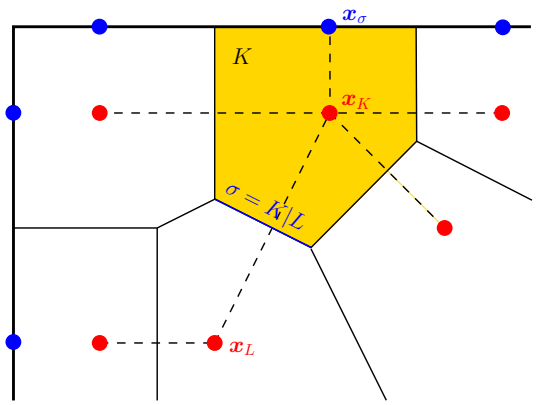


Figure 2.1 – Illustration of an admissible mesh as in Definition 4.

We denote by m_K the d-dimensional Lebesgue measure of the control volume K. The set of the faces is partitioned into two subsets: the set $\mathcal{E}_{\rm int}$ of the interior faces defined by $\mathcal{E}_{\rm int} = \{\sigma \in \mathcal{E} \mid \sigma = K|L \text{ for some } K, L \in \mathcal{T}\}$, and the set $\mathcal{E}_{\rm ext}$ of the exterior faces defined by $\mathcal{E}_{\rm ext} = \{\sigma \in \mathcal{E} \mid \sigma \subset \partial\Omega\}$, which can also be partitioned into $\mathcal{E}^D = \{\sigma \subset \Gamma_D\}$ and $\mathcal{E}^N = \{\sigma \subset \overline{\Gamma}_N\}$. For a given control volume $K \in \mathcal{T}$, we also define $\mathcal{E}_{K,\rm int}$ the set of its faces which belong to $\mathcal{E}_{\rm int}$. For such a face $\sigma \in \mathcal{E}_{K,\rm int}$, we may write $\sigma = K|L$, meaning that $\sigma = \overline{K} \cap \overline{L}$.

Given $\sigma \in \mathcal{E}$, we let

$$d_\sigma = \begin{cases} |\boldsymbol{x}_K - \boldsymbol{x}_L| & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ |\boldsymbol{x}_K - \boldsymbol{x}_\sigma| & \text{if } \sigma \in \mathcal{E}_{\rm ext}, \end{cases} \qquad \text{and} \quad \tau_\sigma = \frac{m_\sigma}{d_\sigma}.$$

We finally introduce the size $h_\mathcal{T}$ and the regularity $\zeta_\mathcal{T}$ (which is assumed to be positive) of a discretization $(\mathcal{T}, \mathcal{E}, (\boldsymbol{x}_K)_{K\in\mathcal{T}})$ of Ω by setting

$$h_\mathcal{T} = \max_{K\in\mathcal{T}} \operatorname{diam}(K), \qquad \zeta_\mathcal{T} = \min_{K\in\mathcal{T}} \min_{\sigma\in\mathcal{E}_K} \frac{d(x_K, \sigma)}{d_\sigma}.$$

Concerning the time discretization of $(0, T)$, we consider an increasing finite family of times $0 = t_0 < t_1 < \ldots, < t_N = T$. We denote by $\Delta t_n = t_n - t_{n-1}$ for $1 \le n \le N$, by $\boldsymbol{\Delta t} = (\Delta t_n)_{1\le n\le N}$, and by $\overline{\boldsymbol{\Delta t}} = \max_{1\le n\le N} \Delta t_n$.

2.2.2 A common basis for the Finite Volume schemes

All the numerical schemes studied in this paper are based on TPFA Finite Volumes. The initial data c_0 is discretized into $(c_K^0)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ by setting

$$c_K^0 = \frac{1}{m_K}\int_K c^0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \forall K \in \mathcal{T}, \tag{2.2.1}$$

while the doping profile c^{dop} is discretized into $\left(c_K^{\mathrm{dop}}\right)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ by

$$c_K^{\mathrm{dop}} = \frac{1}{m_K}\int_K c^{\mathrm{dop}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \forall K \in \mathcal{T}. \tag{2.2.2}$$

Assume that $\boldsymbol{c}^{n-1} = \left(c_K^{n-1}\right)_{K\in\mathcal{T}}$ is given for some $n > 0$, then we have to define how to compute $(\boldsymbol{c}^n, \boldsymbol{\Phi}^n) = (c_K^n, \Phi_K^n)_{K\in\mathcal{T}}$.

First, we introduce some notations. For all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_K$, we define the mirror values $c_{K\sigma}^n$ and $\Phi_{K\sigma}^n$ of c_K^n and Φ_K^n respectively across σ by setting

$$c_{K\sigma}^n = \begin{cases} c_L^n & \text{if } \sigma = K|L \in \mathcal{E}_{\mathrm{int}}, \\ c_K^n & \text{if } \sigma \in \mathcal{E}_{\mathrm{ext}}, \end{cases} \qquad \Phi_{K\sigma}^n = \begin{cases} \Phi_L^n & \text{if } \sigma = K|L \in \mathcal{E}_{\mathrm{int}}, \\ \Phi_K^n & \text{if } \sigma \in \mathcal{E}^N, \\ \Phi_\sigma^n = \frac{1}{m_\sigma}\int_\sigma \Phi^D \mathrm{d}\boldsymbol{\gamma} & \text{if } \sigma \in \mathcal{E}^D. \end{cases} \tag{2.2.3}$$

Given $\boldsymbol{u} = (u_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, we define the oriented and absolute jumps of $\boldsymbol{u}$ across any edge by

$$D_{K\sigma}\boldsymbol{u} = u_{K\sigma} - u_K, \quad D_\sigma \boldsymbol{u} = |D_{K\sigma}\boldsymbol{u}|, \qquad \forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_K.$$

We consider a backward Euler scheme in time and a TPFA finite volume scheme in space. It is written as follows:

$$-\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D_{K\sigma}\boldsymbol{\Phi}^n = m_K\left(c_K^n + c_K^{\mathrm{dop}}\right), \qquad \forall K \in \mathcal{T}, \tag{2.2.4a}$$

$$m_K\frac{c_K^n - c_K^{n-1}}{\Delta t_n} + \sum_{\sigma\in\mathcal{E}_{K,\mathrm{int}}} F_{K\sigma}^n = 0, \qquad \forall K \in \mathcal{T}, \tag{2.2.4b}$$

where $F_{K\sigma}^n$ should be a conservative and consistent approximation of $\frac{1}{\Delta t_n}\int_{t_{n-1}}^{t_n}\int_\sigma \mathbf{J}\cdot \boldsymbol{n}_{K\sigma}$ ($\boldsymbol{n}_{K\sigma}$ denotes the normal to σ outward K). The explicit formulas relating the numerical fluxes $F_{K\sigma}^n$ to the primary unknowns are now the only remaining degree of freedom. Four possible choices are given in the next section.

2.2.3 Numerical fluxes for the conservation of the chemical species

To close the system (2.2.4a)–(2.2.4b), it remains to define the numerical fluxes $F^n_{K\sigma}$.

Due to the no-flux boundary condition we only have to define the inner fluxes. They are defined with a function $\mathcal{F}$ of the primary unknowns $(c^n_K, c^n_L, \Phi^n_K, \Phi^n_L)$:

$$F^n_{K\sigma} = \tau_\sigma \mathcal{F}(c^n_K, c^n_L, \Phi^n_K, \Phi^n_L), \quad \forall K \in \mathcal{T}, \forall \sigma = K|L. \tag{2.2.5}$$

We discuss now four strategies that are based on the four expressions (2.1.2), (2.1.7), (2.1.8), and (2.1.9). They lead to different formulas for $\mathcal{F}$. Three of the discrete fluxes are extensions of the Scharfetter-Gummel scheme [116] and let the Bernoulli function $B(u) = \frac{u}{e^u - 1}$, with $B(0) = 1$, appear in their definition.

All the functions $\mathcal{F}$ defined below verify

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = -\mathcal{F}(c_L, c_K, \Phi_L, \Phi_K) \quad \forall (c_K, c_L, \Phi_K, \Phi_L) \in (0,1)^2 \times \mathbb{R}^2,$$

so that the numerical fluxes are locally conservative, which means

$$F^n_{K\sigma} + F^n_{L\sigma} = 0 \quad \forall \sigma = K|L \in \mathcal{E}_{\text{int}}. \tag{2.2.6}$$

2.2.3.a The centred flux

The so-called centred flux is derived from formula (2.1.2), which suggests the following definition of $\mathcal{F}$:

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = -\frac{c_K + c_L}{2} D_{K\sigma} \left(h(\boldsymbol{c}) + \boldsymbol{\Phi} \right). \tag{C}$$

The associate flux can be seen as a particular case in the TPFA context of the fluxes introduced in [36, 33, 28, 39] in various multipoint flux approximations (MPFA) or finite element contexts. In opposition to the three next schemes, the centred scheme is not based on the Scharfetter-Gummel scheme. We can notice that even if the relation (2.1.10) between the flux and the concentration were be linear (i.e., if $h(c) = \log(c)$ so that $r(c) = c$), $\mathcal{F}$ would be nonlinear with respect to c_K and c_L, and also singular near 0.

2.2.3.b The Sedan flux

The second flux we introduce is named Sedan after the eponymous code SEDAN III [126] [1]. Formula (2.1.7) for the flux $\mathbf{J}$ suggests to use a classical Scharfetter-Gummel scheme, but for a modified potential $\Phi + \nu(c)$ instead of only Φ, leading to the following definition of $\mathcal{F}$:

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = B\Big(D_{K\sigma}(\mathbf{\Phi} + \nu(\boldsymbol{c}))\Big)c_K - B\Big(-D_{K\sigma}(\mathbf{\Phi} + \nu(\boldsymbol{c}))\Big)c_L. \qquad \text{(S)}$$

Remark 2.2.1. We notice that the Sedan flux defined by (S) satisfies

$$\mathcal{F}(c_K, c_L, \Phi, \Phi) = r(c_K) - r(c_L), \quad \forall (c_K, c_L) \in (0,1) \times (0,1), \ \forall \Phi \in \mathbb{R}.$$

It means that when $\mathbf{J} = -\nabla r(c)$, we recover the classical two-point flux approximation:

$$F^n_{K\sigma} = \tau_\sigma(r(c^n_K) - r(c^n_L)), \quad \forall K \in \mathcal{T}, \forall \sigma = K|L.$$

2.2.3.c The activity based flux

The activity based flux we discuss now is a restriction to our simplified model of the flux introduced in [70, 74]. It relies on the expression (2.1.8) of the flux $\mathbf{J}$. Assume that $a(c)$ and $\beta(c)$ are independent one from another (even though this is of course not true), then the flux $\mathbf{J}$ is linear w.r.t. $a(c)$, while $\beta(c)$ is a multiplicative factor. This suggests choosing a particular average for $\beta(c)$ —here the arithmetic mean— and applying the Scharfetter-Gummel scheme to approximate $-\nabla a(c) - a(c)\nabla\Phi$. This yields

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = \frac{\beta(c_K) + \beta(c_L)}{2}\Big\{B(D_{K\sigma}\mathbf{\Phi})a(c_K) - B(-D_{K\sigma}\mathbf{\Phi})a(c_L)\Big\}. \qquad \text{(AB)}$$

1. The online reference contains a link to a tar file http://www-tcad.stanford.edu/oldftp_sw/Sedan-III/relB.8830.tar.Z containing among others the FORTRAN source `diffg.f` which in the lines 53-56 contains

```
if(ferm) then
  dpsin=dpsin+dlog(gamn(ip1)/gamn(i))
  dpsip=dpsip-dlog(gamp(ip1)/gamp(i))
endif
```

Here, `dpsin` and `dpsip` are the arguments of the Bernoulli function and `ferm` is the switch for enabling the degenerate case (Fermi statistics). To our knowledge this is the earliest reference to this scheme.

2.2.3.d The Bessemoulin-Chatard flux

The last numerical flux we consider here is named Bessemoulin-Chatard flux after the author's name of [15]. Formula (2.1.9) for the flux $\mathbf{J}$ suggests that, up to the introduction of a variable diffusion coefficient approximating the quantity $r'(c)$ per face, one can use the Scharfetter-Gummel scheme. Following [15], the approximation $\mathfrak{d}r(c_K, c_L)$ of $r'(c)$ is defined as

$$\mathfrak{d}r(c_K, c_L) = \begin{cases} \dfrac{h(c_K) - h(c_L)}{\log(c_K) - \log(c_L)} & \text{if } c_K \neq c_L, \\ r'(c_K) & \text{if } c_K = c_L. \end{cases}$$

This leads to the following definition of $\mathcal{F}$:

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = \mathfrak{d}r(c_K, c_L) \left\{ B\left(\frac{D_{K\sigma}\mathbf{\Phi}}{\mathfrak{d}r(c_K, c_L)} \right) c_K - B\left(-\frac{D_{K\sigma}\mathbf{\Phi}}{\mathfrak{d}r(c_K, c_L)} \right) c_L \right\}. \tag{BC}$$

2.2.4 Main results and organisation of the paper

We have introduced four schemes defined by (2.2.1)–(2.2.5), supplemented with one of the four definitions of $\mathcal{F}$: (C), (S), (AB), or (BC). Besides numerical comparisons between the different approaches —this will be the purpose of Section 2.5—, we aim at proposing shared pieces of numerical analysis for all the schemes.

All the four schemes proposed above yield a nonlinear system to be solved at each time step. The first theorem proven is this paper concerns the existence of discrete solutions for a given mesh, and the preservation of the physical bounds: boundedness of the concentration between 0 and 1, decay of the energy. The discrete energy functional $E_{\mathcal{T}}$ is the discrete counterpart of the continuous energy functional E, defined by:

$$E_{\mathcal{T}}(\boldsymbol{c}^n, \mathbf{\Phi}^n) = \sum_{K\in\mathcal{T}} m_K H(c_K^n) + \frac{1}{2}\sum_{\sigma\in\mathcal{E}} \tau_\sigma \left(D_\sigma \mathbf{\Phi}^n\right)^2 - \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D\cap\mathcal{E}_K} \tau_\sigma \Phi_\sigma^D D_{K\sigma}\mathbf{\Phi}^n. \tag{2.2.7}$$

As stated in Theorem 2.2.1 below, the nonlinear system corresponding to each scheme admits a solution which preserves the physical bounds on the concentrations and the decay of the energy. The proof of Theorem 2.2.1 will be the purpose of Section 2.3.

Theorem 2.2.1. *Let $(\mathcal{T}, \mathcal{E}, (\boldsymbol{x}_K)_{K\in\mathcal{T}})$ be an admissible mesh and let $\boldsymbol{c}^0$ be defined by (2.2.1). Then, for all $1 \leq n \leq N$, the nonlinear system of equations*

(2.2.3)–(2.2.5), *supplemented either with* (C), (S), (AB), *or* (BC), *has a solution* $(\boldsymbol{c}^n, \boldsymbol{\Phi}^n) \in [0,1]^{\mathcal{T}} \times \mathbb{R}^{\mathcal{T}}$. *Moreover, the solution to the scheme satisfies, for all* $1 \leq n \leq N$,

$$E_{\mathcal{T}}(\boldsymbol{c}^{n,}, \boldsymbol{\Phi}^n) \leq E_{\mathcal{T}}(\boldsymbol{c}^{n-1}, \boldsymbol{\Phi}^{n-1}) \text{ and } 0 < c_K^n < 1, \quad \forall K \in \mathcal{T}.$$

Knowing a discrete solution to the scheme, $(\boldsymbol{c}^n, \boldsymbol{\Phi}^n)_{1\leq n\leq N}$, we can define an approximate solution $(c_{\mathcal{T},\boldsymbol{\Delta t}}, \Phi_{\mathcal{T},\boldsymbol{\Delta t}})$. It is the piecewise constant function defined almost everywhere by

$$c_{\mathcal{T},\boldsymbol{\Delta t}}(t,\boldsymbol{x}) = c_K^n, \quad \Phi_{\mathcal{T},\boldsymbol{\Delta t}}(t,\boldsymbol{x}) = \Phi_K^n \quad \text{if } (t,\boldsymbol{x}) \in (t_{n-1}, t_n] \times K.$$

This definition will be developed in Section 2.4 and supplemented by other reconstruction operators.

Let $(\mathcal{T}_m, \mathcal{E}_m, (\boldsymbol{x}_K)_{K\in\mathcal{T}_m})_{m\geq 1}$ be a sequence of admissible meshes in the sense of Definition 4 such that $h_{\mathcal{T}_m}, \overline{\boldsymbol{\Delta t}}_m \underset{m\to\infty}{\longrightarrow} 0$ while the mesh regularity remains bounded, i.e., $\zeta_{\mathcal{T}_m} \geq \zeta^\star$ for some $\zeta^\star > 0$ not depending on the mesh m. A natural question is the convergence of the associated sequence of approximate solutions $(c_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}, \Phi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m})_{m\geq 1}$ towards a weak solution to the continuous problem. The convergence result is stated in Theorem 2.2.2, only for the centred scheme and the Sedan scheme.

Theorem 2.2.2. *For the centred scheme (inner fluxes defined by* (2.2.5) *and* (C)*) and the Sedan scheme (inner fluxes defined by* (2.2.5) *and* (S)*), a sequence of approximate solutions* $(c_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}, \Phi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m})_{m\geq 1}$ *satisfies, up to a subsequence,*

$$c_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} c \quad \text{a.e. in } Q_T, \qquad \Phi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \Phi \quad \text{in } L^2(Q_T), \tag{2.2.8}$$

where (c, Φ) *is a weak solution to* (2.1.1)-(2.1.6) *in the sense of Definition 3.*

The above theorem deserves some comments. First, the convergence proof carried out in what follows does not encompass the activity based scheme and the Bessemoulin-Chatard scheme for reasons that will appear clearly in the proof later on. This does of course not mean that these schemes do not converge, but only that our analysis does not cover them. Second, the topologies for which the convergence is claimed in (2.2.8) is suboptimal when compared to the results we prove in Section 2.4. However, we choose to keep the statement as simple as possible. The interested reader can refer to Section 2.4 to get finer results, including the convergence of approximate gradients to be defined later on.

Section 2.5 is then devoted to the comparison of the numerical results produced by the different schemes.

2.3 Numerical analysis for fixed meshes

In this section, one aims to show that each scheme admits at least one solution and that the physical bounds are preserved by the schemes. Our approach is based on a topological degree argument [98, 56] to be detailed in Section 2.3.3. It relies on a priori estimates to be stated in Section 2.3.2. Let us start by some preliminary properties of the different functions $\mathcal{F}$, defined either by (C), (S), (AB), or (BC), and some consequences for the inner numerical fluxes $F_{K\sigma}^n$.

2.3.1 Face concentration and face dissipation

For each flux $\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L)$, we want to define a face concentration function $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L)$ satisfying

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = \mathcal{C}(c_K, c_L, \Phi_K, \Phi_L)\,(h(c_K) + \Phi_K - h(c_L) - \Phi_L)\,.$$

Lemma 2.3.1 states that the face concentration functional $\mathcal{C}$ can be continuously defined on $(0,1)\times(0,1)\times\mathbb{R}\times\mathbb{R}$ and that it verifies some bounds. Let us also note that it clearly satisfies $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) = \mathcal{C}(c_L, c_K, \Phi_L, \Phi_K)$.

Lemma 2.3.1. *For a flux $\mathcal{F}$ defined either by (C), (S), (AB) or (BC), the corresponding face concentration functional defined by*

$$\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) = \frac{\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L)}{h(c_K) + \Phi_K - h(c_L) - \Phi_L} \tag{2.3.1}$$

if $h(c_K)+\Phi_K-h(c_L)-\Phi_L \neq 0$ can be extended by continuity on $(0,1)\times(0,1)\times\mathbb{R}\times\mathbb{R}$. Moreover, if $\mathcal{F}$ is defined by (AB) , we have for all $(c_K, c_L, \Phi_K, \Phi_L) \in (0,1)\times(0,1)\times\mathbb{R}\times\mathbb{R}$,

$$\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) \geq \frac{\min(c_K, c_L)}{2} > 0. \tag{2.3.2}$$

In the other cases, $\mathcal{C}$ verifies a stronger result: for all $(c_K, c_L, \Phi_K, \Phi_L) \in (0,1)\times(0,1)\times\mathbb{R}\times\mathbb{R}$,

$$\min(c_K, c_L) \leq \mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) \leq \max(c_K, c_L). \tag{2.3.3}$$

Proof. We first remark that, for the centred flux (C),

$$\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) = \frac{c_K + c_L}{2}.$$

Therefore, $\mathcal{C}$ is well defined in $(0,1)\times(0,1)\times\mathbb{R}\times\mathbb{R}$ and it satisfies the bounds (2.3.3).

The proof is more intricate for the Sedan flux (S) and the Bessemoulin-Chatard flux (BC). It relies on an elementary property of the Bernoulli function, which writes:

$$B(\log(a)-\log(b))a - B(\log(b)-\log(a))b = 0, \qquad \forall (a,b)\in(0,1)^2. \tag{2.3.4}$$

Let us consider first the Bessemoulin-Chatard flux (BC). Applying (2.3.4) with $a = c_K$ and $b = c_L$, we obtain, with $x = \log(c_K/c_L)$ and $y = (\Phi_L - \Phi_K)/\mathfrak{d}r(c_K, c_L)$,

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = \mathfrak{d}r(c_K, c_L)\Big((B(y) - B(x))c_K + (B(-y) - B(-x))c_L\Big).$$

But, we also notice that

$$\mathfrak{d}r(c_K, c_L)(x - y) = h(c_K) + \Phi_K - h(c_L) - \Phi_L,$$

so that (2.3.1) yields

$$\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) = \frac{B(y) - B(x)}{x - y}c_K + \frac{B(-x) - B(-y)}{x - y}c_L \tag{2.3.5}$$

if $h(c_K)+\Phi_K-h(c_L)-\Phi_L \neq 0$, which means $x-y \neq 0$. First, we remark that this definition can be extended if $x-y \to 0$, so that $\mathcal{C}$ is defined in $(0,1)\times(0,1)\times\mathbb{R}\times\mathbb{R}$. Then, as the Bernoulli function is decreasing and satisfies $B(x)-B(-x) = -x$ for all $x \in \mathbb{R}$, which implies

$$\frac{B(y) - B(x)}{x - y} + \frac{B(-x) - B(-y)}{x - y} = 1,$$

we obtain that $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L)$ is a convex combination of c_K and c_L. Therefore, (2.3.3) holds for the Bessemoulin-Chatard flux.

The proof is similar for the Sedan flux (S). Indeed, we still establish (2.3.5), but with $x = \log(c_K/c_L)$ and $y = \Phi_L + \nu(c_L) - \Phi_K - \nu(c_K)$, so that $x - y = h(c_K) + \Phi_K - h(c_L) - \Phi_L$. Here again, $\mathcal{C}$ is well defined in $(0,1)\times(0,1)\times\mathbb{R}\times\mathbb{R}$ and $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L)$ is a convex combination of c_K and c_L, so that (2.3.3) holds for the Sedan flux.

The fact that (2.3.3) does not hold for the activity based flux (AB) is illustrated on Figure 2.2. Nevertheless, one can express the corresponding face concentration

under the form

$$\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) = \frac{\beta(c_K) + \beta(c_L)}{2} \times \left(\frac{B(y) - B(x)}{x - y} a(c_K) + \frac{B(-x) - B(-y)}{x - y} a(c_L) \right),$$

with $x = \log(a(c_K)) - \log(a(c_L))$ and $y = \Phi_L - \Phi_K$. Therefore, $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L)$ is the product of the arithmetic mean of the positive quantities $\beta(c_K)$ and $\beta(c_L)$ with a convex combination of the positive quantities $a(c_K)$ and $a(c_L)$. As a is increasing, this convex combination is bounded by below by $a(\min(c_K, c_L))$. Using the identity $\beta(c)a(c) = c$, we get (2.3.2). □

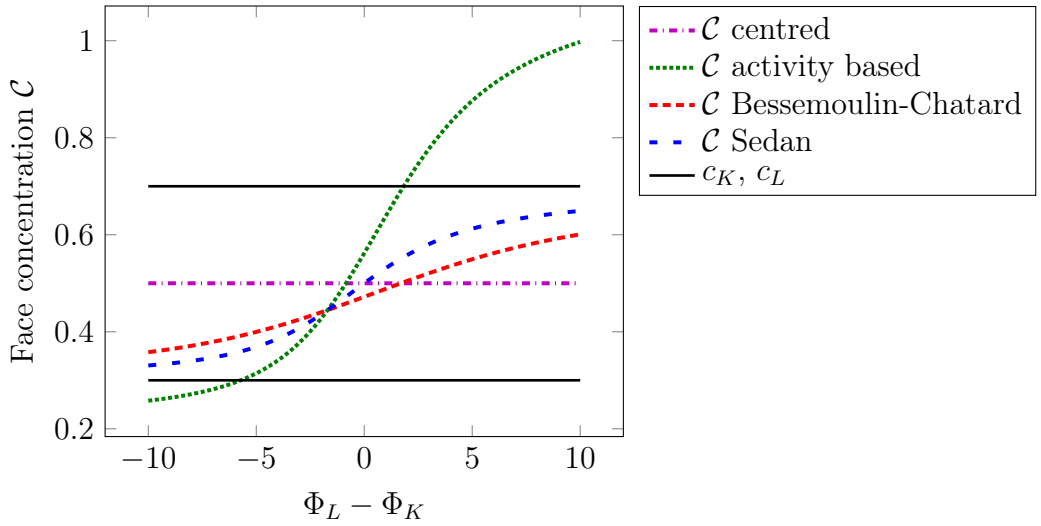


Figure 2.2 – Evolution of the face concentration $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L)$ as a function of the jump of the potential $\Phi_L - \Phi_K$ for the choice $c_K = 0.3$ and $c_L = 0.7$.

Using this result we define one face concentration by internal face and by choice of flux:

$$\mathcal{C}_\sigma^n = \mathcal{C}\left(c_K^n, c_L^n, \Phi_K^n, \Phi_L^n\right) \quad \forall \sigma \in \mathcal{E}_{\text{int}}, \sigma = K|L. \tag{2.3.6}$$

Each flux $F_{K\sigma}^n$ can be rewritten as

$$F_{K\sigma}^n = -\tau_\sigma \mathcal{C}_\sigma^n D_{K\sigma}(h(\boldsymbol{c}^n) + \boldsymbol{\Phi}^n), \quad \forall K \in \mathcal{T}, \forall \sigma = K|L. \tag{2.3.7}$$

We also introduce a face dissipation functional $\mathcal{D} : (0,1) \times (0,1) \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined by

$$\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L) = \mathcal{C}\left(c_K, c_L, \Phi_K, \Phi_L\right) \left|h(c_K) - h(c_L) + \Phi_K - \Phi_L\right|^2. \tag{2.3.8}$$

We set, for each scheme:

$$\mathcal{D}_\sigma^n = \mathcal{D}(c_K^n, c_L^n, \Phi_K^n, \Phi_L^n), \quad \forall \sigma \in \mathcal{E}_{\rm int}, \sigma = K|L. \tag{2.3.9}$$

For $\delta \in (0,1)$ and $M \in \mathbb{R}$, we finally define two functions associated to $\mathcal{D}$, $\Psi_{\delta,M} : (0,1) \to \mathbb{R}$ and $\Upsilon_{\delta,M} : (0,1) \to \mathbb{R}$, by

$$\begin{aligned}
\Psi_{\delta,M}(c_L) &= \inf\{\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L); c_K \in [\delta, 1), (\Phi_K, \Phi_L) \in [-M, M]^2\}, \\
\Upsilon_{\delta,M}(c_L) &= \inf\{\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L); c_K \in (0, 1-\delta], (\Phi_K, \Phi_L) \in [-M, M]^2\}.
\end{aligned} \tag{2.3.10}$$

Note that $\delta \mapsto \Psi_{\delta,M}(c_L)$ and $\delta \mapsto \Upsilon_{\delta,M}(c_L)$ are nondecreasing for all $M \in \mathbb{R}$ and all $c_L \in (0,1)$.

As a by-product of Lemma 2.3.1, we obtain that the face dissipation $\mathcal{D}$ is a nonnegative function as the product of nonnegative quantities. Lemma 2.3.2 is about the coercivity of the face dissipation functional. As its proof is technical, it is given in Appendix 2.B.

Lemma 2.3.2. *The face dissipation functional defined by* (2.3.8) *and either* (C), (S), (AB) *or* (BC) *satisfies the following dissipation property: given* $\delta \in (0,1)$ *and* $M \in \mathbb{R}$, *for* Ψ *and* Υ *as defined in* (2.3.10)*:*

$$\begin{aligned}
\lim_{c_L \to 0} \Psi_{\delta,M}(c_L) &= +\infty, \\
\lim_{c_L \to 1} \Upsilon_{\delta,M}(c_L) &= +\infty.
\end{aligned}$$

2.3.2 Uniform a priori estimates

In all this section, we assume that $(\boldsymbol{c}^n, \boldsymbol{\Phi}^n)_{1 \le n \le N}$ is a solution to the scheme (2.2.3)–(2.2.5) with a numerical flux defined among (C) , (S), (AB), and (BC). We also assume that this solution verifies: $0 < c_K^n < 1$ for all $K \in \mathcal{T}$ and all $1 \le n \le N$. Then the goal of this section is to derive enough a priori estimates on $(\boldsymbol{c}^n, \boldsymbol{\Phi}^n)_{1 \le n \le N}$ in order to show the existence of a weak solution to the nonlinear system induced by the scheme.

The first lemma is the discrete counterpart of the global conservation of mass.

Lemma 2.3.3. *One has*

$$\sum_{K \in \mathcal{T}} m_K c_K^n = \sum_{K \in \mathcal{T}} m_K c_K^{n-1} = \int_\Omega c^0 \mathrm{d}\boldsymbol{x}, \quad \forall 1 \le n \le N.$$

Proof. The first equality is obtained by summing (2.2.4b) over $K \in \mathcal{T}$ and by

using the conservativity of the fluxes (2.2.6). A straightforward induction ensures the second equality thanks to (2.2.1). □

The second a priori estimate is related to energy dissipation and can be seen as a discrete counterpart of Proposition 2.1.1.

Proposition 2.3.1. *Let $\mathcal{D}_\sigma^n$ be defined by (2.3.9), and $E_\mathcal{T}$ by (2.2.7). One has:*

$$\frac{E_\mathcal{T}(\boldsymbol{c}^n, \boldsymbol{\Phi}^n) - E_\mathcal{T}(\boldsymbol{c}^{n-1}, \boldsymbol{\Phi}^{n-1})}{\Delta t_n} \leq -\sum_{\sigma \in \mathcal{E}_{\rm int}} \tau_\sigma \mathcal{D}_\sigma^n \leq 0, \quad \forall 1 \leq n \leq N.$$

Proof. Due to the convexity of H and of $x \mapsto x^2/2$, we have:

$$E_\mathcal{T}(\boldsymbol{c}^n, \boldsymbol{\Phi}^n) - E_\mathcal{T}(\boldsymbol{c}^{n-1}, \boldsymbol{\Phi}^{n-1}) \leq \sum_{K \in \mathcal{T}} m_K (c_K^n - c_K^{n-1}) h(c_K^n) + \sum_{\sigma \in \mathcal{E}} \tau_\sigma D_\sigma \boldsymbol{\Phi}^n D_\sigma (\boldsymbol{\Phi}^n - \boldsymbol{\Phi}^{n-1}) - \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}^D \cap \mathcal{E}_K} \tau_\sigma \Phi_\sigma^D D_{K\sigma} (\boldsymbol{\Phi}^n - \boldsymbol{\Phi}^{n-1}).$$

A discrete integration by parts permits to rewrite the sum of the last two terms, which, combined to the scheme (2.2.4a), leads to

$$E_\mathcal{T}(\boldsymbol{c}^n, \boldsymbol{\Phi}^n) - E_\mathcal{T}(\boldsymbol{c}^{n-1}, \boldsymbol{\Phi}^{n-1}) \leq \sum_{K \in \mathcal{T}} m_K (c_K^n - c_K^{n-1})(h(c_K^n) + \Phi_K^n). \tag{2.3.11}$$

Multiplying the equation (2.2.4b) by $h(c_K^n) + \Phi_K^n$ and summing over $K \in \mathcal{T}$, we obtain that

$$\begin{aligned} \sum_{K \in \mathcal{T}} m_K \frac{c_K^n - c_K^{n-1}}{\Delta t_n} (h(c_K^n) + \Phi_K^n) &= -\sum_{K \in \mathcal{T}} (h(c_K^n) + \Phi_K^n) \sum_{\sigma \in \mathcal{E}_K} F_{K\sigma}^n \\ &= \sum_{\sigma \in \mathcal{E}_{\rm int}} F_{K\sigma}^n (h(c_L^n) + \Phi_L^n - h(c_K^n) + \Phi_K^n) \\ &= -\sum_{\sigma \in \mathcal{E}_{\rm int}} \tau_\sigma \mathcal{C}_\sigma^n \left| D_\sigma \left(h(\boldsymbol{c}^n) + \boldsymbol{\Phi}^n \right) \right|^2 \end{aligned} \tag{2.3.12}$$

Combining (2.3.11) and (2.3.12) provides the desired estimate.

□

The third statement of this section is devoted to a uniform L^∞ estimate of $(\boldsymbol{\Phi}^n)_{1 \leq n \leq N}$. It is a straightforward consequence of the slightly more general Proposition 2.A.1 stated in appendix, together with the a priori bounds $0 < c_K^n < 1$ and $-\|c^{\rm dop}\|_\infty \leq c_K^{\rm dop} \leq \|c^{\rm dop}\|_\infty$.

Lemma 2.3.4. *There exists M_Φ depending only on Φ^D, $c^{\rm dop}$ and Ω such that*

$$|\Phi_K^n| \le M_\Phi, \qquad \forall K \in \mathcal{T},\ \forall 1 \le n \le N.$$

The next lemma concerns the discrete $L^\infty((0,T);H^1(\Omega))$ estimate on the electric potential and the control of the discrete dissipation.

Lemma 2.3.5. *There exists C depending only on Φ^D, $c^{\rm dop}$, Ω and $\zeta_\mathcal{T}$, and C' depending also on c^0 such that:*

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma |D_\sigma \mathbf{\Phi}^n|^2 \le C, \qquad \forall 1 \le n \le N; \tag{2.3.13}$$

$$\left| \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D\cap\mathcal{E}_K} \tau_\sigma \Phi_\sigma^D D_{K\sigma}\mathbf{\Phi}^n \right| \le C, \qquad \forall 1 \le n \le N; \tag{2.3.14}$$

$$\sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \mathcal{D}_\sigma^n \le C'. \tag{2.3.15}$$

Proof. As $\Phi^D \in L^\infty \cap H^1(\Omega)$, it is discretized into $\mathbf{\Phi}^D \in \mathbb{R}^\mathcal{T}$ by setting

$$\Phi_K^D = \frac{1}{m_K}\int_K \Phi^D {\rm d}\boldsymbol{x}, \quad \forall K \in \mathcal{T}, \quad \text{and } \Phi_\sigma^D = \frac{1}{m_\sigma}\int_\sigma \Phi^D {\rm d}\boldsymbol{\gamma}, \qquad \forall \sigma \in \mathcal{E}^D.$$

It satisfies $|\Phi_K^D| \le \|\Phi^D\|_\infty$ for all $K \in \mathcal{T}$. Multiplying (2.2.4a) by $\Phi_K^n - \Phi_K^D$ and summing over $K \in \mathcal{T}$ provides

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\mathbf{\Phi}^n D_{K\sigma}(\mathbf{\Phi}^n - \mathbf{\Phi}^D) = \sum_{K\in\mathcal{T}} m_K (c_K^n + c_K^{\rm dop})(\Phi_K^n - \Phi_K^D). \tag{2.3.16}$$

Using the elementary inequality $a(a-b) \ge \frac{a^2-b^2}{2}$, we get that

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\mathbf{\Phi}^n D_{K\sigma}(\mathbf{\Phi}^n - \mathbf{\Phi}^D) \ge \frac{1}{2}\sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_\sigma \mathbf{\Phi}^n)^2 - \frac{1}{2}\sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_\sigma \mathbf{\Phi}^D)^2.$$

Using the boundedness of c_K^n, $c_K^{\rm dop}$, Φ_K^D, and of Φ_K^n (cf. Lemma 2.3.4), we obtain that the right-hand side of (2.3.16) is bounded:

$$\sum_{K\in\mathcal{T}} m_K (c_K^n + c_K^{\rm dop})(\Phi_K^n - \Phi_K^D) \le C.$$

Following [68, Lemma 13.4], there exists C depending only on $\zeta_{\mathcal{T}}$ such that

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_\sigma \mathbf{\Phi}^D)^2 \leq C \|\Phi^D\|^2_{H^1(\Omega)},$$

which concludes the proof of (2.3.13). Multiplying now the scheme (2.2.4a) by Φ^n_K and summing over $K \in \mathcal{T}$ leads to (2.3.14) by following the same kind of computations. Finally, these two inequality ensures that the functional $\boldsymbol{c}^n \mapsto E_{\mathcal{T}}(\boldsymbol{c}^n, \mathbf{\Phi}[\boldsymbol{c}^n])$ is bounded on $(0,1)^{\mathcal{T}}$. Therefore, Proposition 2.3.1 yields the control of the dissipation (2.3.15). □

As the last step before establishing the existence of a solution to the scheme, we show that the approximate concentrations $\boldsymbol{c}^n$ are bounded away from 0 and 1. Note that contrary to Lemmas 2.3.3 and 2.3.4 and to Proposition 2.3.1, the estimate of the following Lemma is not uniform with respect to mesh size and time step.

Lemma 2.3.6. *There exists $\epsilon > 0$ depending on $\mathcal{T}, \boldsymbol{\Delta t}$, Φ^D, $\bar{c}$, c^{dop} and Ω such that*

$$\epsilon < c^n_K < 1 - \epsilon, \qquad \forall K \in \mathcal{T},\ \forall 1 \leq n \leq N.$$

Proof. The proof follows the idea of [35, Lemma 3.10] (see also [36, Lemma 3.7]). Let us establish the lower bound only since the outline of the proof of the upper bound is similar.

Because of assumption (2.1.4) on the initial data and of the choice (2.2.1) for its discretization, one knows that

$$\frac{1}{m_\Omega} \sum_{K\in\mathcal{T}} m_K c^0_K = \bar{c} \in (0,1).$$

Lemma 2.3.3 ensures the conservation of mass, so that

$$\frac{1}{m_\Omega} \sum_{K\in\mathcal{T}} m_K c^n_K = \bar{c} \in (0,1), \qquad \forall n \geq 1.$$

This implies that there exists $K_0 \in \mathcal{T}$ such that $c^n_{K_0} \geq \bar{c} > 0$. We set $\delta_0 = \bar{c}$.

Denote by $\mathbf{\Phi}[\boldsymbol{c}^n]$ the unique solution to the linear system (2.2.4a). The estimate (2.3.15) of Lemma 2.3.5 ensures that there exists $C_{\mathcal{D}}$ depending (among others) on Δt_n such that

$$\mathcal{D}^n_{\mathcal{T}} = \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \tau_\sigma \mathcal{D}^n_\sigma \leq C_{\mathcal{D}}, \qquad \forall 1 \leq n \leq N. \tag{2.3.17}$$

In particular, for all face $\sigma \in \mathcal{E}_{K_0}$, one gets that $\tau_\sigma \mathcal{D}_\sigma^n \leq C_\mathcal{D}$. Therefore, the concentration $c_{K_1}^n$ in any neighbouring cell K_1 of K_0 is bounded away from 0 by

$$\begin{aligned} c_{K_1}^n &\geq \inf\left\{c_L \in (0,1)\ ;\ \Psi_{c_{K_0}^n, M_\Phi}(c_L) \leq C_\mathcal{D}/\tau_\sigma\right\} \\ &\geq \inf\left\{c_L \in (0,1)\ ;\ \Psi_{\delta_0, M_\Phi}(c_L) \leq C_\mathcal{D}/\min_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma\right\} =: \delta_1 > 0 \end{aligned}$$

thanks to the monotonicity of $\delta \mapsto \Psi_{\delta,M}(c_L)$. Owing to Lemma 2.3.2, the above right-hand side is bounded away from 0 by some quantity that might also depend on $\mathcal{T}$ because of the presence of τ_σ. This lower bound can be set to δ_1, and we can then iterate the procedure to the neighbouring cells of K_1, and so on. Since the mesh is finite, only a finite number of iterations $I_\mathcal{T}$ is needed to cover all the cells, whence a uniform lower bound on c_K^n: $\epsilon = \min_{1\leq i\leq I_\mathcal{T}} \delta_i$, where

$$\delta_{i+1} = \inf\left\{c_L \in (0,1)\ ;\ \Psi_{\delta_i, M_\Phi}(c_L) \leq C_\mathcal{D}/\min_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma\right\} > 0, \quad \delta_0 = \overline{c}.$$

□

2.3.3 Existence of a solution to the schemes

Based on the estimates derived in the previous section, we can establish the existence of at least one solution to each scheme. This completes the proof of Theorem 2.2.1.

Proposition 2.3.2. *Let $\boldsymbol{c}^0$ be defined by (2.2.1). Then, for all $1 \leq n \leq N$, the nonlinear system of equations (2.2.3)–(2.2.5), supplemented either with (C), (S), (AB), or (BC), has a solution $(\boldsymbol{c}^n, \boldsymbol{\Phi}^n) \in \mathbb{R}^\mathcal{T} \times \mathbb{R}^\mathcal{T}$.*

Proof. The proof is a proof by induction; it relies on a topological degree argument [98, 56] at each time step. The idea is to transform continuously our complex nonlinear system into a linear system while guaranteeing that a priori estimates controlling the solution remain valid all along the homotopy. We sketch the main ideas of the proof, making the homotopy (parametrized by $\lambda \in [0,3]$) explicit.

We denote by $\boldsymbol{c}^\star = \boldsymbol{c}^{n-1} \in (0,1)^\mathcal{T}$ the discrete concentration at the previous time step. We are interested in the existence of zeros for a functional

$$\mathcal{H} : \begin{cases} [0,3] \times (0,1)^\mathcal{T} \times \mathbb{R}^\mathcal{T} \to \mathbb{R}^\mathcal{T} \times \mathbb{R}^\mathcal{T} \\ (\lambda, \boldsymbol{c}, \boldsymbol{\Phi}) \mapsto \mathcal{H}(\lambda, \boldsymbol{c}, \boldsymbol{\Phi}) \end{cases}$$

that boils down to the scheme (2.2.4) when $\lambda = 3$. For sake of simplicity, instead of defining $\mathcal{H}$ for the different values of λ, we give a sense to the fact that $\boldsymbol{c}^{(\lambda)}, \boldsymbol{\Phi}^{(\lambda)}$ is solution to $\mathcal{H}(\lambda, \boldsymbol{c}^{(\lambda)}, \boldsymbol{\Phi}^{(\lambda)}) = 0$.

We start with $\lambda \in [0,1]$: $\boldsymbol{c}^{(\lambda)}$ is defined as the solution to the nonlinear system of equation

$$m_K \frac{c_K^{(\lambda)} - c_K^\star}{\Delta t_n} + (1-\lambda) \sum_{\sigma \in \mathcal{E}_{K,\mathrm{int}}} \tau_\sigma (c_K^{(\lambda)} - c_L^{(\lambda)}) + \lambda \sum_{\sigma \in \mathcal{E}_{K,\mathrm{int}}} \tau_\sigma \left(r(c_K^{(\lambda)}) - r(c_L^{(\lambda)}) \right) = 0, \tag{2.3.18}$$

while $\boldsymbol{\Phi}^{(\lambda)} = 0$. Let us remark that for $\lambda = 0$, it boils down to an invertible linear system of equations. Moreover, adapting the proof of Proposition 2.3.1 and using the property $(r(a) - r(b))(h(a) - h(b)) \geq (a-b)(h(a) - h(b))$ for all $(a,b) \in (0,1)^2$, we get:

$$E_{\mathcal{T}}(\boldsymbol{c}^{(\lambda)}, \mathbf{0}) - E_{\mathcal{T}}(\boldsymbol{c}^\star, \mathbf{0}) \leq -\Delta t_n \sum_{\sigma \in \mathcal{E}_{\mathrm{int}}} \tau_\sigma (c_K^{(\lambda)} - c_L^{(\lambda)})(h(c_K^{(\lambda)}) - h(c_L^{(\lambda)})).$$

As the associated dissipation function defined by $\mathcal{D}(c_K, c_L) = (c_K - c_L)(h(c_K) - h(c_L))$ is clearly coercive in the sense of Lemma 2.3.2, we can deduce as in the proof of Lemma 2.3.6 the existence of $\epsilon_1 > 0$ such that $\epsilon_1 < c_K^{(\lambda)} < 1 - \epsilon_1$ for all $K \in \mathcal{T}$ and all $\lambda \in [0,1]$.

For $\lambda \in [1,2]$, one lets our system evolve from the monotone scheme corresponding to $\lambda = 1$ (which, due to Remark 2.2.1 corresponds to the Sedan scheme for the case without electrical potential) to the scheme with the expected numerical fluxes $F_{K\sigma}$. The electrical potential remains fixed to $\boldsymbol{\Phi}^{(\lambda)} = \mathbf{0}$, i.e.,

$$\begin{gathered} m_K \frac{c_K^{(\lambda)} - c_K^\star}{\Delta t_n} + (2-\lambda) \sum_{\sigma \in \mathcal{E}_{K,\mathrm{int}}} \tau_\sigma \left(r(c_K^{(\lambda)}) - r(c_L^{(\lambda)}) \right) + (\lambda - 1) \sum_{\sigma \in \mathcal{E}_{K,\mathrm{int}}} F_{K\sigma}^{(\lambda)} = 0, \\ F_{K\sigma}^{(\lambda)} = \tau_\sigma \mathcal{F}(c_K^{(\lambda)}, c_L^{(\lambda)}, 0, 0). \end{gathered} \tag{2.3.19}$$

with $\mathcal{F}$ defined either by (C), (S), (AB), or (BC). Thanks to Lemma 2.3.6, there exists $\epsilon_2 > 0$ such that $\epsilon_2 < c_K^{(\lambda)} < 1 - \epsilon_2$ for all $K \in \mathcal{T}$ and all $\lambda \in [1,2]$.

During the last step, $\lambda \in [2,3]$, we reactivate progressively the electrical potential while keeping equation (2.2.4b). Defining

$$\Phi_\sigma^{D,(\lambda)} = (\lambda - 2)\Phi_\sigma^D, \qquad \forall \sigma \in \mathcal{E}^D,$$

the solutions $(\boldsymbol{c}^{(\lambda)}, \boldsymbol{\Phi}^{(\lambda)})$ are defined, for all $\lambda \in [2,3]$ as the solution to the non-

linear system: for all $K \in \mathcal{T}$:

$$m_K \frac{c_K^{(\lambda)} - c_K^\star}{\Delta t_n} + \sum_{\sigma \in \mathcal{E}_{K,\mathrm{int}}} F_{K\sigma}^{(\lambda)} = 0, \text{ with } F_{K\sigma}^{(\lambda)} = \tau_\sigma \mathcal{F}\Big(c_K^{(\lambda)}, c_L^{(\lambda)}, (\lambda-2)\Phi_K^{(\lambda)}, (\lambda-2)\Phi_L^{(\lambda)}\Big),$$
$$- \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \mathbf{\Phi}^{(\lambda)} = (\lambda-2) m_K (c_K^{(\lambda)} + c_K^{\mathrm{dop}}).$$

Thanks to Proposition 2.A.1, one has $\left|\Phi_K^{(\lambda)}\right| \leq M_\Phi$ for all $K \in \mathcal{T}$ and all $\lambda \in [2,3]$. Moreover, as in Lemma 2.3.6, we can establish the existence of $\epsilon_3 > 0$ such that $\epsilon_3 < c_K^{(\lambda)} < 1 - \epsilon_3$ for all $K \in \mathcal{T}$ and all $\lambda \in [2,3]$.

Finally, all along the homotopy parametrized by $\lambda \in [0,3]$, the solutions $(\boldsymbol{c}^{(\lambda)}, \mathbf{\Phi}^{(\lambda)})$ remains inside the compact subset $[\epsilon, 1-\epsilon]^{\mathcal{T}} \times [-M_\Phi - 1, M_\phi + 1]^{\mathcal{T}}$ with $\epsilon = \min(\epsilon_1, \epsilon_2, \epsilon_3)$. Thus, the topological degree corresponding to $\mathcal{H}(\lambda, \boldsymbol{c}, \mathbf{\Phi}) = \mathbf{0}$ and to the set $[\epsilon, 1-\epsilon]^{\mathcal{T}} \times [-M_\Phi - 1, M_\phi + 1]^{\mathcal{T}}$ is equal to one all along the homotopy and in particular for $\lambda = 3$. This ensures the existence of (at least) one solution to the scheme (2.2.4). □

2.4 About the convergence towards a weak solution

The goal of this section is to prove Theorem 2.2.2, which states the convergence of the centred scheme (2.2.4), (C), and the Sedan scheme (2.2.4), (S), towards a weak solution to the continuous problem in the sense of Definition 3. Unfortunately, the proof we propose here neither applies to the activity based scheme (2.2.4), (AB), nor the Bessemoulin-Chatard scheme (2.2.4), (BC). This does not mean that these schemes do not converge. Indeed, numerical evidences provided in Section 2.5 seem to show that all the four schemes converge.

We consider here a sequence $\left(\mathcal{T}_m, \mathcal{E}_m, (\boldsymbol{x}_K)_{K \in \mathcal{T}_m}\right)_{m \geq 1}$ of admissible discretization with $h_{\mathcal{T}_m}, \overline{\mathbf{\Delta t}}_m$ tending to 0 as m tends to $+\infty$, while the regularity $\zeta_{\mathcal{T}_m}$ remains uniformly bounded from below by a positive constant $\zeta^\star$. Theorem 2.2.1 provides the existence of a family of discrete solutions

$$(\boldsymbol{c}_m, \mathbf{\Phi}_m)_m = \left((c_K^n)_{K \in \mathcal{T}_m, 1 \leq n \leq N_m}, (\Phi_K^n)_{K \in \mathcal{T}_m, 1 \leq n \leq N_m}\right).$$

To prove Theorem 2.2.2, we first establish in Section 2.4.2 compactness properties on the family of piecewise constant approximate solutions $(c_{\mathcal{T}_m, \mathbf{\Delta t}_m}, \Phi_{\mathcal{T}_m, \mathbf{\Delta t}_m})$ satisfied by the centred scheme and the Sedan scheme. Then we identify the limit as a weak solution in Section 2.4.3.

To enlighten the notations, we remove the subscript m as soon as it is not necessary for understanding.

2.4.1 Reconstruction operators

In order to carry out the analysis of convergence, we introduce some reconstruction operators following the methodology proposed in [64].

The operators $\pi_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}} \to L^\infty(\Omega)$ and $\pi_{\mathcal{T},\boldsymbol{\Delta t}} : \mathbb{R}^{\mathcal{T}\times N} \to L^\infty((0,T)\times\Omega)$ are defined respectively by

$$\pi_{\mathcal{T}}\boldsymbol{u}(\boldsymbol{x}) = u_K \quad \text{if } \boldsymbol{x} \in K, \qquad \forall \boldsymbol{u} = (u_K)_{K\in\mathcal{T}},$$

and

$$\pi_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}(t,\boldsymbol{x}) = u_K^n \quad \text{if } (t,\boldsymbol{x}) \in (t_{n-1},t_n]\times K, \qquad \forall \boldsymbol{u} = (u_K^n)_{K\in\mathcal{T},1\le n\le N}.$$

These operators allow passing from the discrete solution $(\boldsymbol{c}^n,\boldsymbol{\Phi}^n)_{1\le n\le N}$ to the approximate solution since

$$c_{\mathcal{T},\boldsymbol{\Delta t}} = \pi_{\mathcal{T},\boldsymbol{\Delta t}}\left(\boldsymbol{c}^n\right)_n, \qquad \Phi_{\mathcal{T},\boldsymbol{\Delta t}} = \pi_{\mathcal{T},\boldsymbol{\Delta t}}\left(\boldsymbol{\Phi}^n\right)_n.$$

To carry out the analysis, we further need to introduce approximate gradient reconstruction. Since the boundary conditions play a crucial role in the definition of the gradient, we need to enrich the discrete solution by face values $(c_\sigma^n)_{\sigma\in\mathcal{E}_{\rm ext},1\le n\le N}$ and $(\Phi_\sigma^n)_{\sigma\in\mathcal{E}_{\rm ext},1\le n\le N}$ defined by $c_\sigma^n = c_{K\sigma}^n$ and $\Phi_\sigma^n = \Phi_{K\sigma}^n$. With a slight abuse of notations, we still denote by $\boldsymbol{c}^n = ((c_K^n)_{K\in\mathcal{T}},(c_\sigma^n)_{\sigma\in\mathcal{E}_{\rm ext}})$ and $\boldsymbol{\Phi}^n = ((\Phi_K^n)_{K\in\mathcal{T}},(\Phi_\sigma^n)_{\sigma\in\mathcal{E}_{\rm ext}})$ the elements of $(0,1)^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}$ and $\mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}$ containing both the cell values and the exterior faces values of the concentration and the potential respectively.

For $\sigma = K|L \in \mathcal{E}_{\rm int}$, we denote by Δ_σ the diamond cell corresponding to σ, that is the interior of the convex hull of $\sigma\cup\{\boldsymbol{x}_K,\boldsymbol{x}_L\}$. For $\sigma\in\mathcal{E}_{\rm ext}$, the diamond cell Δ_σ is defined as the interior of the convex hull of $\sigma\cup\{\boldsymbol{x}_K\}$. The approximate gradient $\nabla_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}} \to L^2(\Omega)^d$ we use in the analysis is merely weakly consistent (unless $d=1$) and takes its source in [48, 66]. It is piecewise constant on the diamond cells Δ_σ, and it is defined as follows:

$$\nabla_{\mathcal{T}}\boldsymbol{u}(\boldsymbol{x}) = -d\frac{D_{K\sigma}\boldsymbol{u}}{d_\sigma}\boldsymbol{n}_{K\sigma} \quad \text{if } \boldsymbol{x}\in\Delta_\sigma, \qquad \forall \boldsymbol{u}\in\mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}.$$

We also define $\nabla_{\mathcal{T},\boldsymbol{\Delta t}} : \mathbb{R}^{(\mathcal{T}\cup\mathcal{E}_{\rm ext})\times N} \to L^2(Q_T)^d$ by setting

$$\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}(t,\cdot) = \nabla_{\mathcal{T}}\boldsymbol{u}^n \quad \text{if } t\in(t_{n-1},t_n], \qquad \forall \boldsymbol{u} = (\boldsymbol{u}^n)_{1\le n\le N} \in \mathbb{R}^{(\mathcal{T}\cup\mathcal{E}_{\rm ext})\times N}.$$

Let us recall now some key properties to be used in the analysis. First, for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}$,

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\boldsymbol{u} D_{K\sigma}\boldsymbol{v} = \frac{1}{d}\int_\Omega \nabla_{\mathcal{T}}\boldsymbol{u}\cdot\nabla_{\mathcal{T}}\boldsymbol{v}\mathrm{d}\boldsymbol{x}.$$

This implies in particular that

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma |D_\sigma \boldsymbol{u}|^2 = \frac{1}{d}\int_\Omega |\nabla_{\mathcal{T}}\boldsymbol{u}|^2\mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{u}\in\mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}. \tag{2.4.1}$$

2.4.2 Compactness properties for the approximate concentration

The goal here is to take advantage of the a priori estimates established in Section 2.3.2 to recover enough compactness for the sequences of approximate solutions.

Lemma 2.4.1. *Let $(\boldsymbol{c}_m, \boldsymbol{\Phi}_m)$ be the family of discrete solutions defined either by the centred scheme or by the Sedan scheme. There exists C depending only on Φ^D, Ω, $\zeta^\star$, c_0, $c^{\rm dop}$ and T, such that*

$$\iint_{Q_T} |\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} r(\boldsymbol{c}_m)|^2 + \left(\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} r(\boldsymbol{c}_m)\right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \le C.$$

Proof. We get rid of the subscript m for the ease of reading. We will split the proof in two parts, first we focus on the proof of:

$$\iint_{Q_T} |\nabla_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c})|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \le C. \tag{2.4.2}$$

Thanks to (2.4.1), we have

$$\begin{aligned}\iint_{Q_T} |\nabla_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c})|^2 &= d\sum_{n=1}^N \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma |D_\sigma(r(\boldsymbol{c}^n))|^2,\\ &= d\sum_{n=1}^N \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \left(\widetilde{\mathcal{C}}_\sigma^n\right)^2 |D_\sigma(h(\boldsymbol{c}^n))|^2,\end{aligned}$$

where we have defined the mean face concentrations $\left(\widetilde{\mathcal{C}}_\sigma^n\right)_{\sigma\in\mathcal{E}_{\rm int},1\le n\le N}$ by setting

$$\widetilde{\mathcal{C}}_\sigma^n = \frac{D_\sigma r(\boldsymbol{c}^n)}{D_\sigma h(\boldsymbol{c}^n)} \text{ if } c_K^n \neq c_L^n \quad \text{and} \quad \widetilde{\mathcal{C}}_\sigma^n = c_K^n \text{ otherwise}, \qquad \forall \sigma = K|L. \tag{2.4.3}$$

As noticed by (2.C.1), $\widetilde{\mathcal{C}}_\sigma^n$ is a mean value of c_K^n and c_L^n; so that $\widetilde{\mathcal{C}}_\sigma^n \in (0,1)$ for all $\sigma \in \mathcal{E}_{\rm int}$. Moreover, Lemma 2.C.1 proved in Appendix 2.C ensures that there exists $G > 0$ such that

$$\frac{\widetilde{\mathcal{C}}_\sigma^n}{\mathcal{C}_\sigma^n} \leq G, \qquad \forall \sigma \in \mathcal{E}_{\rm int},\ \forall n \in \{1,\dots,N\}. \tag{2.4.4}$$

Note that the above estimate may not hold for the Bessemoulin-Chatard scheme, as shown in Remark 2.C.1. Then, thanks to the classical $(a+b)^2 \leq 2a^2 + 2b^2$ inequality, we obtain

$$\begin{aligned}\iint_{Q_T} |\nabla_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c})|^2 \leq & 2dG \sum_{n=1}^N \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \mathcal{C}_\sigma^n |D_\sigma(h(\boldsymbol{c}^n) + \boldsymbol{\Phi}^n)|^2 \\ & + 2d \sum_{n=1}^N \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma |D_\sigma \boldsymbol{\Phi}^n|^2.\end{aligned}$$

Therefore, Lemma 2.3.5 yield the desired bound (2.4.2).

We now focus on the proof of:

$$\iint_{Q_T} \left(\pi_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c})\right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq C. \tag{2.4.5}$$

Noticing that for $c^* = \frac{1+\bar{c}}{2} > \bar{c}$:

$$r(c) \leq (r(c) - r(c^*))^+ + r(c^*),$$

we have, using $(a+b)^2 \leq 2(a^2+b^2)$:

$$\iint_{Q_T} |\pi_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c})|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq 2 \iint_{Q_T} |(\pi_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c}) - r(c^*))^+|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + 2r(c^*)^2 m(\Omega) T. \tag{2.4.6}$$

Let $t \in [0,T]$ and $u = (\pi_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c}) - r(c^*))^+(t)$. We intend to show that we have a L^2 bound on u following ideas of [3, Appendix A.1]. As u is nonnegative, we have:

$$\int_\Omega |u - \bar{u}|^2 = \int_{u=0} \bar{u}^2 + \int_{\Omega\setminus\{u=0\}} |u-\bar{u}|^2 \geq m(\{u=0\})\bar{u}^2, \tag{2.4.7}$$

where $\bar{u} = \fint_\Omega u$. Using Poincaré-Wirtinger inequality (see [16, Theorem 5] or [86,

Theorem 2.1]), we have:

$$\int_\Omega |u-\bar{u}|^2 \leq \frac{C}{\zeta_{\mathcal{T}}} \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma (D_\sigma u)^2. \tag{2.4.8}$$

If we had a lower bound on $m(\{u=0\})$, the equations (2.4.8) and (2.4.7) would yield an upper bound on $\bar{u}$. By definition of u and monotonicity of r, u is zero if and only if c is smaller than c^*. Using the monotonicity of integration and Lemma 2.3.3, we have:

$$c^*(m(\Omega)-m(\{u=0\})) = \int_{c>c^*} c^* \leq \int_\Omega \pi_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{c}(t) = m(\Omega)\bar{c}.$$

Hence, as $c^* = (1+\bar{c})/2$,

$$m(\Omega)\frac{1-\bar{c}}{1+\bar{c}} \leq m(\{u=0\}).$$

Finally, we have:

$$\int_\Omega u^2 \leq 2\left(\int_\Omega |u-\bar{u}|^2 + \bar{u}^2\right) \leq C \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma (D_\sigma u)^2.$$

Using the definition of u, we have:

$$\sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma (D_\sigma u)^2 \leq \int_\Omega |\nabla_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c})(t)|^2.$$

Hence, integrating in time, and using (2.4.6):

$$\iint_{Q_T} |\pi_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c})|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq C \iint_{Q_T} |\nabla_{\mathcal{T},\boldsymbol{\Delta t}} r(\boldsymbol{c})(t)|^2 + C.$$

We then deduce (2.4.5) from (2.4.2). This concludes the proof of Lemma 2.4.1. □

Proposition 2.4.1. *Let $(\boldsymbol{c}_m, \boldsymbol{\Phi}_m)$ be the family of discrete solutions defined either by the centred scheme or by the Sedan scheme. In both cases, there exists $c \in L^\infty(Q_T;[0,1])$ with $r(c) \in L^2((0,T);H^1(\Omega))$ such that, up to a subsequence,*

$$\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \boldsymbol{c}_m \underset{m\to\infty}{\longrightarrow} c \quad \text{a.e. in } Q_T, \tag{2.4.9}$$

$$\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} r(\boldsymbol{c}_m) \underset{m\to\infty}{\longrightarrow} \nabla r(c) \quad \text{weakly in } L^2(Q_T). \tag{2.4.10}$$

Remark 2.4.1. The limit c obtained in Proposition 2.4.1 could *a priori* depend on the chosen subsequence or be different for the centred scheme and the Sedan scheme. In Section 2.4.3, we will identify each limit as a weak solution to the initial problem.

Proof. Since $0 < \pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{c}_m < 1$ for all $m \geq 1$, there exists c in $L^\infty(Q_T;[0,1])$ such that $\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{c}_m$ tends to c in the $L^\infty(Q_T)$ weak-$\star$ sense. We still have to establish the almost everywhere convergence as well as the fact that $r(c)$ belongs to $L^2((0,T);H^1(\Omega))$. To this end, we make use of the black box [9, Theorem 3.9] which provides both the almost everywhere convergence and the identification of the limit of $\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} r(\boldsymbol{c}_m)$ as $r(c)$. We already have Lemma 2.4.1 at hand and c is bounded in L^∞, so that, owing to [9], it is sufficient to prove that there exists some C not depending on m such that, for all $\boldsymbol{\varphi}_m = (\varphi_K,\varphi_\sigma)_{K,\sigma} \in \mathbb{R}^{(\mathcal{T}_m\cup\mathcal{E}_{\mathrm{ext},m})\times N_m}$,

$$\left|\sum_n \Delta t_n \sum_{K\in\mathcal{T}_m} m_K \frac{c_K^n - c_K^{n-1}}{\Delta t_n}\varphi_K^n\right| \leq C\|\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\varphi}_m\|_{L^\infty(Q_T)}.$$

We would then have, among other things, the desired convergence (2.4.9). Using (2.2.4b) and the writing (2.3.7) of the fluxes, we obtain that

$$\begin{aligned}\left|\sum_n \Delta t_n \sum_{K\in\mathcal{T}_m} m_K \frac{c_K^n - c_K^{n-1}}{\Delta t_n}\varphi_K^n\right| &= \left|\sum_n \Delta t_n \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \tau_\sigma \mathcal{C}_\sigma^n D_{K\sigma}(h(\boldsymbol{c}^n)+\boldsymbol{\Phi}^n) D_{K\sigma}\boldsymbol{\varphi}\right| \\ &\leq \left(\sum_n \Delta t_n \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \tau_\sigma \mathcal{C}_\sigma^n \left|D_\sigma(h(\boldsymbol{c}^n)+\boldsymbol{\Phi}^n)\right|^2\right)^{1/2} \\ &\quad \times \left(\sum_n \Delta t_n \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \tau_\sigma |D_\sigma\boldsymbol{\varphi}^n|^2\right)^{1/2} \\ &\leq C\|\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\varphi}_m\|_{L^2(Q_T)} \leq C\|\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\varphi}_m\|_{L^\infty(Q_T)},\end{aligned}$$

thanks to the boundedness of the dissipation in Lemma 2.3.5 which is a consequence of Proposition 2.3.1.

Since $\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} r(\boldsymbol{c}_m)$ is bounded in $L^2(Q_T)^d$, it converges weakly in $L^2(Q_T)^d$ towards some $\boldsymbol{U}$. The identification of $\boldsymbol{U}$ as $\nabla r(c)$ is classical (see for instance [48, Sec. 4], [66] or [62, Lemma 6.5]). □

We have two kinds of face values at hand: $\left(\mathcal{C}_\sigma^n\right)_{\sigma\in\mathcal{E}_{\mathrm{int}},1\leq n\leq N}$ and $\left(\widetilde{\mathcal{C}}_\sigma^n\right)_{\sigma\in\mathcal{E}_{\mathrm{int}},1\leq n\leq N}$ defined respectively by (2.3.6) and (2.4.3). Based on this, we can reconstruct two approximate concentration profiles $c_{\mathcal{E},\boldsymbol{\Delta t}}$ and $\widetilde{c}_{\mathcal{E},\boldsymbol{\Delta t}}$ that are piecewise constant on

the diamond cells by setting

$$c_{\mathcal{E},\boldsymbol{\Delta t}}(t,\boldsymbol{x}) = \begin{cases} \mathcal{C}_\sigma^n & \text{if } (t,\boldsymbol{x}) \in (t_{n-1},t_n] \times \Delta_\sigma, \quad \sigma \in \mathcal{E}_{\rm int}, \\ c_K^n & \text{if } \boldsymbol{x} \in \Delta_\sigma, \quad \sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K, \end{cases} \tag{2.4.11}$$

and

$$\widetilde{c}_{\mathcal{E},\boldsymbol{\Delta t}}(t,\boldsymbol{x}) = \begin{cases} \widetilde{\mathcal{C}}_\sigma^n & \text{if } (t,\boldsymbol{x}) \in (t_{n-1},t_n] \times \Delta_\sigma, \quad \sigma \in \mathcal{E}_{\rm int}, \\ c_K^n & \text{if } \boldsymbol{x} \in \Delta_\sigma, \quad \sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K. \end{cases} \tag{2.4.12}$$

Lemma 2.4.2. *For the centred scheme and the Sedan scheme, there holds*

$$c_{\mathcal{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} c \quad \text{in } L^p(Q_T) \text{ for all } p \in [1,\infty), \tag{2.4.13}$$

$$\widetilde{c}_{\mathcal{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} c \quad \text{in } L^p(Q_T) \text{ for all } p \in [1,\infty), \tag{2.4.14}$$

where c is as in Proposition 2.4.1.

Proof. We only prove (2.4.13) since the proof of (2.4.14) is similar. Here again, we get rid of m for clarity. Since $c_{\mathcal{T},\boldsymbol{\Delta t}}$ converges almost everywhere to c and remains bounded between 0 and 1, it converges in $L^p(Q_T)$. $c_{\mathcal{E},\boldsymbol{\Delta t}}$ is also uniformly bounded, hence it suffices to show that $\|c_{\mathcal{E},\boldsymbol{\Delta t}} - c_{\mathcal{T},\boldsymbol{\Delta t}}\|_{L^1(Q_T)}$ tends to 0. Denoting by $\Delta_{K\sigma}$ the half-diamond cell which is defined as the interior of the convex hull of $\sigma \cup \{\boldsymbol{x}_K\}$ for $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, one has

$$\begin{aligned} \|c_{\mathcal{E},\boldsymbol{\Delta t}} - c_{\mathcal{T},\boldsymbol{\Delta t}}\|_{L^1(Q_T)} &\le \sum_{n=1}^N \Delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_{\Delta_{K\sigma}} |c_K^n - \mathcal{C}_\sigma^n| \\ &\le \frac{h_\mathcal{T}}{d} \sum_{n=1}^N \Delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_\sigma |c_K^n - \mathcal{C}_\sigma^n|, \end{aligned}$$

where we have used the geometric relation $m(\Delta_{K\sigma}) = \frac{1}{d} m_\sigma \mathrm{dist}(\boldsymbol{x}_K,\sigma) \le \frac{h_\mathcal{T}}{d} m_\sigma$. For the internal faces, Lemma 2.3.1 (use (2.C.1) instead for $\widetilde{\mathcal{C}}_\sigma^n$) implies that

$$|c_K^n - \mathcal{C}_\sigma^n| + |c_L^n - \mathcal{C}_\sigma^n| = |c_K^n - c_L^n|, \qquad \forall \sigma = K|L.$$

Therefore, we obtain that

$$\begin{aligned}\|c_{\mathcal{E},\boldsymbol{\Delta t}} - c_{\mathcal{T},\boldsymbol{\Delta t}}\|_{L^1(Q_T)} \leq & \frac{h_{\mathcal{T}}}{d} \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} m_\sigma D_\sigma \boldsymbol{c}^n \\ \leq & \frac{h_{\mathcal{T}}}{d} \left(\sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} m_\sigma d_\sigma\right)^{1/2} \left(\sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma |D_\sigma \boldsymbol{c}^n|^2\right)^{1/2}.\end{aligned}$$

Since $|r(a) - r(b)| > |a - b|$ for all $a, b \in (0,1)$, we deduce from Lemma 2.4.1 that

$$\|c_{\mathcal{E},\boldsymbol{\Delta t}} - c_{\mathcal{T},\boldsymbol{\Delta t}}\|_{L^1(Q_T)} \leq C h_{\mathcal{T}}.$$

□

2.4.3 Convergence towards a weak solution

Proposition 2.4.2. *Let c be as in Proposition 2.4.1 and let $\Phi \in L^\infty(Q_T) \cap L^\infty((0,T); H^1(\Omega))$ be the solution to the Poisson equation (2.1.3) with boundary conditions (2.1.6). Then, for the centred scheme and the Sedan scheme, there holds*

$$\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \boldsymbol{\Phi}_m \underset{m\to\infty}{\longrightarrow} \Phi \text{ in } L^2(Q_T) \text{ and in the } L^\infty(Q_T) \text{ weak-}\star \text{ sense}, \tag{2.4.15}$$

and

$$\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \boldsymbol{\Phi}_m \underset{m\to\infty}{\longrightarrow} \nabla\Phi \quad \text{in the } L^\infty((0,T); L^2(\Omega)^d) \text{ weak-}\star \text{ sense}. \tag{2.4.16}$$

Proof. The existence of some $\Phi \in L^\infty(Q_T)$ such that (2.4.15) holds is a straightforward consequence of Lemma 2.3.4, whereas the existence of some vector field $\boldsymbol{U}$ in $L^\infty((0,T); L^2(\Omega)^d)$ such that $\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \boldsymbol{\Phi}$ tends to $\boldsymbol{U}$ as m tends to ∞ follows from Lemma 2.3.5 together with (2.4.1). For the proof of the identification $\boldsymbol{U} = \nabla\Phi$, we refer to [48, 66, 62].

We show now that Φ satisfies the Poisson equation (2.1.3). Let $\psi \in C_c^\infty([0,T] \times \{\Omega \cup \Gamma^N\})$, then define $\psi_K^n = \psi(\boldsymbol{x}_K, t_n)$ and $\psi_\sigma^n = \psi(\boldsymbol{x}_\sigma, t_n)$ for $1 \leq n \leq N$, $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_{\rm ext}$. Following [63] (see [52] for a practical example), one can reconstruct a second approximate gradient operator $\widehat{\nabla}_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}} \to L^\infty(\Omega)^d$ such that

$$\int_\Omega \nabla_{\mathcal{T}} \boldsymbol{u} \cdot \widehat{\nabla}_{\mathcal{T}} \boldsymbol{v} \mathrm{d}\boldsymbol{x} = \sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma} \boldsymbol{u} D_{K\sigma} \boldsymbol{v}, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{T}},$$

and which is strongly consistent, i.e.,

$$\widehat{\nabla}_{\mathcal{T}}\boldsymbol{\psi}^n \underset{h_{\mathcal{T}}\to 0}{\longrightarrow} \nabla\psi(\cdot,t_n) \text{ uniformly in } \overline{\Omega}, \quad \forall n\in\{1,\dots,N\}, \tag{2.4.17}$$

thanks to the smoothness of ψ. The scheme (2.2.4a) then reduces to

$$\int_\Omega \nabla_{\mathcal{T}}\boldsymbol{\Phi}^n\cdot\widehat{\nabla}_{\mathcal{T}}\boldsymbol{\psi}^n\mathrm{d}\boldsymbol{x} = \int_\Omega \pi_{\mathcal{T}}(\boldsymbol{c}^n+c^{\mathrm{dop}})\pi_{\mathcal{T}}\boldsymbol{\psi}^n\mathrm{d}\boldsymbol{x}, \quad \forall n\in[\![1,N]\!],\ \forall\boldsymbol{\psi}\in\mathbb{R}^{(\mathcal{T}\cup\mathcal{E}_{\mathrm{ext}})\times N}.$$

Integrating with respect to time over $(0,T)$ and passing to the limit $h_{\mathcal{T}},\overline{\boldsymbol{\Delta t}}\to 0$ thanks to Proposition 2.4.1, (2.4.16) and (2.4.17) then yields

$$\iint_{Q_T}\nabla\Phi\cdot\nabla\psi\mathrm{d}\boldsymbol{x}\mathrm{d}t = \iint_{Q_T}(c+c^{\mathrm{dop}})\psi\mathrm{d}\boldsymbol{x}\mathrm{d}t, \qquad \forall\psi\in C_c^\infty([0,T]\times\Omega\cup\Gamma^N).$$

In particular, (2.1.15) holds for almost every $t\in(0,T)$. Concerning the boundary conditions for Φ, the fact that $\Phi=\Phi^D$ on $(0,T)\times\Gamma^D$. can be proved for instance following the lines of [23, Section 4].

It remains to check that $\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\Phi}_m$ strongly converges towards Φ in $L^2(Q_T)$. To this end, we make use of a discrete Aubin-Simon lemma [78] in the particular setting of [31, Lemma 9]. Since we have a discrete $L^\infty(H^1)$ estimate at hand thanks to Lemma 2.3.5, it suffices to show that there exists C not depending on m such that, for all $n\geq 1$ and all $\boldsymbol{\varphi}\in\mathbb{R}^{\mathcal{T}_m}$, we have

$$\sum_{K\in\mathcal{T}_m} m_K(\Phi_K^n-\Phi_K^{n-1})\varphi_K = \int_\Omega \pi_{\mathcal{T}_m}\left(\boldsymbol{\Phi}^n-\boldsymbol{\Phi}^{n-1}\right)\pi_{\mathcal{T}_m}\boldsymbol{\varphi} \leq \Delta t_n C\|\pi_{\mathcal{T}_m}\boldsymbol{\varphi}\|_{L^2}. \tag{2.4.18}$$

By linearity of (2.2.4a) we have:

$$-\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma D_{K\sigma}\left(\boldsymbol{\Phi}^n-\boldsymbol{\Phi}^{n-1}\right) = m_K(c_K^n-c_K^{n-1}), \qquad \forall K\in\mathcal{T}_m.$$

Using (2.2.4b) and (2.3.7) there holds

$$\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma D_{K\sigma}\left(\boldsymbol{\Phi}^n-\boldsymbol{\Phi}^{n-1}\right) = -\Delta t_n\sum_{\sigma\in\mathcal{E}_{K,\mathrm{int}}}\tau_\sigma\mathcal{C}_\sigma^n D_{K\sigma}(h(\boldsymbol{c}^n)+\boldsymbol{\Phi}^n), \qquad \forall K\in\mathcal{T}_m. \tag{2.4.19}$$

Let $\boldsymbol{\varphi}\in\mathbb{R}^{\mathcal{T}_m}$, then define $\boldsymbol{\psi}\in\mathbb{R}^{\mathcal{T}_m}$ as the solution of the linear system

$$-\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma D_{K\sigma}\boldsymbol{\psi} = m_K\varphi_K, \qquad \forall K\in\mathcal{T}_m, \tag{2.4.20}$$

where we have set $\psi_{K\sigma} = \psi_L$ if $\sigma = K|L \in \mathcal{E}_{\rm int}$, and $\psi_{K\sigma} = 0$ if $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\rm ext}$. Multiplying (2.4.19) by ψ_K, summing over $K \in \mathcal{T}_m$, performing discrete integration by parts and using (2.4.20) yields

$$\sum_{K\in\mathcal{T}_m} m_K(\Phi_K^n - \Phi_K^{n-1})\varphi_K = \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \mathcal{C}_\sigma^n D_{K\sigma}(h(\boldsymbol{c}^n) + \boldsymbol{\Phi}^n) D_{K\sigma}\boldsymbol{\psi}.$$

Using successively Cauchy Schwarz inequality and [68, Lemma 9.2] we get

$$\begin{aligned}\sum_{K\in\mathcal{T}_m} m_K(\Phi_K^n - \Phi_K^{n-1})\varphi_K \leq& \Delta t_n \Big(\sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \mathcal{C}_\sigma^n \big(D_\sigma(h(\boldsymbol{c}^n) + \boldsymbol{\Phi}^n)\big)^2 \Big)^{\frac12} \Big(\sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \big(D_{K\sigma}\boldsymbol{\psi}\big)^2 \Big)^{\frac12} \\ \leq& C\Delta t_n \Big(\sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \mathcal{C}_\sigma^n \big(D_\sigma(h(\boldsymbol{c}^n) + \boldsymbol{\Phi}^n)\big)^2 \Big)^{\frac12} \|\pi_{\mathcal{T}_m}\boldsymbol{\varphi}\|_{L^2(\Omega)}.\end{aligned}$$

Then the control on the dissipation established in Proposition 2.3.1 allows to recover (2.4.18), hence the relative compactness in $L^2(Q_T)$ of $(\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\Phi}_m)_m$. □

Remark 2.4.2 (Enhanced convergence properties). The convergence described in Proposition 2.4.2 is suboptimal. One can establish the strong convergence of $\widehat{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\Phi}_m$ towards $\nabla\Phi$, where the gradient reconstruction operator $\widehat{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}$ is the extension to the time-space domain Q_T of the operator $\widehat{\nabla}_{\mathcal{T}_m}$ used in the proof of Proposition 2.4.2. We refer to [63] for details on these enhanced convergence properties.

Proposition 2.4.3. *Let c, Φ be as in Propositions 2.4.1 and 2.4.2, then the weak formulation (2.1.14) holds.*

Proof. Let $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$, then define $\varphi_K^n = \varphi(\boldsymbol{x}_K, t_n)$ for all $n \in \{0,\dots,N\}$ and $K \in \mathcal{T}$. Multiplying (2.2.4b) by $\Delta t_n \varphi_K^{n-1}$, then summing over $K \in \mathcal{T}$ and $n \in \{1,\dots,N\}$ and using expression (2.3.7) for the fluxes leads to

$$T_1 + T_2 + T_3 = 0, \tag{2.4.21}$$

where we have set

$$T_1 = \sum_{n=1}^{N} \sum_{K\in\mathcal{T}} m_K (c_K^n - c_K^{n-1}) \varphi_K^{n-1},$$
$$T_2 = \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \mathcal{C}_\sigma^n D_{K\sigma} h(\boldsymbol{c}^n) D_{K\sigma} \boldsymbol{\varphi}^{n-1},$$
$$T_3 = \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \mathcal{C}_\sigma^n D_{K\sigma} \boldsymbol{\Phi}^n D_{K\sigma} \boldsymbol{\varphi}^{n-1}.$$

The term T_1 can be rewritten as

$$T_1 = \sum_{n=1}^{N} \Delta t_n \sum_{K\in\mathcal{T}} m_K c_K^n \frac{\varphi_K^{n-1} - \varphi_K^n}{\Delta t_n} - \sum_{K\in\mathcal{T}} m_K c_K^0 \varphi_K^0,$$

so that it follows from the convergence of $\pi_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{c}$ towards c and of $\pi_{\mathcal{T}} \boldsymbol{c}^0$ towards c^0 together with the regularity of φ that

$$T_1 \underset{m\to\infty}{\longrightarrow} -\iint_{Q_T} c \partial_t \varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega c^0 \varphi(0,\cdot)\mathrm{d}\boldsymbol{x}. \tag{2.4.22}$$

On the other hand, the term T_3 can be rewritten as

$$T_3 = \iint_{Q_T} c_{\mathcal{E},\boldsymbol{\Delta t}} \nabla_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{\Phi} \cdot \widehat{\nabla}_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{\varphi} \mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

where $\widehat{\nabla}_{\mathcal{T},\boldsymbol{\Delta t}}$ is the strongly consistent gradient reconstruction operator introduced in the proof of Proposition 2.4.2 and in Remark 2.4.2. In particular, due to the smoothness of φ, $\widehat{\nabla}_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{\varphi}$ converges uniformly towards $\nabla\varphi$. Therefore, it follows from Lemma 2.4.2 and Proposition 2.4.2 that

$$T_3 \underset{m\to\infty}{\longrightarrow} \iint_{Q_T} c \nabla\Phi \cdot \nabla\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{2.4.23}$$

Define the term

$$\widetilde{T}_2 = \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \widetilde{\mathcal{C}}_\sigma^n D_{K\sigma} h(\boldsymbol{c}^n) D_{K\sigma} \boldsymbol{\varphi}^{n-1} = \iint_{Q_T} \nabla_{\mathcal{T},\Delta t} r(\boldsymbol{c}) \cdot \widehat{\nabla}_{\mathcal{T},\Delta t} \boldsymbol{\varphi} \mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

then it follows from Proposition 2.4.1 that

$$\widetilde{T}_2 \underset{m\to\infty}{\longrightarrow} \iint_{Q_T} \nabla r(c)\cdot\nabla\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Therefore, it only remains to show that $|T_2-\widetilde{T}_2|$ tends to 0 to conclude the proof of Proposition 2.4.3. Thanks to the triangle and Cauchy-Schwarz inequalities, one has

$$\begin{aligned}|T_2-\widetilde{T}_2| &\le \sum_{n=1}^{N}\Delta t_n\sum_{\sigma\in\mathcal{E}}\tau_\sigma\left|\mathcal{C}_\sigma^n-\widetilde{\mathcal{C}}_\sigma^n\right| D_\sigma h(\boldsymbol{c}^n)D_\sigma\boldsymbol{\varphi}^{n-1}\\ &\le \left(\sum_{n=1}^{N}\Delta t_n\sum_{\sigma\in\mathcal{E}}\tau_\sigma\mathcal{C}_\sigma^n|D_\sigma h(\boldsymbol{c}^n)|^2\right)^{1/2}\\ &\qquad\times\left(\sum_{n=1}^{N}\Delta t_n\sum_{\sigma\in\mathcal{E}}\tau_\sigma\frac{(\mathcal{C}_\sigma^n-\widetilde{\mathcal{C}}_\sigma^n)^2}{\mathcal{C}_\sigma^n}|D_\sigma\boldsymbol{\varphi}^{n-1}|^2\right)^{1/2}.\end{aligned}$$

The first term in the right-hand side is uniformly bounded thanks to Lemma 2.3.5, using the ideas of the proof of (2.4.2)). Thus our problem amounts to show that

$$\mathcal{R}:=\sum_{n=1}^{N}\Delta t_n\sum_{\sigma\in\mathcal{E}}\tau_\sigma\frac{(\mathcal{C}_\sigma^n-\widetilde{\mathcal{C}}_\sigma^n)^2}{\mathcal{C}_\sigma^n}|D_\sigma\boldsymbol{\varphi}^{n-1}|^2\underset{m\to\infty}{\longrightarrow}0. \tag{2.4.24}$$

Let us reformulate $\mathcal{R}$ as

$$\mathcal{R}:=\sum_{n=1}^{N}\Delta t_n\sum_{\sigma\in\mathcal{E}}\tau_\sigma|\mathcal{C}_\sigma^n-\widetilde{\mathcal{C}}_\sigma^n|\left|1-\frac{\widetilde{\mathcal{C}}_\sigma^n}{\mathcal{C}_\sigma^n}\right||D_\sigma\boldsymbol{\varphi}^{n-1}|^2.$$

Thanks to (2.4.4), the quantity $\left|1-\frac{\widetilde{\mathcal{C}}_\sigma^n}{\mathcal{C}_\sigma^n}\right|$ is uniformly bounded, whereas the regularity of φ implies that $D_\sigma\boldsymbol{\varphi}^{n-1}\le\|\nabla\varphi\|_\infty d_\sigma$. Putting this in the above expression of $\mathcal{R}$, we obtain that

$$\mathcal{R}\le C\|c_{\mathcal{E},\boldsymbol{\Delta t}}-\widetilde{c}_{\mathcal{E},\boldsymbol{\Delta t}}\|_{L^1(Q_T)}\underset{m\to\infty}{\longrightarrow}0,$$

thanks to Lemma 2.4.2. □

Remark 2.4.3. This convergence proof does not hold for the Bessemoulin-Chatard scheme because we lack compactness properties: Lemma 2.C.1, yielding equation (2.4.4), is not satisfied for this scheme (see Remark 2.C.1), which affects successively the proofs of Lemma 2.4.1, Proposition 2.4.1 and therefore Theorem 2.2.2.

The activity based scheme does not satisfy the bounds (2.3.3) of Lemma 2.3.1. This implies gaps in the proof of Lemma 2.4.2, and the convergence of the scheme stated in Theorem 2.2.2 is not established.

2.5 Numerical comparison of the schemes

The numerical examples [72] have been implemented in the Julia language [17] based on the package `VoronoiFVM.jl` [73] which realizes the implicit Euler Voronoi finite volume method for nonlinear diffusion-convection-reaction equations on simplicial grids. The resulting nonlinear systems of equations are solved using Newton's method with parameter embedding. An advantage of the implementation in Julia is the availability of `ForwardDiff.jl` [115], an automatic differentiation package. This package allows the assembly of analytical Jacobians based on a generic implementation of nonlinear parameter functions without the need to write source code for derivatives.

2.5.1 1D time evolution and convergence test

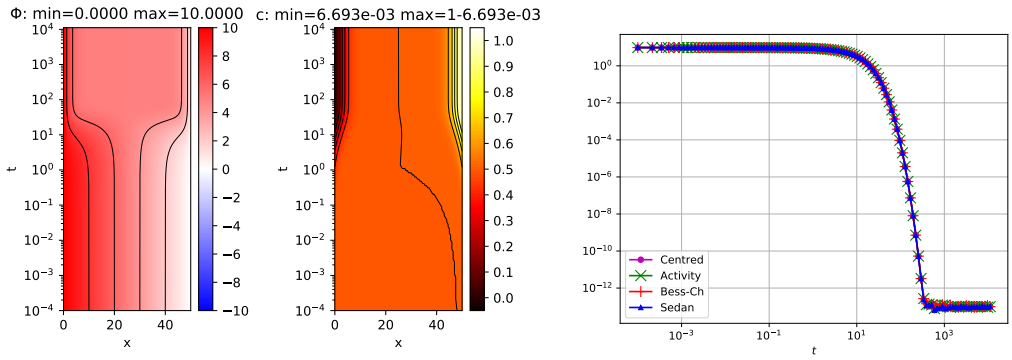


Figure 2.3 – Left: time evolution of solution on domain $\Omega = (0, 50)$ with constant initial value $c = \frac{1}{2}$, Dirichlet boundary conditions $\Phi(0) = 10$, $\Phi(50) = 0$, $c^{\mathrm{dop}} = -\frac{1}{2}$ and homogeneous Neumann boundary conditions for c. Right: Evolution of the relative free energy according to (2.1.11).

The first group of examples considers the problem as described by (2.1.1)-(2.1.3) in a one-dimensional domain with Dirichlet boundary conditions for Φ and homogeneous Neumann boundary conditions for c. We regard the time evolution from a zero potential Φ and constant concentrations c_0. In all examples, we assume a constant doping concentration $c^{\mathrm{dop}} = -\frac{1}{2}$. Calculations have been performed

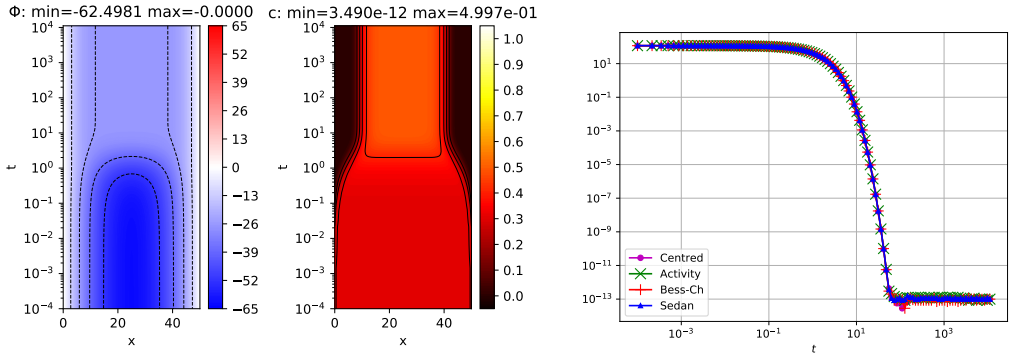


Figure 2.4 – Left: time evolution of solution on domain $\Omega = (0, 50)$ with constant initial value $c = 0.3$, Dirichlet boundary conditions $\Phi(0) = 0$, $\Phi(50) = 0$, $c^{\text{dop}} = -\frac{1}{2}$ and homogeneous Neumann boundary conditions for c. Right: Evolution of the relative free energy according to (2.1.11).

with subdivision of the domain $\Omega = (0, 50)$ into 100 control volumes. Time steps have been chosen in a geometric progression $t_i = t_1 * \delta^i$ with $\delta = 1.15$ and $t_1 = 10^{-4}$.

In the first example (Fig. 2.3), $c_0 = 0.5$, and the initial amount of charge carriers exactly matches the amount of doping. With the start of time evolution, at $x = 0$ a potential of 10 is applied leading to a redistribution of the charge carrier concentration which for large t approaches a steady state with two space charge regions at the boundaries with opposite charge and an electroneutral region with $c = 0.5$ in the center of the domain. We remark that c stays in the range $(0, 1)$, and that the energy (2.1.11) decreases during time evolution for all four schemes discussed in this paper. We also remark that for zero applied potential, the constant values $\Phi = 0$ and $c = 0.5$ would comprise a solution for all $t > 0$.

Fig. 2.4 considers the case $c_0 = 0.3$. The available amount of charge carriers is not able to compensate for the amount of doping. At the end of the time evolution, the charge carriers are concentrated in the center of the domain, establishing an electroneutral region. At both boundaries, depletion boundary layers create equally charged space charge regions due to the lack of charge carriers able to compensate the doping.

Fig. 2.5 considers the case $c_0 = 0.7$ which in sense is symmetric to the previous one. There is again an electroneutral region in the center, and this time, "superfluous" charge carriers are forced to enrichment boundary layers.

Fig. 2.6 provides a comparison of the convergence behaviour for the test case discussed in Fig. 2.3. We compare the solutions at a moment of time where we observe a rather large descent of the relative free energy based on a reference solution obtained on a fine grid of 40960 nodes using scheme (S). No visible

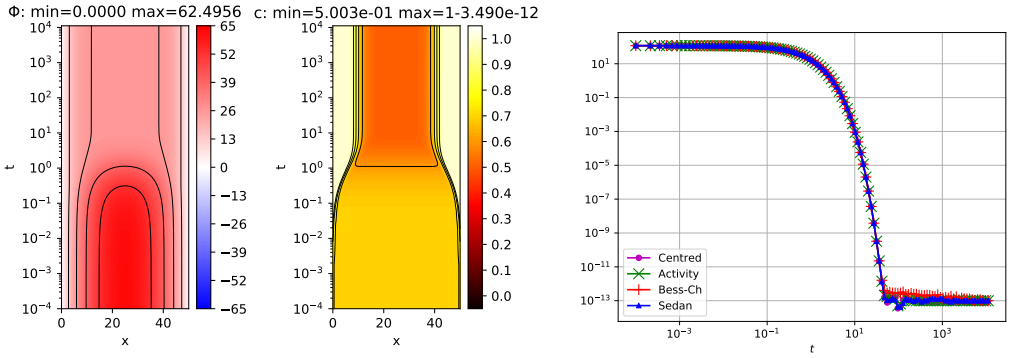


Figure 2.5 – Left: time evolution of solution on domain $\Omega = (0, 50)$ with constant initial value $c = 0.7$, Dirichlet boundary conditions $\Phi(0) = 0$, $\Phi(50) = 0$, $c^{\mathrm{dop}} = -\frac{1}{2}$ and homogeneous Neumann boundary conditions for c. Right: Evolution of the relative free energy according to (2.1.11).

difference in the plot have been found when using one of the other schemes to obtain the reference solution.

We observe first order convergence in the H^1 norm and second order convergence in the L^2 norm. No significant difference between the results for the various schemes.

2.5.2 1D stationary convergence test

In order to reveal the behaviour of the various schemes under more extreme conditions, this convergence test outside of thermodynamic equilibrium includes regions of the solution with concentrations extremely close to 0 and 1, respectively, enforced by inhomogeneous Dirichlet boundary conditions for the concentration, thus leaving the realm of the analysis in this paper. Once again, we assume $\Omega = (0, L)$ with $L = 50$, $c^{\mathrm{dop}} = -\frac{1}{2}$. We set boundary values $\Phi(0) = \Phi(L) = 0$ for the electrostatic potential, and $c(0) = 10^{-3}, c(L) = 1 - 10^{-3}$. We calculate a reference solution using the scheme (S) on a fine grid of 40960 nodes with grid spacing $h \simeq 1.22 \cdot 10^{-3}$, see Fig. 2.7. We use this solution as a surrogate for an analytical solution in a numerical investigation of the convergence rates of the different schemes. While no visible differences have been detected when using the schemes (AB) or (BC) for reference, for the slower converging scheme (C) as reference flux one would need a finer reference mesh to obtain similar results.

The result is shown in Fig. 2.8. We observe, that both in the H^1 and the L^2 norms, the schemes based on the modification of the Scharfetter-Gummel idea behave significantly better than the centred scheme. This is probably due to the

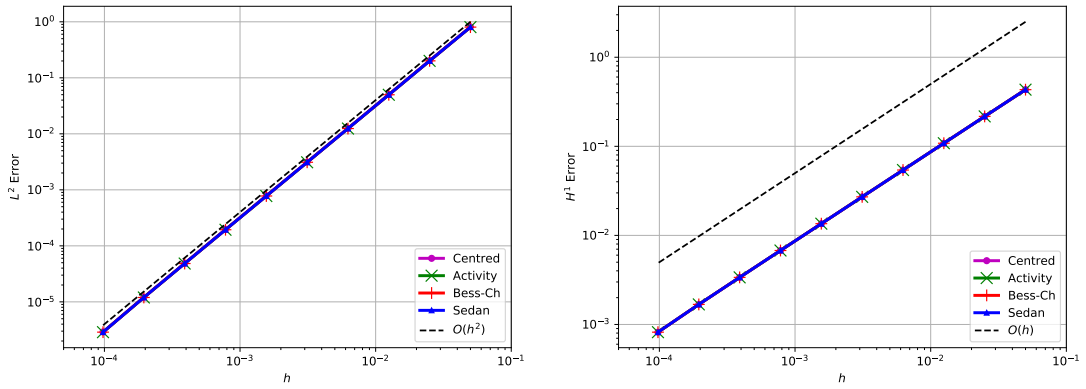


Figure 2.6 – Convergence behaviour of the different schemes for the case depicted in Fig. 2.3: comparison of solutions at $t = 10$. Left: L^2-error, right: H^1 error. Correspondence to the equation in the paper: "centred": (C), "Sedan": (S), "Activity": (AB), "Bess-Ch": (BC).

Dirichlet boundary condition close to 0 where the function $c \mapsto h(c)$ appearing explicitly in the centred scheme is singular. Judging from the L^2 error plot in Fig. 2.8 (left), the scheme (S) converges better than all the others. Asymptotically, all schemes show the same standard behaviour: we observe second order convergence in the L^2 norm and first order convergence in the H^1-norm.

2.5.3 2D Unipolar Field Effect Transistor

As a second example, we consider a unipolar field effect transistor. The domain is $\Omega = (0, L) \times (0, H)$ with $L = 10^{-1}$, $H = 10^{-2}$. We let $c^{\mathrm{dop}} = -\frac{1}{2}$, and set the following boundary conditions at the contacts:

$$\begin{aligned}
\begin{pmatrix} \Phi \\ c \end{pmatrix} &= \begin{pmatrix} -5 \\ \frac{1}{2} \end{pmatrix} \text{ at } \Gamma_{source} = (0, 0.2 \cdot L) \times H, \\
\begin{pmatrix} \Phi \\ c \end{pmatrix} &= \begin{pmatrix} 5 \\ \frac{1}{2} \end{pmatrix} \text{ at } \Gamma_{drain} = (0.8 \cdot L, L) \times H, \\
\begin{pmatrix} \nabla\Phi \cdot \mathbf{n} \\ \mathbf{J} \cdot \mathbf{n} \end{pmatrix} &= \begin{pmatrix} -\frac{1}{d}(\Phi - U_{gate}) \\ 0 \end{pmatrix} \text{ at } \Gamma_{gate} = (0.3 \cdot L, 0.7 \cdot L) \times H, \\
\begin{pmatrix} \nabla\Phi \cdot \mathbf{n} \\ \mathbf{J} \cdot \mathbf{n} \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ at } \partial\Omega \setminus (\Gamma_{gate} \cup \Gamma_{source} \cup \Gamma_{drain}).
\end{aligned}$$

Here, $\Phi_{gate} \in (-50, 50)$ is the gate voltage, and $d = 0.1 \cdot H$ is the gate thickness. We introduce a slightly anisotropic rectangular grid $n_x \times n_y$ with $n_x = 10 \times 2^{n_{ref}}$

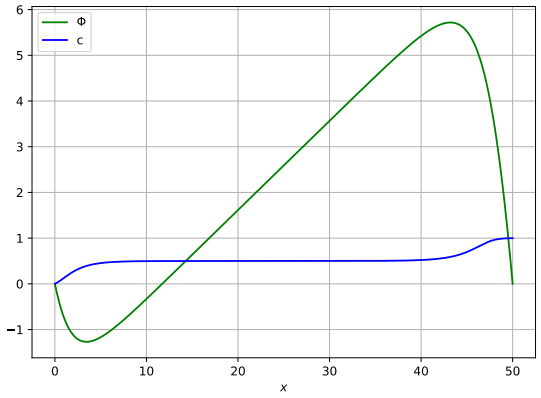


Figure 2.7 – Stationary solution with Dirichlet boundary conditions for c and Φ.

and $n_y = 5 \times 2^{n_{ref}}$, where n_{ref} is the refinement level. Each cell in the rectangular grid is subdivided into two triangles, see Fig. 2.9 (left). From the resulting triangle mesh, the Voronoi tessellation is obtained.

With fixed source and drain voltages, we vary the gate voltage U_{gate} from 50 to -50. At $U_{gate} = 50$, the positive applied potential pushes away the positively charged carriers from the channel – the region under the gate contact, see Fig. 2.10. The resulting lack of charge carriers results in a near zero current. With decreasing gate voltage, more and more charge carriers are allowed into the channel, leading to an increase in the current. When the gate voltage decreases further, charge carriers are attracted to the gate contact and fill up the channel. Due to the degeneration, their concentration cannot exceed 1. As a result, we observe a saturation of the current close to some maximum value for gate voltages approaching -50, see Fig. 2.10.

All schemes under consideration except (AB) represent this saturation behaviour quite well already at rather coarse grids, see Fig. 2.9 (right). This appears to be in line with earlier investigations of the scheme based on activity averaging [69] which hint that its asymptotic behaviour for large electric fields is not satisfactory.

To get an idea about the convergence in this case, we produce a reference solution on a grid with 821121 nodes using the scheme (S) and compare the calculated I-V curves. The behaviour of the error in the I-V curves is shown in Fig. 2.13. While all four schemes exhibit convergence of order at least $O(h)$, the activity based scheme (AB) converges with a constant approximately one order of magnitude larger than the others.

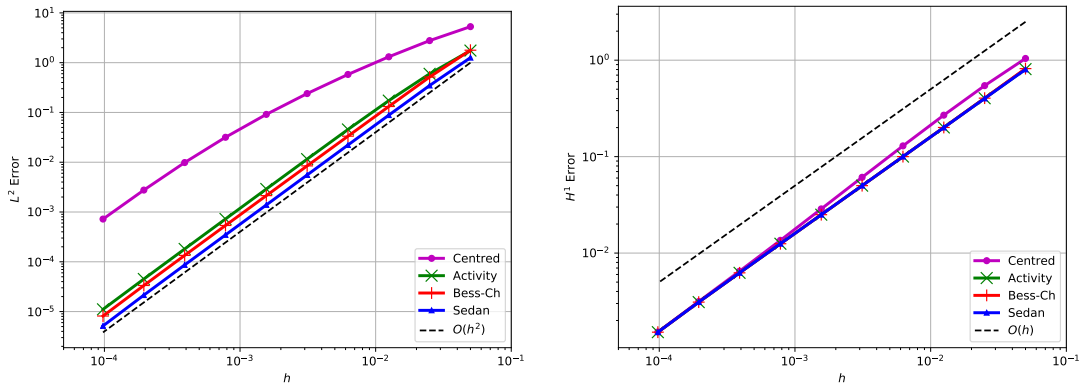


Figure 2.8 – Convergence behaviour of the different schemes. Left: L^2-error, right: H^1 error. Correspondence to the equation in the paper: "centred": (C), "Sedan": (S), "Activity": (AB), "Bess-Ch": (BC).

2.6 Conclusion

Four finite volume numerical schemes for a degenerate unipolar drift-diffusion model have been studied both from a numerical and theoretical point of view. Three of them – the schemes (AB), (S) and (BC) – can be seen as generalizations of the classical Scharfetter-Gummel scheme [116] inspired by different ways to express the degeneracy of the carrier density in the continuous model. Existence of the discrete solution and monotone decrease of the relative free energy have been proven for all four of them. We were able to prove rigorously the convergence to a solution of the continuous problem for only two of the schemes, namely (S) and (C). However, numerical experiments suggest that all the four schemes converge and are of order two with respect to space, even though some particular test cases show limitations for the schemes (AB) and (C). Besides, the extension of the scheme (BC) to more complex physics involving several conservation laws is not straightforward. Moreover, a robust implementation of scheme (BC) requires additional efforts to handle the case of constant concentrations. The present study suggests a preference for scheme (S) in practical applications as long as the mobility is linear. In the case of nonlinear mobilities (like e.g. $c(1-c)$), the extension of the schemes (AB), (S) and (BC) is unclear and a scheme based on (C) seems to be a good option.

Acknowledgements. This work was partially supported by Labex CEMPI (ANR-11-LABX-0007-01). Claire Chainais-Hillairet was also supported by project MoHyCon (ANR-17-CE40-0027-01). Finally, the authors warmly thank the anony-

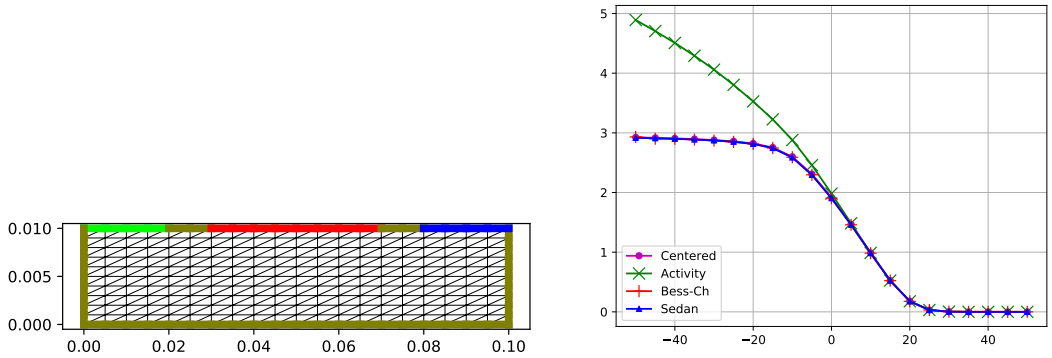


Figure 2.9 – Discretization grid of refinement level $n_{ref} = 1$ (left) and corresponding I-V curves for different discretization schemes (right).

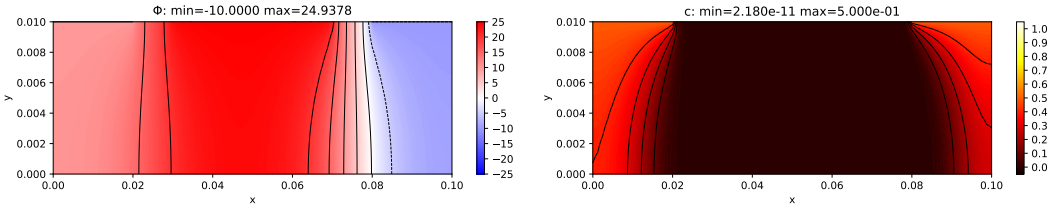


Figure 2.10 – Electrostatic potential (left) and concentration (right) for closed gate ($U_{gate} = 50$).

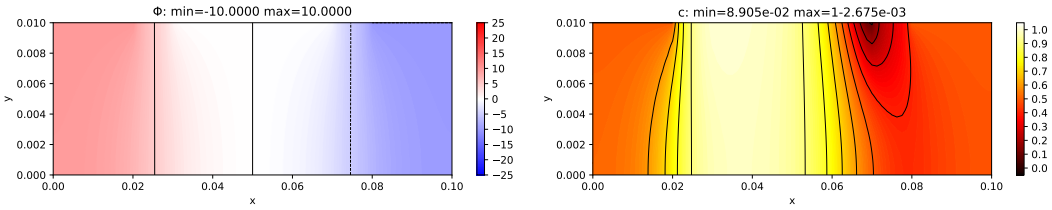


Figure 2.11 – Electrostatic potential (left) and concentration (right) for $U_{gate} = 0$).

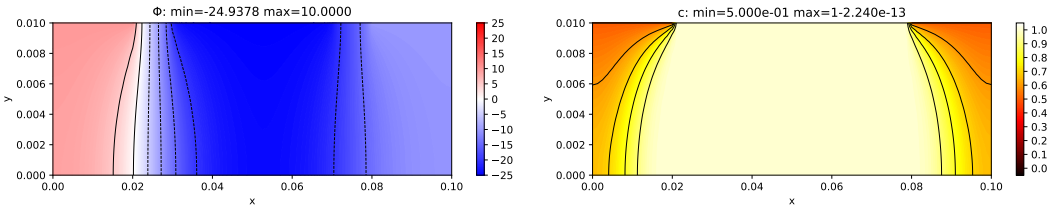


Figure 2.12 – Electrostatic potential (left) and concentration (right) for open gate ($U_{gate} = -50$), with concentration in the channel reaching the saturation value 1.

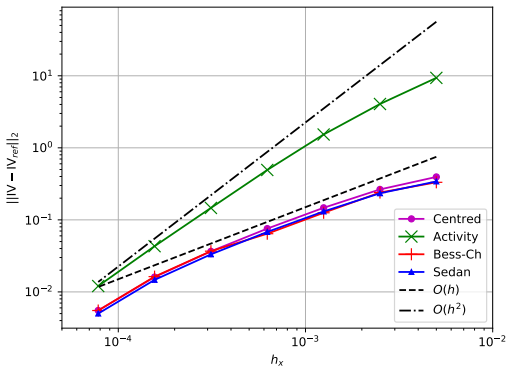


Figure 2.13 – Convergence of the I-V curves calculated using the different discretization schemes.

mous referees for their constructive feedback.

2.A L^∞ bound on the TPFA FV approximate Poisson equation

It is well known that the solution to the Poisson equation

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = u^D & \text{on } \Gamma^D, \\ \nabla u \cdot \boldsymbol{n} = 0 & \text{on } \Gamma^N, \end{cases} \tag{2.A.1}$$

is bounded in $L^\infty(\Omega)$ provided $f \in L^\infty(\Omega)$ and $u^D \in L^\infty(\partial\Omega)$. The goal of this appendix is to get a discrete counterpart of this estimate for TPFA finite volume approximations of (2.A.1). The data u^D and f are discretized into

$$u_\sigma^D = \frac{1}{m_\sigma}\int_\sigma u^D(\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\gamma}, \qquad f_K = \frac{1}{m_K}\int_K f\mathrm{d}\boldsymbol{x}, \qquad \sigma \in \mathcal{E}^D,\ K \in \mathcal{T}.$$

and the classical TPFA finite volume scheme is:

$$-\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D_{K\sigma} u = m_K f_K, \quad \forall K \in \mathcal{T}.$$

The associate linear system of equations can be written as

$$\mathbb{L}\boldsymbol{u} = \boldsymbol{b}, \tag{2.A.2}$$

with $\boldsymbol{u} = (u_K, u_\sigma)_{K\in\mathcal{T},\sigma\in\mathcal{E}^D}$ (let us note that we keep the Dirichlet nodes in the set of unknowns), $\boldsymbol{b} = \left(f_K, u_\sigma^D\right)_{K\in\mathcal{T},\sigma\in\mathcal{E}^D}$ and $\mathbb{L} \in \mathbb{R}^{(\mathcal{T}\cup\mathcal{E}^D)\times(\mathcal{T}\cup\mathcal{E}^D)}$ is the sparse symmetric definite positive matrix defined by

$$\mathbb{L}_{\sigma,\sigma} = 1, \qquad \mathbb{L}_{\sigma,\ell} = 0 \text{ if } \ell \neq \sigma, \qquad \sigma \in \mathcal{E}^D,$$

$$\mathbb{L}_{K,K\sigma} = -\frac{\tau_\sigma}{m_K}, \qquad \mathbb{L}_{K,K} = \frac{1}{m_K}\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma, \qquad K \in \mathcal{T}.$$

In the above definition of $\mathbb{L}$, ℓ denotes an arbitrary index in $\mathcal{T}\cap\mathcal{E}^D$, whereas $K\sigma$ denotes the mirror index of K w.r.t. the faces $\sigma \in \mathcal{E}_K$, i.e., $K\sigma = L$ if $\sigma = K|L \in \mathcal{E}_K \cap \mathcal{E}_{\text{int}}$ and $K\sigma = \sigma$ if $\sigma \in \mathcal{E}_K \cap \mathcal{E}^D$.

The goal of this section is to derive an ℓ^∞ bound on the solution $\boldsymbol{u}$ to the linear system (2.A.2) which is uniform w.r.t. the mesh. This is the purpose of the

following proposition.

Proposition 2.A.1. *There exists C depending only on Ω such that*

$$|u_K| \le C \max\left\{\|u^D\|_{L^\infty(\partial\Omega)}, \|f\|_{L^\infty(\Omega)}\right\}, \qquad \forall K \in \mathcal{T}.$$

Proof. The proof we propose here is an extension to the context of TPFA Finite Volumes of the proof of Hackbusch [88] for Finite Differences. An alternative proof of Proposition 2.A.1 based on Stampacchia's truncation estimates is sketched in [76].

The definitions of u_σ^D and f_K ensure that

$$\|\boldsymbol{b}\|_\infty \le \max\left\{\|u^D\|_{L^\infty(\partial\Omega)}, \|f\|_{L^\infty(\Omega)}\right\},$$

so that

$$\|\boldsymbol{u}\|_\infty \le \|\mathbb{L}^{-1}\|_\infty \|\boldsymbol{b}\|_\infty \le \|\mathbb{L}^{-1}\|_\infty \max\left\{\|u^D\|_{L^\infty(\partial\Omega)}, \|f\|_{L^\infty(\Omega)}\right\}.$$

Therefore, it only remains to check that $\|\mathbb{L}^{-1}\|_\infty \le C$ for some C not depending on $\mathcal{T}$.

The matrix $\mathbb{L}$ is a M-matrix (see [88, Definition 4.8]). Therefore, owing to [88, Theorem 4.24], if we can exhibit some vector $\boldsymbol{w} \in \mathbb{R}^{\mathcal{T}\cup\mathcal{E}^D}$ such that $\mathbb{L}\boldsymbol{w} \ge \mathbf{1}$, then $\|\mathbb{L}^{-1}\|_\infty \le \|\boldsymbol{w}\|_\infty$. Define the function $w : \overline{\Omega} \to \mathbb{R}$ by

$$w(\boldsymbol{x}) = 1 + \frac{1}{d}\left(\sup_{\boldsymbol{y}\in\Omega} |\boldsymbol{y}|^2 - |\boldsymbol{x}|^2\right) \ge 1, \qquad \boldsymbol{x} \in \overline{\Omega},$$

and the vector $\boldsymbol{w} = (w_K, w_\sigma)$ by $w_K = w(\boldsymbol{x}_K)$, $K \in \mathcal{T}$, and $w_\sigma = w(\boldsymbol{x}_\sigma)$, $\sigma \in \mathcal{E}^D$.

The estimate on the Dirichlet nodes is straightforward:

$$(\mathbb{L}\boldsymbol{w})_\sigma = w_\sigma \ge 1, \qquad \forall \sigma \in \mathcal{E}^D.$$

Now, we focus on the inner nodes $K \in \mathcal{T}$. Since $\sum_{\ell\in\mathcal{T}\cup\mathcal{E}^D} \mathbb{L}_{K,\ell} = \sum_{\sigma\in\mathcal{E}_K} \mathbb{L}_{K,K\sigma} = 0$, one has

$$\begin{aligned}
(\mathbb{L}\boldsymbol{w})_K =& \frac{1}{d}\sum_{\sigma\in\mathcal{E}_K} \mathbb{L}_{K,K\sigma}|\boldsymbol{x}_{K\sigma}|^2 = \frac{1}{dm_K}\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\left(|\boldsymbol{x}_K|^2 - |\boldsymbol{x}_{K\sigma}|^2\right)\\
=& \frac{1}{dm_K}\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\left(|\boldsymbol{x}_K - \boldsymbol{x}_{K\sigma}|^2 + 2\boldsymbol{x}_K\cdot(\boldsymbol{x}_K - \boldsymbol{x}_{K\sigma})\right)\\
=& \frac{1}{dm_K}\sum_{\sigma\in\mathcal{E}_K} m_\sigma d_\sigma + \frac{2}{dm_K}\boldsymbol{x}_K \cdot \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma(\boldsymbol{x}_K - \boldsymbol{x}_{K\sigma}).
\end{aligned}$$

Because of the geometric relation $m_\sigma d_\sigma = dm_{\Delta_\sigma}$, and since $K \subset \bigcup_{\sigma\in\mathcal{E}_K} \Delta_\sigma$, there holds

$$\frac{1}{dm_K}\sum_{\sigma\in\mathcal{E}_K} m_\sigma d_\sigma = \frac{1}{m_K}\sum_{\sigma\in\mathcal{E}_K} m_{\Delta_\sigma} \geq 1.$$

On the other hand, the second term vanishes since

$$\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma(\boldsymbol{x}_K - \boldsymbol{x}_{K\sigma}) = -\sum_{\sigma\in\mathcal{E}_K} m_\sigma \boldsymbol{n}_{K\sigma} = \mathbf{0}.$$

Therefore, $(\mathbb{L}\boldsymbol{w})_K \geq 1$ for all $K \in \mathcal{T}$. As a consequence,

$$\|\mathbb{L}^{-1}\|_\infty \leq \|\boldsymbol{w}\|_\infty = 1 + \frac{1}{d}\sup_{\boldsymbol{y}\in\Omega}|\boldsymbol{y}|^2 \leq 1 + \frac{\mathrm{diam}(\Omega)^2}{4d}.$$

The last estimate comes from the fact that one can choose the origin for $\boldsymbol{y}$ arbitrarily. □

2.B Proof of Lemma 2.3.2

Step 1. Let $\delta \in (0,1)$ and $M \in \mathbb{R}$. We start with the proof of

$$\lim_{c_L\to 1} \Upsilon_{\delta,M}(c_L) = +\infty, \tag{2.B.1}$$

where

$$\Upsilon_{\delta,M}(c_L) = \inf\left\{\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L); c_K \in (0, 1-\delta], (\Phi_K, \Phi_L) \in [-M, M]^2\right\}.$$

We recall that

$$\begin{aligned}\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L) &= \mathcal{C}(c_K, c_L, \Phi_K, \Phi_L)|h(c_K) + \Phi_K - h(c_L) - \Phi_L|^2 \\ &= \mathcal{F}(c_K, c_L, \Phi_K, \Phi_L)(h(c_K) + \Phi_K - h(c_L) - \Phi_L),\end{aligned}$$

and we notice that the diffusion force blows up:

$$\lim_{c_L\to 1}\inf\left\{\left|h(c_K) - h(c_L) + \Phi_K - \Phi_L\right|; c_K \in (0, 1-\delta], (\Phi_K, \Phi_L) \in [-M, M]^2\right\} = +\infty. \tag{2.B.2}$$

Therefore, we can get (2.B.1) by proving that either $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L)$ or simply $\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L)$ stays bounded away from 0, uniformly in $c_K \in (0, 1-\delta]$, $(\Phi_K, \Phi_L) \in [-M, M]^2$, for $c_L \geq 1/2$.

For the centred flux, we have that, for all $(c_K, \Phi_K, \Phi_L) \in (0, 1-\delta] \times [-M, M]^2$, $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) = (c_K + c_L)/2 \geq c_L/2$. This yields (2.B.1). For the three other

schemes, we remark that, for any $\alpha \in (0, 1-\delta)$, we have

$$\Upsilon_{\delta,M}(c_L) = \min(\Upsilon^{\alpha,1}_{\delta,M}(c_L), \Upsilon^{\alpha,2}_{\delta,M}(c_L)),$$

where

$$\Upsilon^{\alpha,1}_{\delta,M}(c_L) = \inf\left\{\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L); c_K \in (0,\alpha), (\Phi_K, \Phi_L) \in [-M, M]^2\right\},$$
$$\Upsilon^{\alpha,2}_{\delta,M}(c_L) = \inf\left\{\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L); c_K \in [\alpha, 1-\delta], (\Phi_K, \Phi_L) \in [-M, M]^2\right\}.$$

The Lemma 2.3.1 ensures that, independently of the choice of the numerical flux, we have at least $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) \geq \min(c_K, c_L)/2$, so that $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) \geq \alpha/2$ for all $(c_K, c_L, \Phi_K, \Phi_L) \in [\alpha, 1-\delta] \times [1/2, 1) \times [-M, M]^2$ if $\alpha \in (0, 1-\delta)$.Therefore, for all $\alpha \in (0, 1-\delta)$, we have

$$\lim_{c_L \to 1} \Upsilon^{\alpha,2}_{\delta,M}(c_L) = +\infty.$$

It remains to prove that for a given $\alpha \in (0, 1-\delta)$ we also have

$$\lim_{c_L \to 1} \Upsilon^{\alpha,1}_{\delta,M}(c_L) = +\infty. \tag{2.B.3}$$

Because of the monotonicity of $\delta \mapsto \Upsilon_{\delta,M}(c_L)$, we can restrict our attention to the case $\delta \leq 1/2$, so that we can seek for $\alpha \in (0, 1/2]$.

For the Bessemoulin-Chatard flux, we have

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = \mathfrak{d}r(c_K, c_L)\left\{B\left(\frac{\Phi_L - \Phi_K}{\mathfrak{d}r(c_K, c_L)}\right)c_K - B\left(\frac{\Phi_K - \Phi_L}{\mathfrak{d}r(c_K, c_L)}\right)c_L\right\},$$

with $\mathfrak{d}r(c_K, c_L) \geq 1$. Using the monotonicity of the Bernoulli function and the bounds on Φ_K and Φ_L, we get:

$$B(2M) \leqslant B\left(\pm\frac{\Phi_L - \Phi_K}{\mathfrak{d}r(c_K, c_L)}\right) \leqslant B(-2M).$$

Hence, for $\alpha = \frac{B(2M)}{4B(-2M)}$:

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) \leqslant \mathfrak{d}r(c_K, c_L)\left\{B(-2M)\alpha - B(2M)\frac{1}{2}\right\} \leqslant -\frac{B(2M)}{4}.$$

Then, thanks to (2.B.2), we deduce (2.B.3) for $\alpha = \frac{B(2M)}{4B(-2M)}$ and therefore (2.B.1).

For the Sedan flux, we use similarly the monotonicity of the function B and ν,

so that

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) \leq B\left(-2M + \nu(\frac{1}{2}) - \nu(\alpha)\right)\alpha - B\left(2M - \nu(\frac{1}{2}) + \nu(\alpha)\right)\frac{1}{2},$$
$$\forall c_K \in (0,\alpha), c_L \in (\frac{1}{2},1), (\Phi_K,\Phi_L) \in [-M,M]^2.$$

The right-hand side of the last inequality tends to $-B\left(2M - \nu(\frac{1}{2})\right)\frac{1}{2}$ when α tends to 0. The negativity of this limit means that for a given α small enough the flux remains bounded away from 0 so that we deduce (2.B.3) and therefore (2.B.1).

For the activity based flux, we also use the monotonicity of a and β, which yields

$$\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) \leq \frac{1}{4}\left(B(-2M)a(\alpha) - B(2M)a(\frac{1}{2})\right),$$
$$\forall c_K \in (0,\alpha), c_L \in (\frac{1}{2},1), (\Phi_K,\Phi_L) \in [-M,M]^2.$$

The right-hand side has a negative limit when α tends to 0. Thus, it remains bounded away from 0 for a given $\alpha < 1/2$, and we deduce (2.B.3) and therefore (2.B.1).

Step 2. We now focus on the proof of

$$\lim_{c_L \to 0} \Psi_{\delta,M}(c_L) = +\infty \tag{2.B.4}$$

where

$$\Psi_{\delta,M}(c_L) = \inf\left\{\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L); c_K \in [\delta,1), (\Phi_K,\Phi_L) \in [-M,M]^2\right\}.$$

We use similar arguments than in Step 1. First, the diffusion force still blows up:

$$\lim_{c_L \to 0} \inf\left\{\left|h(c_K) - h(c_L) + \Phi_K - \Phi_L\right|; c_K \in [\delta,1), (\Phi_K,\Phi_L) \in [-M,M]^2\right\} = +\infty. \tag{2.B.5}$$

For the centred flux, we have: $\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) = (c_K + c_L)/2 \geq \delta/2$ hence (2.B.4). For the other fluxes, we split using a parameter α again

$$\Psi_{\delta,M}(c_L) = \min(\Psi^{\alpha,1}_{\delta,M}(c_L), \Psi^{\alpha,2}_{\delta,M}(c_L)),$$

where

$$\Psi^{\alpha,1}_{\delta,M}(c_L) = \inf\left\{\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L); c_K \in [\delta, \alpha], (\Phi_K, \Phi_L) \in [-M, M]^2\right\};$$
$$\Psi^{\alpha,2}_{\delta,M}(c_L) = \inf\left\{\mathcal{D}(c_K, c_L, \Phi_K, \Phi_L); c_K \in (\alpha, 1), (\Phi_K, \Phi_L) \in [-M, M]^2\right\}.$$

Using the symmetry of the flux $\mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) = -\mathcal{F}(c_L, c_K, \Phi_L, \Phi_K)$ and following the proof of (2.B.3), we get that for $\alpha = 1/2$,

$$\lim_{c_L \to 0} \Psi^{\alpha,2}_{\delta,M}(c_L) = +\infty$$

We now have to prove that, for $\alpha = 1/2$,

$$\lim_{c_L \to 0} \Psi^{\alpha,1}_{\delta,M}(c_L) = +\infty \tag{2.B.6}$$

To this end, we will show bounds on the flux. The set $[\delta, \alpha] \times [-M, M]^2$ is compact, and the flux functions are continuous. It is sufficient to show a positive lower bound for the limit at any $(c^*, \Phi^*, \Phi_*) \in [\delta, \alpha] \times [-M, M]^2$:

$$l^* = \lim_{(c_K, c_L, \Phi_K, \Phi_L) \to (c^*, 0, \Phi^*, \Phi_*)} \mathcal{F}(c_K, c_L, \Phi_K, \Phi_L) > 0.$$

For the Sedan scheme, we have:

$$l^* = B\left(\Phi_* - \Phi^* - \nu(c^*)\right)\right) c^* \geq \delta B(2M).$$

For the Bessemoulin-Chatard scheme, we have: $\lim_{(c_K, c_L) \to (c^*, 0)} \mathfrak{d}r(c_K, c_L) = 1$, hence:

$$l^* = B(\Phi_* - \Phi^*)c^* \geq \delta B(2M).$$

For the activity based scheme we have:

$$l^* = \frac{\beta(c^*) + 1}{2} B(\Phi_* - \Phi^*)a(c^*) \geq \frac{a(\delta)}{2} B(2M).$$

As these limits are bounded away from zero we have (2.B.6) hence (2.B.4). This concludes the proof of Lemma 2.3.2.

2.C Comparison of face concentration functionals

For each scheme, we have defined a face concentration functional $\mathcal{C} : (0,1) \times (0,1) \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. We introduce a second face concentration functional $\widetilde{\mathcal{C}}$:

$(0,1)\times(0,1)\to\mathbb{R}$, defined by

$$\widetilde{\mathcal{C}}(c_K,c_L)=\frac{r(c_K)-r(c_L)}{h(c_K)-h(c_L)} \text{ if } c_K\neq c_L \text{ and } c_K \text{ otherwise.}$$

As $r'(c)=ch'(c)$, it is clear that

$$\min(c_K,c_L)\leq\widetilde{\mathcal{C}}(c_K,c_L)\leq\max(c_K,c_L) \text{ for all } (c_K,c_L)\in(0,1)\times(0,1). \quad (2.C.1)$$

Lemma 2.C.1 states a comparison between $\mathcal{C}$ and $\widetilde{\mathcal{C}}$ for the centred and the Sedan schemes.

Lemma 2.C.1. *For the centred scheme and the Sedan scheme, there exists $G>0$, depending only on M, such that for all $(c_K,c_L,\Phi_K,\Phi_L)\in(0,1)\times(0,1)\times[-M,M]\times[-M,M]$,*

$$\frac{\widetilde{\mathcal{C}}(c_K,c_L)}{\mathcal{C}(c_K,c_L,\Phi_K,\Phi_L)}\leq G. \quad (2.C.2)$$

Remark 2.C.1. For the Bessemoulin-Chatard scheme, the bound (2.C.2) does not hold. Let us consider that $\Phi_K=\Phi_L=\Phi$, then with the notations $x=\log(c_K/c_L)$, and $y=\log(\frac{1-c_L}{1-c_K})$, we have:

$$\frac{\tilde{\mathcal{C}}(c_K,c_L)}{\mathcal{C}(c_K,c_L,\Phi,\Phi)}=\frac{xy}{(c_K-c_L)(x+y)}.$$

For $(c_K,c_L)\to(1,0)$, x and y tends to $+\infty$, hence the blow up of the ratio.

Proof. The case of the centred scheme defined by (C) is the easiest one, since

$$\mathcal{C}(c_K,c_L,\Phi_K,\Phi_L)=\frac{c_K+c_L}{2}\geq\frac{1}{2}\max(c_K,c_L),$$

so that (2.C.2) holds with $G=2$ thanks to (2.C.1).

Let us now focus on the Sedan scheme defined by (S). We can introduce the function $\mathcal{G}:(0,1)\times(0,1)\to\mathbb{R}$ defined by

$$\mathcal{G}(c_K,c_L)=\frac{\widetilde{\mathcal{C}}(c_K,c_L)}{\min_{(\Phi_K,\Phi_L)\in[-M,M]^2}\mathcal{C}(c_K,c_L,\Phi_K,\Phi_L)}.$$

It is a continuous function which satisfies the symmetry property $\mathcal{G}(c_K,c_L)=\mathcal{G}(c_L,c_K)$ and the consistency $\mathcal{G}(c_K,c_K)=1$.

Because of the symmetry and the consistency properties, we can assume without loss of generality that $c_K>c_L$. Using the average properties (2.3.3) and (2.C.1),

one obtains that

$$\mathcal{G}(c_K, c_L) \leq \frac{c_K}{c_L} \leq \frac{1}{c_L}, \tag{2.C.3}$$

so that we only have to check that $\mathcal{G}(c_K, c_L)$ remains uniformly bounded as c_L tends to 0 to prove (2.C.2). To that extent, we compute explicitly the minimum of $\mathcal{C}$. We recall that we have, using (2.3.5):

$$\mathcal{C}(c_K, c_L, \Phi_K, \Phi_L) = tc_K + (1-t)c_L, \qquad t = \frac{B(y) - B(x)}{x - y},$$

where $y = \Phi_L + \nu(c_L) - \Phi_K - \nu(c_K)$ and $x = \log(c_K/c_L)$. Using the assumption $c_K > c_L$, $\mathcal{C}$ is minimal when t is minimal. As B is convex, and x fixed, this happens for y maximal, i.e.

$$y = 2M + \nu(c_L) - \nu(c_K).$$

Using this result, one can expand

$$\mathcal{G}(c_K, c_L) = \frac{(h(c_K) - h(c_L) - 2M)(r(c_K) - r(c_L))}{\big(B(2M + \nu(c_L) - \nu(c_K))c_K - B(-2M - \nu(c_L) + \nu(c_K))c_L\big)(h(c_K) - h(c_L))}$$

Therefore we study the limit of $\mathcal{G}$ when (c_K, c_L) tends to (1,0), (0,0) and $(c^*, 0)$ with $c^* \in (0,1)$.

We first consider the limit $(c_K, c_L) \to (1,0)$. We have the following equivalences when $(c_K, c_L) \to (1,0)$:

$$\begin{aligned} h(c_K) - h(c_L) &\sim -\log(1-c_K) - \log(c_L) \\ h(c_K) - h(c_L) - 2M &\sim -\log(1-c_K) - \log(c_L) \\ r(c_K) - r(c_L) &\sim -\log(1-c_K) \\ B(y)c_K - B(-y)c_L &\sim -c_K \log(1-c_K) \end{aligned}$$

This yields:

$$\lim_{(c_K,c_L)\to(1,0)} \mathcal{G}(c_K, c_L, \Phi_K, \Phi_L) = 1. \tag{2.C.4}$$

With similar arguments, we compute the limit $(c_K, c_L) \to (c^*, 0)$ with $c^* \in (0,1)$. We get:

$$\lim_{(c_K,c_L)\to(c^*,0)} \mathcal{G}(c_K, c_L) = \frac{r(c^*)}{B(y^*)c^*}, \tag{2.C.5}$$

with $y^* = 2M + \log(1-c^*)$.

In the neighbourhood of $(0,0)$, the behaviour is more complex, as the limit of $\log(c_K/c_L)$ is not defined and $\mathcal{G}$ does not have a limit. However, thanks to (2.C.3), $\mathcal{G}(c_K, c_L)$ stays bounded if c_K/c_L stays bounded while $(c_K, c_L) \to (0,0)$. It remains

to consider the case where $(c_K, c_L) \to (0,0)$ while $c_K/c_L \to \infty$. In this case, we have:

$$\frac{(h(c_K) - h(c_L) - 2M)}{h(c_K) - h(c_L)} \to 1,$$
$$r(c_K) - r(c_L) \sim -c_L + c_K$$
$$B(y)c_K - B(-y)c_L \sim B(2M)c_K - B(-2M)c_L,$$

and

$$\lim_{\substack{(c_K,c_L)\to(0,0)\\ c_K/c_L\to\infty}} \mathcal{G}(c_K, c_L) = \frac{1}{B(2M)}.$$

We conclude that $\mathcal{G}(c_K, c_L)$ stays bounded when (c_K, c_L) is in the neighbourhood of $(0,0)$. Combined with (2.C.3), (2.C.4) and (2.C.5), this concludes the proof of Lemma 2.C.1. □

2.D Some notations

In this Section, we recall the definition of some notations used along the paper. Table 2.1 gives the definition of the different quantities involved at the continuous level, while Table 2.2 gives the definition of the functions involved in the study of the numerical schemes.

$h(c)$	$\log \dfrac{c}{1-c}$	chemical potential
$\nu(c)$	$-\log(1-c)$	excess chemical potential
$r(c)$	$-\log(1-c)$	diffusion enhancement
$a(c)$	$\dfrac{c}{1-c}$	activity coefficient
$\beta(c)$	$1-c$	inverse activity coefficient
$H(c)$	$c\log(c)+(1-c)\log(1-c)$	entropy density
$E(c,\Phi)$	$\displaystyle\int_\Omega\left\{H(c)+\frac{1}{2}\lvert\nabla\Phi\rvert^2\right\}\mathrm{d}\boldsymbol{x}-\int_{\Gamma_D}\Phi^D\nabla\Phi\cdot\boldsymbol{n}\mathrm{d}\boldsymbol{\gamma}$	free energy

Table 2.1 – Definition of the different functions involved in the continuous problem.

$B(x)$	$\dfrac{x}{e^x-1}$	Bernoulli function
$\mathfrak{d}r(c_K,c_L)$	$\begin{cases}\dfrac{h(c_K)-h(c_L)}{\log(c_K)-\log(c_L)} & \text{if } c_K \neq c_L,\\ r'(c_K) & \text{if } c_K = c_L.\end{cases}$	approximation of r' consistant with the thermal equilibrium
$\mathcal{F}(c_K,c_L,\Phi_K,\Phi_L)$	$\dfrac{F_{KL}}{\tau_{KL}}$	numerical flux intensity
$\mathcal{C}(c_K,c_L,\Phi_K,\Phi_L)$	$\dfrac{\mathcal{F}(c_K,c_L,\Phi_K,\Phi_L)}{h(c_K)+\Phi_K-h(c_L)-\Phi_L}$	face concentration
$\widetilde{\mathcal{C}}(c_K,c_L)$	$\begin{cases}\dfrac{r(c_K)-r(c_L)}{h(c_K)-h(c_L)} & \text{if } c_K \neq c_L\\ c_K & \text{otherwise}\end{cases}$	face concentration compliant with the weak solution
$\mathcal{D}(c_K,c_L,\Phi_K,\Phi_L)$	$\mathcal{F}(c_K,c_L,\Phi_K,\Phi_L)$ $\times (h(c_K)+\Phi_K-h(c_L)-\Phi_L)$	face entropy dissipation

Table 2.2 – Definition of the different functions involved in the numerical schemes.

Chapter 3

Entropy and convergence analysis for two finite volume schemes for a Nernst-Planck-Poisson system with ion volume constraints.

Ce chapitre est un travail en collaboration avec Jürgen Führmann. Il est disponible sur HAL et a été soumis pour publication.

In this paper, we consider a drift-diffusion system with cross-coupling through the chemical potentials comprising a model for the motion of finite size ions in liquid electrolytes. The drift term is due to the self-consistent electric field maintained by the ions and described by a Poisson equation. We design two finite volume schemes based on different formulations of the fluxes. We also provide a stability analysis of these schemes and an existence result for the corresponding discrete solutions. A convergence proof is proposed for non-degenerate solutions. Numerical experiments show the behavior of these schemes.

Outline of the current chapter

3.1 Introduction **88**
3.1.1 The Nernst-Planck-Poisson system with finite ionic volumes . . . 89
3.1.2 Key properties of the continuous system . . . 91
3.1.3 Positioning and outline . . . 94
3.2 Discretization and main Theorems **94**
3.2.1 Discretization of $(0,T)\times\Omega$ . . . 94
3.2.2 A common setting for the Finite Volume schemes . . . 96
3.2.3 Numerical fluxes for the conservation equations . . . 97
3.2.4 Main theorems . . . 99
3.3 Fixed Mesh analysis **100**
3.3.1 Analysis of numerical flux based functions . . . 100
3.3.2 *A priori* estimates . . . 102
3.3.3 Existence of solutions . . . 106
3.4 Convergence **108**
3.4.1 Reconstruction operators . . . 108
3.4.2 Compactness . . . 109
3.4.3 Identification . . . 122
3.5 Numerical Examples **126**
3.5.1 Species redistribution in a one-dimensional cell filled with binary electrolyte . . . 126
3.5.2 1D stationary convergence test . . . 127
3.5.3 An electrolytic diode . . . 127
3.A Chemical free energy density and chemical potentials **133**
3.B Proof of Lemma 3.3.2 **134**
3.B.1 Limit of $\Psi_{\delta,\epsilon,M,i}$ . . . 135
3.B.2 Limit of $\Upsilon_{\delta,M}$ . . . 136
3.C Study of a numerical scheme for $h_i = \log(c_i) - \alpha\log(c_0)$ **137**
3.D A simple convergence lemma **139**

3.1 Introduction

Proper modeling of the motion of ions in electrolytes – mixtures of a solvent and N ionic species which can be described by their concentrations c_i – and associated simulations are crucial in the development of efficient batteries, fuel cells, and many other applications commonly considered as key technologies for the 21st century. The classical Nernst-Planck equation is a linear system which for given

electrostatic potential Φ, charge number z_i and diffusion coefficient D_i describes the evolution of the ion concentration c_i via

$$\partial_t c_i - \operatorname{div}(D_i N_i) = 0, \qquad N_i = \nabla c_i + z_i c_i \nabla \Phi = c_i \nabla \left(\log(c_i) + z_i \Phi\right).$$

The self-consistent electrostatic potential is described by the Poisson equation

$$-\nabla \cdot \lambda^2 \nabla \Phi = \sum_{i=1}^{N} z_i c_i.$$

This model assumes that ions are infinitely small and that the ions of a given species i interact neither with the solvent nor with other ionic species. However, in reality, ion sizes are finite, and ion motion is only possible with a simultaneous displacement of solvent molecules. Moreover, the effective size of ions is increased by the fact that in a polar solvent like water, they are surrounded by a solvation shell consisting of a certain number of solvent molecules. The inclusion of these effects into the model is particularly important for concentrated electrolytes and in electrode boundary layers with high ion concentrations.

Historically, there have been many, often independent attempts to fix this situation, see e.g. the review in [12], the discussion in [70] or [99]. A comprehensive model of ideal mixtures of solvated ions has been derived in [60, 59]. In [70, 71], a two point flux finite volume discretization approach for these problems has been derived. Various variants of ionic flux approximations have been investigated for the unipolar case, where only one ionic species is considered, in Chapter 2, with the result that the flux approximation approach introduced in [70] has several more accurate alternatives. For two of them, we have been able to find appropriate generalizations to the case of several ionic species. These are introduced and analyzed in the present paper.

In the sequel of Section 3.1, the continuous problem is formulated, and several key properties of the continuous system are discussed. Among these is the decay of an entropy functional for positive solutions.

3.1.1 The Nernst-Planck-Poisson system with finite ionic volumes

Consider a bounded connected polytopal domain $\Omega \subset \mathbb{R}^d$, and finite simulation horizon $T > 0$. We model the evolution of the concentration c_0 of a solvent and N

dissolved species: c_i, $i \in [\![1, N]\!]$. The mixture satisfies a volume filling constraint

$$\sum_{i=0}^{N} v_i c_i = 1,$$

where v_i are the molar volumes of the species. We will use this constraint using ratios of molar volumes $k_i = \frac{v_i}{v_0}$:

$$\sum_{i=0}^{N} k_i c_i = \frac{1}{v_0}. \tag{3.1.1}$$

The coefficients $(k_1, \dots, k_N)$ are parameters of the problem and k_0 is by definition equal to 1. As the molar volumes are not the same, the total concentration

$$\overline{c} := \sum_{i=0}^{N} c_i \tag{3.1.2}$$

is not uniform. The set of positive concentrations $c_i, i \in [\![1, N]\!]$ such that c_0 is positive is denoted by

$$\mathcal{A} = \left\{ (c_1, ..., c_N) \in (0, +\infty)^N \middle| c_0 := \frac{1}{v_0} - \sum_{i=1}^{N} k_i c_i > 0 \right\}.$$

We also introduce the topological adherence of $\mathcal{A}$:

$$\bar{\mathcal{A}} = \left\{ (c_1, ..., c_N) \in [0, +\infty)^N \middle| c_0 := \frac{1}{v_0} - \sum_{i=1}^{N} k_i c_i \geq 0 \right\}.$$

For the sake of clarity, we will let $C = (c_1, ..., c_N) \in \mathcal{A}$ and often consider c_0 and $\overline{c}$ as functions of C thanks to (3.1.1) and (3.1.2) without clearly expressing the dependency. The dissolved species follow a conservation equation:

$$\partial_t c_i - \operatorname{div} D_i \mathcal{N}_i = 0, \qquad \mathcal{N}_i = c_i \nabla \left(h_i(C) + \widetilde{z}_i \Phi \right) \qquad \forall i \in [\![1, N]\!]. \tag{3.1.3}$$

where $\widetilde{z}_i = z_i - k_i z_0$ the reduced charge number and $D_i > 0$ the diffusion coefficient are parameters of the problem, while $h_i(C)$ the chemical potential depends on all the concentrations through:

$$h_i(C) = \log \frac{c_i}{\overline{c}} - k_i \log \frac{c_0}{\overline{c}} \qquad \forall i \in [\![1, N]\!]. \tag{3.1.4}$$

This system is supplemented with Poisson equation for the potential:

$$-\lambda^2 \Delta \Phi = c^{\text{dop}} + \sum_{i=0}^{N} z_i c_i. \tag{3.1.5}$$

To simplify the computations, we let $c^{\text{dop}} = \frac{z_0}{v_0} + \widetilde{c}^{\text{dop}}$ and see that:

$$c^{\text{dop}} + \sum_{i=0}^{N} z_i c_i = \widetilde{c}^{\text{dop}} + \sum_{i=1}^{N} \widetilde{z}_i c_i.$$

To avoid unnecessary complications of the notations, we will drop the tildas for the reduced molar charges as the real molar charges do not appear anymore. Moreover, to simplify the proofs, we will assume that the solvent carries no charge, hence $z_0 = 0$ and $\widetilde{c}^{\text{dop}} = 0$. Treatment of nonzero $\widetilde{c}^{\text{dop}}$ one can find in Chapter 2.

As in Chapter 2, we consider a Dirichlet boundary condition for the potential on a non-negligible part of the boundary $\Gamma_{\text{D}} \subset \partial\Omega$ and homogeneous Neumann boundary condition on $\Gamma_{\text{N}} = \partial\Omega \setminus \Gamma_{\text{D}}$:

$$\Phi = \Phi^D \qquad \text{on } (0,T) \times \Gamma_D, \qquad\qquad \nabla\Phi \cdot n = 0 \qquad \text{on } (0,T) \times \Gamma_N, \tag{3.1.6}$$

where Φ^D is assumed to be constant in time and in $H^1(\Omega) \cap L^\infty(\Omega)$.

The system is supplemented with the following no flux boundary conditions for the concentrations:

$$c_i \nabla \left(h_i(C) + z_i \Phi\right) \cdot n = 0 \qquad \text{on } (0,T) \times \partial\Omega, \text{ for all } i \in [\![1, N]\!], \tag{3.1.7}$$

and with an initial condition C^0 satisfying:

$$C^0 \in L^\infty(\Omega, \bar{\mathcal{A}}) \qquad \text{and} \qquad \int_\Omega c_i^0 > 0 \qquad \forall i \in [\![0, N]\!]. \tag{3.1.8}$$

3.1.2 Key properties of the continuous system

In this section, we attempt to exhibit the properties of a smooth enough solution (C, Φ) to the system (3.1.3)–(3.1.8) so that calculations are justified. The first property is the conservation of mass. In other words, thanks to (3.1.3), C satisfies for any $t \in [0,T], i \in [\![1, N]\!]$:

$$\int_\Omega c_i(0,x) = \int_\Omega c_i(t,x).$$

Moreover, we need the concentrations to be positive for (3.1.4) to have a sense. In the discrete setting, we will show that the concentrations belong to $\mathcal{A}$. In the continuous setting, it will be assumed. We hint that it might be possible to do it using the entropy method [92] and the flux formulation proposed in [70]. Indeed, another key property of the system is the dissipation of a free energy. In this case, the chemical free energy density $H(C)$ is defined as follows:

$$H(C) := \sum_{i=0}^{N} c_i \log\left(\frac{c_i}{\overline{c}}\right) = \sum_{i=0}^{N} c_i \log c_i - \overline{c}\log\overline{c}.$$

This function is convex, however, the addition of the term $-\overline{c}\log\overline{c}$ makes the proof quite intricate. This point is detailed in Appendix 3.A along with the proof of the following equations:

$$\partial_{c_i} H(c_1, ..., c_N) = h_i(C), \qquad \forall i \in [\![1, N]\!],\ C = (c_1, ..., c_N) \in \mathcal{A}, \tag{3.1.9}$$

$$\frac{-\log(N+1)}{v_0 \min k_i} \leq H(C) \leq 0 \qquad \forall C \in \mathcal{A}. \tag{3.1.10}$$

The total free energy is formed by the integral of the chemical free energy density and electrical terms:

$$E(C, \Phi) = \int_\Omega H(C) + \lambda^2 \frac{|\nabla\Phi|^2}{2} \mathrm{d}x - \lambda^2 \int_{\Gamma_D} \Phi_D \nabla\Phi \cdot n.$$

Proposition 3.1.1. *Let (C, Φ) be smooth solutions of (3.1.3)–(3.1.8) such that $C(t, x) \in \mathcal{A}$. For such solutions, E is a convex Lyapunov functional. Moreover, we have:*

$$\partial_t E + \int_\Omega \sum_{i=1}^{n} D_i c_i |\nabla h_i(C) + z_i \Phi|^2 = 0. \tag{3.1.11}$$

Proof. We have using chain rules and (3.1.9):

$$\partial_t \int_\Omega H(c_1, ..., c_N)\mathrm{d}x = \int_\Omega \sum_{i=1}^{N} h_i(C)\partial_t c_i \mathrm{d}x. \tag{3.1.12}$$

We also have using chain rules and integrating by part:

$$\partial_t \int_\Omega \frac{|\nabla\Phi|^2}{2}\mathrm{d}x = \int_{\partial\Omega} \Phi\partial_t\left(\nabla\Phi \cdot n\right) - \int_\Omega \Phi\partial_t \Delta\,\Phi \mathrm{d}x.$$

Notice that we have $\nabla\Phi \cdot n = 0$ on Γ_N and $\Phi = \Phi_D$ on Γ_D. Using equation (3.1.5),

we have:

$$\partial_t \lambda^2 \left(\int_\Omega \frac{|\nabla \Phi|^2}{2} \mathrm{d}x - \int_{\Gamma_D} \Phi_D \nabla \Phi \cdot n \right) = \int_\Omega \Phi \sum_{i=1}^N z_i \partial_t c_i \mathrm{d}x.$$

Using this equation and (3.1.12), we have:

$$\partial_t E = \sum_{i=1}^N \int_\Omega (h_i(C) + z_i \Phi) \partial_t c_i \mathrm{d}x.$$

Using now equation (3.1.3) and integration by parts, we have the desired equation (3.1.11). Due to the non-negativity of $D_i c_i$, E is a Lyapunov functional. Its convexity follows from the assumption $C \in \mathcal{A}$ (see Lemma 3.A.1). □

Finally, we introduce a notion of weak solution that relies on a reformulation of the fluxes:

$$\mathcal{N}_i = \nabla c_i - k_i c_i \nabla \log c_0 + (k_i - 1) c_i \nabla \log \overline{c} + z_i c_i \nabla \Phi,$$

and the space of H^1 functions satisfying the Dirichlet boundary conditions for the potential:

$$\mathcal{H}_{\Gamma^D} = \{ f \in H^1(\Omega), f_{|_{\Gamma_D}} = 0 \} \qquad \text{and} \qquad Q_T = (0,T) \times \Omega.$$

More precisely:

Definition 5. *A couple (C, Φ) is a weak solution of (3.1.3)–(3.1.8) if*

- *$C \in L^\infty((Q_T; \overline{\mathcal{A}})$ with $\log(c_0) \in L^2((0,T); H^1(\Omega))$;*
- *$\Phi - \Phi^D \in L^\infty((0,T), \mathcal{H}_{\Gamma^D})$;*
- *for all $\varphi \in C_c^\infty([0,T) \times \overline{\Omega})^N$, $i \in [\![1, N]\!]$*

$$\begin{aligned} &\iint_{Q_T} c_i \partial_t \varphi_i \mathrm{d}x\mathrm{d}t + \int_\Omega c_i^0 \varphi_i(0,x) \mathrm{d}x \\ &\quad - \iint_{Q_T} \big(\nabla c_i + c_i \nabla \big(-k_i \log c_0 + (k_i - 1) \log \overline{c} + z_i \Phi \big) \big) \cdot \nabla \varphi_i \mathrm{d}x\mathrm{d}t = 0; \end{aligned} \tag{3.1.13}$$

- *for all $\psi \in \mathcal{H}_{\Gamma^D}$ and almost all $t \in (0,T)$,*

$$\lambda^2 \int_\Omega \nabla \Phi(t,x) \cdot \nabla \psi(x) \mathrm{d}x = \int_\Omega \psi(x) \sum_{i=1}^N z_i c_i(t,x) \mathrm{d}x. \tag{3.1.14}$$

3.1.3 Positioning and outline

The structure of cross-diffusion systems challenges the maximum principle-based methods. In this paper we aim to discretize the system (3.1.3)–(3.1.8). For $N = 1$ this system is a nonlinear drift-diffusion problem and several discretizations have been proposed in Chapter 2. We focus on the extension of these schemes to the more general setting with $N > 1$ while adapting the proofs to tackle the challenges introduced by cross-diffusion.

More precisely, in Section 3.2, the two point flux based finite volume discretization with two variants of the flux approximation is introduced. The main theorems about the existence of discrete solutions and the convergence of approximate solutions are stated. Existence, free energy decay, and positivity of concentrations are proven in Section 3.3, whereas the convergence is proven in Section 3.4. Several 1D and 2D numerical examples showcasing the proven properties of the discretization scheme are discussed in Section 3.5.

3.2 Discretization and main Theorems

In this section, we propose two discretizations of (3.1.3)–(3.1.8) and discrete counterparts of the continuous properties. First, in Section 3.2.1, we state the requirements on the mesh and fix some notations. Then in Section 3.2.2, we describe the common setting for the two schemes to be studied in this paper. These schemes, presented in Section 3.2.3, rely on so-called two-point flux approximations of different formulations of $\mathcal{N}_i$. Then in Section 3.2.4, we state our two main results. The first one, namely Theorem 3.2.1, focuses on the existence of a solution to the nonlinear system corresponding to the schemes for a given mesh, and the dissipation of the energy at the discrete level. More precisely, one establishes that all the studied schemes satisfy a discrete counterpart to Proposition 3.1.1. Our second main result, namely Theorem 3.2.2, is devoted to the convergence of the schemes as the time step and the mesh size tend to 0.

3.2.1 Discretization of $(0, T) \times \Omega$

In this paper, we perform a parallel study of two numerical schemes based on two-point flux approximation (TPFA) finite volume schemes. As explained in [61, 67], this approach appears to be very efficient for isotropic continuous problems when one has the freedom to choose a suitable mesh fulfilling the so-called orthogonality condition [89, 68]. We recall here the definition of such a mesh, which is illustrated in Figure 3.1.

Definition 6. *An admissible mesh of Ω is a triplet $\left(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}}\right)$ such that the following conditions are fulfilled.*

(i) The set $\mathcal{T}$ is finite and each control volume (or cell) $K \in \mathcal{T}$ is non-empty, open, polyhedral, and convex. We assume that

$$K \cap L = \emptyset \quad \text{if } K, L \in \mathcal{T} \text{ with } K \neq L, \qquad \text{while} \quad \bigcup_{K\in\mathcal{T}} \overline{K} = \overline{\Omega}.$$

(ii) Each face $\sigma \in \mathcal{E}$ is closed and is contained in a hyperplane of $\mathbb{R}^d$, with positive $(d-1)$-dimensional Hausdorff (or Lebesgue) measure denoted by $m_\sigma = \mathcal{H}^{d-1}(\sigma) > 0$. We assume that $\mathcal{H}^{d-1}(\sigma \cap \sigma') = 0$ for $\sigma, \sigma' \in \mathcal{E}$ unless $\sigma' = \sigma$. For all $K \in \mathcal{T}$, we assume that there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \sigma$. Moreover, we suppose that $\bigcup_{K\in\mathcal{T}} \mathcal{E}_K = \mathcal{E}$. Given two distinct control volumes $K, L \in \mathcal{T}$, the intersection $\overline{K} \cap \overline{L}$ either reduces to a single face $\sigma \in \mathcal{E}$ denoted by $K|L$, or its $(d-1)$-dimensional Hausdorff measure is 0.

(iii) The cell centers $(x_K)_{K\in\mathcal{T}}$ belong to their cell: $x_K \in K$, and are such that, if $K, L \in \mathcal{T}$ share a face $K|L$, then the vector $x_L - x_K$ is orthogonal to $K|L$.

(iv) For the boundary faces $\sigma \subset \partial\Omega$, we assume that either $\sigma \subset \overline{\Gamma}_D$ or $\sigma \subset \overline{\Gamma}_N$. For $\sigma \subset \partial\Omega$ with $\sigma \in \mathcal{E}_K$ for some $K \in \mathcal{T}$, we assume additionally that there exists $x_\sigma \in \sigma$ such that $x_\sigma - x_K$ is orthogonal to σ.

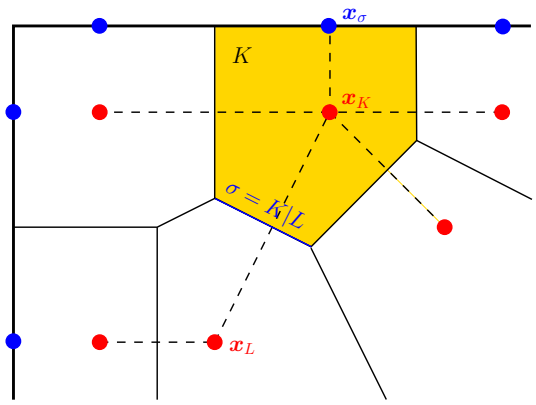


Figure 3.1 – Illustration of an admissible mesh as in Definition 6.

We denote by m_K the d-dimensional Lebesgue measure of the control volume K. The set of the faces is partitioned into two subsets: the set $\mathcal{E}_{\rm int}$ of the interior faces defined by $\mathcal{E}_{\rm int} = \{\sigma \in \mathcal{E} \mid \sigma = K|L \text{ for some } K, L \in \mathcal{T}\}$, and the set $\mathcal{E}_{\rm ext}$ of the exterior faces defined by $\mathcal{E}_{\rm ext} = \{\sigma \in \mathcal{E} \mid \sigma \subset \partial\Omega\}$, which can also be partitioned into $\mathcal{E}^{\rm D} = \{\sigma \subset \overline{\Gamma}_{\rm D}\}$ and $\mathcal{E}^{\rm N} = \{\sigma \subset \overline{\Gamma}_{\rm N}\}$.

Given $\sigma \in \mathcal{E}$, we let

$$d_\sigma = \begin{cases} |x_K - x_L| & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ |x_K - x_\sigma| & \text{if } \sigma \in \mathcal{E}_{\rm ext}, \end{cases} \qquad \text{and} \quad \tau_\sigma = \frac{m_\sigma}{d_\sigma}.$$

We finally introduce the size $h_\mathcal{T}$ and the regularity $\zeta_\mathcal{T}$ (which is assumed to be positive) of a discretization $(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}})$ of Ω by setting

$$h_\mathcal{T} = \max_{K\in\mathcal{T}} \operatorname{diam}(K), \qquad \zeta_\mathcal{T} = \min_{K\in\mathcal{T}} \min_{\sigma\in\mathcal{E}_K} \frac{\mathrm{d}(x_K, \sigma)}{d_\sigma}.$$

Concerning the time discretization of $(0, T)$, we consider an increasing finite family of times $0 = t_0 < t_1 < \ldots, < t_{N_T} = T$. We denote by $\Delta t_n = t_n - t_{n-1}$ for $1 \le n \le N_T$, by $\boldsymbol{\Delta t} = (\Delta t_n)_{1\le n\le N_T}$, and by $h_{\boldsymbol{\Delta t}} = \max_{1\le n\le N_T} \Delta t_n$. We will use boldface notations for vectors whose number of components is dependent on the mesh while keeping the uppercase notation $\boldsymbol{C}$ when we also consider different species.

3.2.2 A common setting for the Finite Volume schemes

The initial data C^0 which belongs to $L^\infty(\Omega, \bar{\mathcal{A}})$ thanks to (3.1.8) is discretized into $(C^0_K)_{K\in\mathcal{T}} \in \bar{\mathcal{A}}^\mathcal{T}$ by setting

$$c^0_{K,i} = \fint_K c^0_i(x)\mathrm{d}x \qquad \forall K \in \mathcal{T}, i \in [\![1, N]\!]. \tag{3.2.1}$$

Notice that previous equation also holds for $i = 0$ and that this discretization satisfies:

$$\sum_{K\in\mathcal{T}} m_K c^0_{K,i} = \int_\Omega c^0_i(x)\mathrm{d}x > 0 \qquad i \in [\![0, N]\!] \qquad \text{and} \qquad C^0_K \in \bar{\mathcal{A}} \qquad \forall K \in \mathcal{T}. \tag{3.2.2}$$

Assume that $\boldsymbol{C}^{n-1} = \left(c^{n-1}_{K,i}\right)_{K\in\mathcal{T}, i\in[\![0,N]\!]}$ is given for some $n > 0$, then we have to define how to compute $(\boldsymbol{C}^n, \boldsymbol{\Phi}^n) = (C^n_K, \Phi^n_K)_{K\in\mathcal{T}}$. First, we introduce some notations. For all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_K$, we define the mirror values $C^n_{K\sigma}$ and $\Phi^n_{K\sigma}$ of C^n_K and Φ^n_K respectively across σ by setting

$$C^n_{K\sigma} = \begin{cases} C^n_L & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ C^n_K & \text{if } \sigma \in \mathcal{E}_{\rm ext}, \end{cases} \qquad \Phi^n_{K\sigma} = \begin{cases} \Phi^n_L & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ \Phi^n_K & \text{if } \sigma \in \mathcal{E}^N, \\ \Phi^D_\sigma = \fint_\sigma \Phi^D \mathrm{d}\boldsymbol{\gamma} & \text{if } \sigma \in \mathcal{E}^D. \end{cases} \tag{3.2.3}$$

Given $\boldsymbol{u} = (u_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, we define the oriented and absolute jumps of $\boldsymbol{u}$ across any edge by

$$D_{K\sigma}\boldsymbol{u} = u_{K\sigma} - u_K, \quad D_\sigma \boldsymbol{u} = |D_{K\sigma}\boldsymbol{u}|, \qquad \forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_K.$$

We may now use these operators to describe our scheme. The potential is approximated using the classic TPFA scheme for the Poisson equation:

$$-\lambda^2 \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D_{K\sigma}\boldsymbol{\Phi}^n = m_K \sum_{i=1}^N z_i c_{K,i}^n, \qquad \forall K \in \mathcal{T}. \tag{3.2.4a}$$

The conservation equation is approximated using a backward-Euler scheme in time:

$$m_K \frac{c_{K,i}^n - c_{K,i}^{n-1}}{\Delta t_n} + \sum_{\sigma\in\mathcal{E}_K} F_{K\sigma,i}^n = 0, \qquad \forall K \in \mathcal{T}, i \in [\![1, N]\!], \tag{3.2.4b}$$

where $F_{K\sigma,i}^n$ should be a conservative and consistent approximation of the integral $-\frac{D_i}{\Delta t_n}\int_{t_{n-1}}^{t_n}\int_\sigma \mathcal{N}_i \cdot n_{K\sigma}$ ($n_{K\sigma}$ denotes the normal to σ outward K). Finally, the concentration of the solvent is computed using a discrete version of the volume filling constraint:

$$c_{K,0}^n = \frac{1}{v_0} - \sum_{i=1}^N k_i c_{K,i}^n, \qquad \forall K \in \mathcal{T}. \tag{3.2.4c}$$

It remains to define the numerical fluxes $F_{K\sigma,i}^n$. Two possible choices are given in the next section.

3.2.3 Numerical fluxes for the conservation equations

To close the system (3.2.4), we have to define the numerical fluxes $F_{K\sigma,i}^n$. As we intend to use two point flux approximations, they should be of the form:

$$F_{K\sigma,i}^n = \begin{cases} 0 & \text{if } \sigma \in \mathcal{E}_{\text{ext}} \\ \tau_\sigma D_i \mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) & \text{if } \sigma = K|L \in \mathcal{E}_{\text{int}} \end{cases} \tag{3.2.5}$$

For the sake of readability, we have chosen to define the flux functions $\mathcal{F}_i$ for unitary D_i. Thus this constant should rarely appear in the functional inequalities of the following sections. To preserve the conservation of mass, all the flux functions $\mathcal{F}_i$ defined afterward satisfy an anti-symmetry property:

$$\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) = -\mathcal{F}_i(C_L, C_K, \Phi_L, \Phi_K) \qquad \forall C_K, C_L \in \mathcal{A}, \Phi_K, \Phi_L \in \mathbb{R}, \tag{3.2.6}$$

so that the fluxes are locally conservative, i.e.:

$$F_{K,\sigma} + F_{L,\sigma} = 0 \qquad \forall \sigma = K|L \in \mathcal{E}_{\text{int}}.$$

3.2.3.a The centered flux

The first numerical flux we consider is based on the original expression of the flux (3.1.3):

$$\mathcal{N}_i = D_i c_i \nabla \Big(h_i(c) + z_i \Phi \Big).$$

The gradient and edge concentration are independently discretized :

$$\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) = \frac{c_{K,i} + c_{L,i}}{2} \left(h_i(C_K) - h_i(C_L) \right) + z_i (\Phi_K - \Phi_L) \Big). \tag{C}$$

This flux is a straightforward generalization of the eponymous flux presented in Chapter 2. As such it is also similar to the fluxes introduced in [36, 33, 28, 39, 32].

3.2.3.b The Sedan flux

The other flux under study is also a generalization of the Sedan flux presented in Chapter 2. It originates from and is named after the SEDAN III semiconductor device simulation code [126] and is used to handle the case of degenerated semiconductors in semiconductor device simulators, see [123, 120]. The scheme relies on the introduction of the excess chemical potential

$$\nu_i(C) := h_i(C) - \log(c_i) = -\log(\bar{c}) - k_i \log \frac{c_0}{\bar{c}}.$$

This excess potential characterizes the non-ideality of the electrolyte leading to the following equivalent continuous flux formulation:

$$\mathcal{N}_i = D_i \Big[\nabla c_i + c_i \nabla \left(z_i \Phi + \nu_i(C) \right) \Big].$$

The Scharfetter-Gummel-inspired discretization [116] of this expression of the flux leads to the so-called Sedan flux:

$$\begin{aligned} \mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) = B\big(z_i \Phi_L + \nu_i(C_L) - z_i \Phi_K - \nu_i(C_K)\big) c_{K,i} \\ - B\big(z_i \Phi_K + \nu_i(C_K) - z_i \Phi_L - \nu_i(C_L)\big) c_{L,i}, \end{aligned} \tag{S}$$

where $B(x) = \frac{x}{e^x - 1}$ for all $x \neq 0$ is the Bernoulli function. Notice that B can be extended by $B(0) = 1$ and is in C^∞.

Remark. In chapter 2 we studied two other schemes. One was based on the diffusion enhancement and discretization ideas originating from [15]. The extension of this so-called Bessemoulin-Chatard scheme to the multi-species case appears to be not feasible due to the intrinsic use of one-dimensional chain rules. The other scheme based on activity variables and the averaging of the inverse activity coefficient was introduced for the multi-species case in [70]. Numerical analysis of such a scheme is more intricate and would likely not be satisfactory as we were not able to prove convergence in chapter 2. Moreover, unless more sophisticated inverse activity coefficient averaging strategies are available, this scheme is considerably less accurate compared to all the others discussed in chapter 2.

3.2.4 Main theorems

We have proposed two schemes (3.2.4), (3.2.5) supplemented with either (C) or (S). Both schemes are nonlinear systems. Solutions to this nonlinear system should satisfy discrete equivalents of the properties listed in Section 3.1.2, namely conservation of mass and energy-dissipation. For the latter, we introduce the discrete energy functional $E_{\mathcal{T}}$ as a discrete counterpart of the continuous energy functional E. It is defined by:

$$E_{\mathcal{T}}(\boldsymbol{C}^n, \boldsymbol{\Phi}^n) = \sum_{K\in\mathcal{T}} m_K H(C_K^n) + \frac{\lambda^2}{2}\sum_{\sigma\in\mathcal{E}} \tau_\sigma \left(D_\sigma \boldsymbol{\Phi}^n\right)^2 - \lambda^2 \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D\cap\mathcal{E}_K} \tau_\sigma \Phi_\sigma^D D_{K\sigma}\boldsymbol{\Phi}^n. \tag{3.2.7}$$

The first theorem proven in this paper focuses on the existence of discrete solutions for a given mesh, and the preservation of the physical bounds: non negative concentrations, and the properties of Section 3.1.2.

Theorem 3.2.1. *Let $(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}})$ be an admissible mesh and let $\boldsymbol{C}^0$ be defined by (3.2.1). Then, for all $1 \le n \le N_T$, the nonlinear system of equations (3.2.4), (3.2.5) supplemented with either (C) or (S) has a solution*

$$(\boldsymbol{C}^n, \boldsymbol{\Phi}^n) \in \mathcal{A}^{\mathcal{T}} \times \mathbb{R}^{\mathcal{T}}.$$

Moreover, the solution to the scheme satisfies, for all $1 \le n \le N_T$,

$$E_{\mathcal{T}}(\boldsymbol{C}^n, \boldsymbol{\Phi}^n) - E_{\mathcal{T}}(\boldsymbol{C}^{n-1}, \boldsymbol{\Phi}^{n-1}) \le \Delta t_n \sum_{i=1}^{N}\sum_{\sigma\in\mathcal{E}} F_{K\sigma,i}^n D_{K\sigma}(h_i(\boldsymbol{C}^n) + z_i\boldsymbol{\Phi}^n), \tag{3.2.8}$$

and

$$\sum_{K\in\mathcal{T}} c_{K,i} m_K = \int_\Omega c_i^0(x)\mathrm{d}x \qquad \forall i \in [\![0, N]\!]. \tag{3.2.9}$$

The proof of this theorem is the purpose of Section 3.3. Knowing a discrete solution to the scheme, $(\boldsymbol{C}^n, \boldsymbol{\Phi}^n)_{1\leq n\leq N}$, we can define an approximate solution $(C_{\mathcal{T},\boldsymbol{\Delta t}}, \Phi_{\mathcal{T},\boldsymbol{\Delta t}})$. It is the piecewise constant function defined almost everywhere by

$$C_{\mathcal{T},\boldsymbol{\Delta t}}(t,x) = C_K^n, \quad \Phi_{\mathcal{T},\boldsymbol{\Delta t}}(t,x) = \Phi_K^n \quad \text{if } (t,x) \in (t_{n-1}, t_n] \times K.$$

This definition will be developed in Section 3.4 and supplemented by other reconstruction operators.

Using this existence result, we let $(\mathcal{T}_m, \mathcal{E}_m, (x_K)_{K\in\mathcal{T}_m})_{m\geq 1}$, $(\boldsymbol{C}_m^n, \boldsymbol{\Phi}_m^n) \in \mathcal{A}^{\mathcal{T}} \times \mathbb{R}^{\mathcal{T}}$, be a sequence of admissible meshes in the sense of Definition 6 and associated approximate solution. We assume that $h_{\mathcal{T}_m}, h_{\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} 0$ while the mesh regularity remains bounded, i.e., $\zeta_{\mathcal{T}_m} \geq \zeta^\star$ for some $\zeta^\star > 0$ not depending on m. A natural question is the convergence of $(C_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}, \Phi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m})$ towards a weak solution to the continuous problem. The convergence result is stated in Theorem 3.2.2 which will be proved in Section 3.4.

Theorem 3.2.2. *For the two schemes under study, a sequence of approximate solutions $(\boldsymbol{C}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}, \boldsymbol{\Phi}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m})_{m\geq 1}$ satisfies, up to a subsequence:*

$$C_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} C \quad \text{in } L^2(Q_T)^{N+1}, \qquad \Phi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \Phi \quad \text{in } L^2(Q_T). \tag{3.2.10}$$

Moreover if $\inf_{Q_T} c_0 > 0$, (C, Φ) is a weak solution of (3.1.3)–(3.1.8) in the sense of Definition 5.

3.3 Fixed Mesh analysis

In this section, we intend to prove Theorem 3.2.1. To this end, we will use a topological degree argument in Section 3.3.3. This topological degree relies on properties of the fluxes and *a priori* estimates detailed respectively in the following section and in Section 3.3.2. The methodology of this proof is very similar to the one done in chapter 2. The key changes and improvements are concentrated in Proposition 3.3.2, Lemmas 3.3.2 and 3.3.5.

3.3.1 Analysis of numerical flux based functions

In this section, we introduce several functions derived from $\mathcal{F}_i$. As in Chapter 2, the first functions of interest models the free energy dissipation for each species $i \in [\![1, N]\!]$:

$$\mathcal{D}_i(C_K, C_L, \Phi_K, \Phi_L) := -\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L)\left(h_i(C_K) + z_i\Phi_K - h_i(C_L) - z_i\Phi_L\right).$$

We also introduce the local free energy dissipation $\mathcal{D} := \sum_{i=1}^{N} \mathcal{D}_i$. In addition to this function, we can define a reconstruction of the concentration at the interfaces. This is the purpose of the following lemma:

Lemma 3.3.1. *For a flux $\mathcal{F}_i$ defined either by (C) or (S), the corresponding face concentration functions defined by*

$$\mathcal{C}_i(C_K, C_L, \Phi_K, \Phi_L) = \frac{\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L)}{h_i(C_K) + z_i\Phi_K - h_i(C_L) - z_i\Phi_L} \qquad \forall i \in [\![1, N]\!] \quad (3.3.1)$$

if $h_i(C_K)+z_i\Phi_K-h_i(C_L)-z_i\Phi_L \neq 0$ can be extended by continuity on $\mathcal{A}\times\mathcal{A}\times\mathbb{R}\times\mathbb{R}$. Moreover, for all $(C_K, C_L, \Phi_K, \Phi_L) \in \mathcal{A} \times \mathcal{A} \times \mathbb{R} \times \mathbb{R}$, and for all $i \in [\![1, N]\!]$:

$$\min(c_{K,i}, c_{L,i}) \leq \mathcal{C}_i(C_K, C_L, \Phi_K, \Phi_L) \leq \max(c_{K,i}, c_{L,i}). \quad (3.3.2)$$

Proof. The proof of the extension by continuity and the average property (3.3.2) is highly similar to [29, Lemma 3.1]. For the centered scheme defined by (C), we have by definition:

$$\mathcal{C}_i(C_K, C_L, \Phi_K, \Phi_L) = \frac{c_{K,i} + c_{L,i}}{2},$$

hence the extension by continuity and equation (3.3.2).

For the Sedan scheme, defined by (S), we introduce $x_i = \log(c_{K,i}/c_{L,i})$ and $y_i = z_i\Phi_L + \nu_i(C_L) - z_i\Phi_K - \nu_i(C_K)$ and notice that:

$$\begin{aligned} h_i(C_K) + z_i\Phi_K - h_i(C_L) - z_i\Phi_L &= x_i - y_i, \\ \mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) &= B(y_i)c_{K,i} - B(-y_i)c_{L,i}. \end{aligned} \quad (3.3.3)$$

Using the following property of the Bernoulli function:

$$B(\log(a) - \log(b))a - B(\log(b) - \log(a))b = 0, \qquad \forall (a, b) \in (0, +\infty)^2,$$

we have:

$$\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) = (B(y_i) - B(x_i))c_{K,i} - (B(-y_i) - B(-x_i))c_{L,i}. \quad (3.3.4)$$

Finally using (3.3.3) and the differentiability of B, we have the desired extension on $\mathcal{A} \times \mathcal{A} \times \mathbb{R} \times \mathbb{R}$. We also have equation (3.3.2) thanks to the monotony of B and the relation $B(x) - B(-x) = -x$ for all $x \in \mathbb{R}$. □

Thanks to this lemma, $\mathcal{D}_i$ rewrites:

$$\mathcal{D}_i(C_K, C_L, \Phi_K, \Phi_L) = \mathcal{C}_i(C_K, C_L, \Phi_K, \Phi_L)\Big(h_i(C_K) + z_i\Phi_K - h_i(C_L) - z_i\Phi_L\Big)^2. \quad (3.3.5)$$

This new formulation along with (3.3.2) grants the non-negativity of $\mathcal{D}_i$ and $\mathcal{D}$. The following lemma gives more detailed information on the behavior of $\mathcal{D}$:

Lemma 3.3.2. *Let for $\delta, \epsilon, M, c > 0$, $i \in [\![1, N]\!]$:*

$$\begin{aligned}\Psi_{\delta,\epsilon,M,i}(c) &:= \inf_{\substack{(C_K,C_L)\in\mathcal{A}^2,\ (\Phi_K,\Phi_L)\in[-M,M]^2 \\ c_{K,0},c_{L,0}>\epsilon,\ c_{K,i}\geq\min(\delta,\frac{0.5}{k_i v_0}),\ c_{L,i}<c}} \mathcal{D}_i(C_K, C_L, \Phi_K, \Phi_L), \\ \Upsilon_{\delta,M}(c) &:= \inf_{\substack{(C_K,C_L)\in\mathcal{A}^2,\ (\Phi_K,\Phi_L)\in[-M,M]^2 \\ c_{K,0}\geq\min(\delta,\frac{0.5}{v_0}),\ c_{L,0}<c}} \mathcal{D}(C_K, C_L, \Phi_K, \Phi_L).\end{aligned} \tag{3.3.6}$$

We have, for all $\delta, \epsilon, M > 0$:

$$\lim_{c\to 0^+} \Upsilon_{\delta,M}(c) = +\infty \qquad \textit{and} \qquad \lim_{c\to 0^+} \Psi_{\delta,\epsilon,M,i}(c) = +\infty \qquad \forall i \in [\![1, N]\!].$$

As the proof of this lemma is purely technical it has been relegated to appendix 3.B.

3.3.2 *A priori* estimates

In this section, we intend to establish uniform *a priori* estimates on the concentration and the potential, in order to prove the existence of solutions that satisfies the properties of Theorem 3.2.1.

We assume that we dispose of $(\boldsymbol{C}^n, \boldsymbol{\Phi}^n)_{n\in[\![0,N_{\max}]\!]}$ solution of (3.2.1), (3.2.4), (3.2.5) supplemented with either (C) or (S) in $\overline{\mathcal{A}}^{\mathcal{T}} \times \mathbb{R}^{\mathcal{T}}$. Where $\overline{\mathcal{A}}$, the adherence of $\mathcal{A}$ is the set of non-negative concentrations $c_0, ...c_N$ satisfying the volume filling constraint. The first *a priori* estimate is the conservation of mass (3.2.9):

Lemma 3.3.3. *For all n in $[\![0, N_{max}]\!]$, i in $[\![0, N]\!]$ we have:*

$$\sum_{K\in\mathcal{T}} m_K c_{K,i}^n = \int_\Omega c_i^0(x)\mathrm{d}x.$$

The proof is straightforward and classical thanks to the local conservativity of the fluxes, the no flux boundary conditions, and the discretization choice for $\boldsymbol{C}^0$.

We can also build a discrete equivalent to Theorem 3.1.1 using $E_{\mathcal{T}}$ defined in (3.2.7) and the dissipation function $\mathcal{D}_i$. This is the purpose of the following proposition:

Proposition 3.3.1. *For all n in $[\![0, N_{max}]\!]$, we have*

$$E_{\mathcal{T}}(\boldsymbol{C}^n, \boldsymbol{\Phi}^n) - E_{\mathcal{T}}(\boldsymbol{C}^{n-1}, \boldsymbol{\Phi}^{n-1}) \leq -\Delta t_n \sum_i D_i \sum_{\sigma=K|L\in\mathcal{E}_{int}} \tau_\sigma \mathcal{D}_i(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n). \tag{3.3.7}$$

Remark 3.3.1. Thanks to (3.2.5) and the definition of $\mathcal{D}_i$, (3.3.7) and (3.2.8) are equivalents.

Proof. The proof is fairly classical once noticed that thanks to Lemma 3.A.1, H is convex (thus $E_{\mathcal{T}}$ too). The inequality $f(a) - f(b) \leq f'(a)(a-b)$ yields:

$$\begin{aligned} E_{\mathcal{T}}(\boldsymbol{C}^n, \boldsymbol{\Phi}^n) - E_{\mathcal{T}}(\boldsymbol{C}^{n-1}, \boldsymbol{\Phi}^{n-1}) \leq \sum_{K\in\mathcal{T}} \sum_{i=1}^N m_K (c_{K,i}^n - c_{K,i}^{n-1}) h_i(C_K^n) + \\ \lambda^2 \sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\boldsymbol{\Phi}^n D_{K\sigma}(\boldsymbol{\Phi}^n - \boldsymbol{\Phi}^{n-1}) - \lambda^2 \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D\cap\mathcal{E}_K} \tau_\sigma \Phi_\sigma^D D_{K\sigma}(\boldsymbol{\Phi}^n - \boldsymbol{\Phi}^{n-1}). \end{aligned} \tag{3.3.8}$$

Notice that the left-hand side is the term of interest, we will then focus on the reformulation of the right-hand side. We multiply equation (3.2.4b) by $h_i(C_K) + z_i\Phi_K$ and we sum over the cells and species in order to get the following three-terms formula:

$$\begin{aligned} \underbrace{\sum_{K\in\mathcal{T}} \sum_{i=1}^N m_K \frac{c_{K,i}^n - c_{K,i}^{n-1}}{\Delta t_n} h_i(C_K^n)}_{T_{\text{chem}}} + \underbrace{\sum_{K\in\mathcal{T}} \Phi_K \sum_{i=1}^N m_K z_i \frac{c_{K,i}^n - c_{K,i}^{n-1}}{\Delta t_n}}_{T_{\text{el}}} \\ + \underbrace{\sum_{K\in\mathcal{T}} \sum_{i=1}^N \left(\sum_{\sigma\in\mathcal{E}_K} F_{K\sigma,i}^n \right) (h_i(C_K) + z_i\Phi_K)}_{T_{\text{diss}}} = 0. \end{aligned} \tag{3.3.9}$$

The term concerning the chemical energy, $\Delta t_n T_{\text{chem}}$, appears directly in (3.3.8), thus we focus on T_{el}. Using equation (3.2.4a), we have:

$$\Delta t_n T_{\text{el}} = \lambda^2 \sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\boldsymbol{\Phi}^n D_{K\sigma}(\boldsymbol{\Phi}^n - \boldsymbol{\Phi}^{n-1}) - \lambda^2 \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D\cap\mathcal{E}_K} \tau_\sigma \Phi_\sigma^D D_{K\sigma}(\boldsymbol{\Phi}^n - \boldsymbol{\Phi}^{n-1}), \tag{3.3.10}$$

which is the second line of equation (3.3.8). For T_{diss}, an integration by parts

yields:

$$T_{\text{diss}} = -\sum_{i=1}^{N}\sum_{\sigma\in\mathcal{E}} F^n_{K\sigma,i} D_{K\sigma}(h_i(\boldsymbol{C}) + z_i\boldsymbol{\Phi}).$$

Using this equation and equations (3.3.10), (3.3.9) in (3.3.8), we have (3.2.8):

$$E_{\mathcal{T}}(\boldsymbol{C}^n, \boldsymbol{\Phi}^n) - E_{\mathcal{T}}(\boldsymbol{C}^{n-1}, \boldsymbol{\Phi}^{n-1}) \leq \Delta t_n \sum_{i=1}^{N}\sum_{\sigma\in\mathcal{E}} F^n_{K\sigma,i} D_{K\sigma}(h_i(\boldsymbol{C}^n) + z_i\boldsymbol{\Phi}^n),$$

which concludes the proof thanks to the preliminary remark. □

In the following lemma, we will show several bounds on the potential $\boldsymbol{\Phi}$ and then take advantage of them to get a bound on the free energy dissipation:

Lemma 3.3.4. *There exist M_Φ depending only on λ, Φ^D, Ω, $\frac{1}{v_0}$, $(k_1, \dots, k_N)$, $(z_1, \dots, z_N)$, and another constant M_* depending also on $\zeta_{\mathcal{T}}$ such that:*

$$\|\boldsymbol{\Phi}^n\|_\infty \leq M_\Phi, \qquad \forall 1 \leq n \leq N_{max}, \tag{3.3.11}$$

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma |D_\sigma \boldsymbol{\Phi}^n|^2 \leq M_*, \qquad \forall 1 \leq n \leq N_{max}, \tag{3.3.12}$$

$$\left| \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D\cap\mathcal{E}_K} \tau_\sigma \Phi^D_\sigma D_{K\sigma}\boldsymbol{\Phi}^n \right| \leq M_*, \qquad \forall 1 \leq n \leq N_{max}, \tag{3.3.13}$$

$$\sum_{n=1}^{N_{max}} \Delta t_n \sum_{i=1}^{N} D_i \sum_{\sigma=K|L\in\mathcal{E}_{int}} \tau_\sigma \mathcal{D}_i(C^n_K, C^n_L, \Phi^n_K, \Phi^n_L) \leq M_*. \tag{3.3.14}$$

Proof. The proof of (3.3.11) is a straightforward application of [29, Proposition A.1]. As the proof of (3.3.12) is detailed in [29, Lemma 3.6], we focus on the proof of (3.3.13), assuming (3.3.12).

Multiplying equation (3.2.4a) by Φ^n_K and summing over $K \in \mathcal{T}$ yields, using (3.2.3):

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma \left(D_\sigma \boldsymbol{\Phi}^n\right)^2 - \sum_{\sigma\in\mathcal{E}^D} \tau_\sigma \Phi^D_\sigma D_{K\sigma}\boldsymbol{\Phi}^n = \sum_{K\in\mathcal{T}} \Phi^n_K m_K \sum_{i=1}^{N} z_i c^n_{K,i}.$$

Using equation (3.3.12), (3.3.11), and $\boldsymbol{C}^n \in \overline{\mathcal{A}}^{\mathcal{T}}$, we have the desired result. The

last result is based on (3.3.7). Summing that equation, we have:

$$\sum_{n=1}^{N_{\max}} \Delta t_n \sum_{i=1}^{N} D_i \sum_{\sigma=K|L\in\mathcal{E}_{\text{int}}} \tau_\sigma \mathcal{D}_i(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) \leq E_{\mathcal{T}}(\boldsymbol{C}^0, \boldsymbol{\Phi}^0) - E_{\mathcal{T}}(\boldsymbol{C}^{N_{\max}}, \boldsymbol{\Phi}^{N_{\max}}) \tag{3.3.15}$$

We have thanks to equations (3.1.10), (3.3.12), and (3.3.13):

$$E_{\mathcal{T}}(\boldsymbol{C}^{N_{\max}}, \boldsymbol{\Phi}^{N_{\max}}) \geq -|\Omega| \frac{\log(N+1)}{v_0 \min k_i} - \lambda^2 M_* \qquad \text{and} \qquad E_{\mathcal{T}}(\boldsymbol{C}^0, \boldsymbol{\Phi}^0) \leq \frac{3}{2}\lambda^2 M_*,$$

so that (3.3.15) becomes:

$$\sum_{n=1}^{N_{\max}} \Delta t_n \sum_{i=1}^{N} D_i \sum_{\sigma=K|L\in\mathcal{E}_{\text{int}}} -\tau_\sigma \mathcal{D}_i(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) \leq \frac{5}{2}\lambda^2 M_* + |\Omega| \frac{\log(N+1)}{v_0 \min k_i}.$$

Hence the desired result, up to the choice of a bigger constant M_*. □

Finally, we use the free energy dissipation result (3.3.14), and the estimates on the free energy dissipation functional to improve the assumption $\boldsymbol{C}^n \in \overline{\mathcal{A}}^{\mathcal{T}}$.

Lemma 3.3.5. *There exist $\epsilon_0, \epsilon_1, ..., \epsilon_N$ positive, depending on, among other things, C^0 and decreasing with $\min \boldsymbol{\Delta t}$ and $\min_{\sigma\in\mathcal{E}} \tau_\sigma$ such that:*

$$c_{K,i}^n \geq \epsilon_i \qquad \forall K \in \mathcal{T}, n \in [\![1, N_{max}]\!], i \in [\![0, N]\!]$$

Proof. The proof follows the idea of [35, Lemma 3.10] (see also [36, Lemma 3.7], [29, Lemma 3.7]). We start with the proof for $i = 0$ and a fixed time step n using $\Upsilon_{\delta, M_\Phi}$, then treat the case of $i \in [\![1, N]\!]$ using $\Psi_{\delta,\epsilon_0,M_\Phi,i}$ and finally notice that no assumptions were made on n.

Thanks to assumption (3.1.8) on the initial concentrations, and Lemma 3.3.3, we dispose of $K \in \mathcal{T}$ such that:

$$c_{K,0}^n \geq \fint_\Omega c_0^0 \mathrm{d}x =: \delta_0 > 0$$

We let $\delta_1 = \Upsilon^{-1}_{\delta_0, M_\Phi}(\frac{M_*}{\min \boldsymbol{\Delta t} \min_{i\in[1,N]} D_i \min_{\sigma\in\mathcal{E}} \tau_\sigma})$ where M_* is as in Lemma 3.3.4. It is well defined thanks to the monotony of Υ and Lemma 3.3.2. Moreover, we have for every cell L sharing an edge with K:

$$c_{L,0}^n \geq \delta_1 > 0,$$

thanks to the positivity of $\mathcal{D}_i$ and equation (3.3.14). Similarly we recursively

define:

$$\delta_{l+1} = \Upsilon^{-1}_{\delta_l, M_\Phi}\left(\frac{M_*}{\min \boldsymbol{\Delta t} \min_{i\in[\![1,N]\!]} D_i \min_{\sigma\in\mathcal{E}} \tau_\sigma}\right) \qquad \forall l \in \mathbb{N}^*, \tag{3.3.16}$$

and notice that thanks to the connectivity of Ω there exist l such that, for all $L \in \mathcal{T}$:

$$c^n_{L,0} \geq \delta_l.$$

Hence a possible choice for ϵ_0. As explained above, the proof is exactly the same for $i \in [\![1, N]\!]$, with the use of $\Psi_{\delta,\epsilon_0,M_\Phi,i}$ instead of Υ_{δ,M_Φ} in equation (3.3.16) and does not depend on the time step $n \geq 1$. □

3.3.3 Existence of solutions

Using the estimates of the previous section we can establish the existence of a solution to our numerical scheme. Thanks to Proposition 3.3.1 and Lemmas 3.3.3 and 3.3.5, this will conclude the proof of Theorem 3.2.1.

Proposition 3.3.2. *Let $\boldsymbol{C}^0$ be defined by (3.2.1). Then, for all $1 \leq n \leq N_T$, the nonlinear system of equations (3.2.4), (3.2.5) supplemented with either (C) or (S) has a solution $(\boldsymbol{C}^n, \boldsymbol{\Phi}^n) \in \mathcal{A}^\mathcal{T} \times \mathbb{R}^\mathcal{T}$.*

Proof. As in [29, Proposition 3.8], we use induction and a topological degree argument to transform continuously the non-linear system (3.2.4), (3.2.5) to a linear one. However, the path presented in Chapter 2 is no longer valid as we do not have a monotony property on h_i. The homotopy follows 3 steps. The first one is sketched in Appendix 3.C, the second one changes the discretization while maintaining k_i, D_i to 1 and the potential to zero. The last step corresponds to the activation of the potential and the remaining nonlinearities.

Following these ideas, we follows the zeros of a homotopy $\mathcal{H}$:

$$\mathcal{H} : \begin{cases} [0,3] \times \mathcal{A}^\mathcal{T} \times \mathbb{R}^\mathcal{T} \to (\mathbb{R}^N)^\mathcal{T} \times \mathbb{R}^\mathcal{T} \\ (\alpha, \boldsymbol{C}, \boldsymbol{\Phi}) \mapsto \mathcal{H}(\alpha, \boldsymbol{C}, \boldsymbol{\Phi}), \end{cases}$$

which should be our scheme for $\alpha = 3$ and the heat equation for $\alpha = 0$.

At every step, $\boldsymbol{c}_0$ is eliminated thanks to (3.2.4c).

Step 1: implementation of the solvent effects using an *ad hoc* scheme.

For $\alpha \in [0,1]$, $\mathcal{H} = 0$ means that for all $K \in \mathcal{T}$, $i \in [\![1, N]\!]$:

$$\frac{c_{K,i} - c_{K,i}^{n-1}}{\Delta t_n} m_K + \sum_{\sigma = K|L \in \mathcal{E}_{\text{int}}} \tau_\sigma \frac{c_{K,i} - c_{L,i}}{\log(c_{K,i}/c_{L,i})} \Big(\log(c_{L,i}) - \log(c_{K,i}) + \alpha \big(\log(c_{K,0}) - \log(c_{L,0}) \big) \Big) = 0,$$

$$-\lambda^2 \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \mathbf{\Phi} = 0,$$

where Φ_α^D is set to zero. As expressed in Lemma 3.C.1 we dispose of ϵ_1 such that the zeros of $\mathcal{H}$ have a concentration that is bounded away from zero by ϵ_1.

Step 2: change of scheme without potential and for identical species.

We change the discretization of $c_i \nabla \log(c_i/c_0)$. For $\alpha \in [1,2]$, $\mathcal{H} = 0$ rewrites:

$$\frac{c_{K,i} - c_{K,i}^{n-1}}{\Delta t_n} m_K + (2-\alpha) \sum_{\sigma = K|L \in \mathcal{E}_{K,\text{int}}} \tau_\sigma \frac{c_{K,i} - c_{L,i}}{\log(c_{K,i}/c_{L,i})} \Big(\log(c_{L,i}) - \log(c_{K,i}) + \big(\log(c_{K,0}) - \log(c_{L,0}) \big) \Big)$$

$$+ (\alpha - 1) \sum_{\sigma = K|L \in \mathcal{E}_{K,\text{int}}} \tau_\sigma \mathcal{F}_i(C_K, C_L, 0, 0) = 0,$$

$$-\lambda^2 \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \mathbf{\Phi} = 0$$

where Φ_α^D is again set to zero and $k_{i,\alpha}$ to 1. Here again we dispose of ϵ_2 such that the zeros of $\mathcal{H}$ have a concentration that is bounded away from zero by ϵ_2.

Step 3: activation of the potential and the difference between the species

For $\alpha \in [2,3]$, $\mathcal{H} = 0$ means:

$$\frac{c_{K,i} - c_{K,i}^{n-1}}{\Delta t_n} m_K + \big(3 - \alpha + (\alpha - 2) D_i \big) \sum_{\sigma = K|L \in \mathcal{E}_{K,\text{int}}} \tau_\sigma \mathcal{F}_i(C_K, C_L, (3-\alpha)\Phi_K, (3-\alpha)\Phi_L) = 0$$

$$-\lambda^2 \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \mathbf{\Phi} = m_K \sum_{i=1}^{N} (3-\alpha) z_i c_i,$$

where Φ_α^D is set to $(3-\alpha)\Phi_D$ and $k_{i,\alpha}$ to $3 - \alpha + (\alpha - 2)k_i$. Thanks to Lemma 3.3.5, we dispose of ϵ_3 such that the zeros of $\mathcal{H}$ have a concentration that is bounded away from zero by ϵ_3.

Conclusion

Using a topological degree argument [98, 56], we can derive the existence of solutions for $\alpha = 3$ from the non zero topological degree at $\alpha = 0$ and the uniform bounds on the concentration: $\min(\epsilon_1, \epsilon_2, \epsilon_3)$ and the potential: M_Φ. □

3.4 Convergence

In this section we prove Theorem 3.2.2, which states the convergence of our schemes towards a weak solution. We consider a sequence $\left(\mathcal{T}_m, \mathcal{E}_m, (x_K)_{K\in\mathcal{T}_m}\right)_{m\geq 1}$ of admissible meshes with $h_{\mathcal{T}_m}, h_{\boldsymbol{\Delta t}_m}$ tending to 0 as m tends to $+\infty$, while the regularity $\zeta_{\mathcal{T}_m}$ remains uniformly bounded from below by a positive constant $\zeta^\star$.

Thanks to Theorem 3.2.1, we have a family of discrete solutions $(\boldsymbol{C}_m, \boldsymbol{\Phi}_m)_m$. We will first propose different reconstructions of approximate solutions in Section 3.4.1, then we show several compactness properties in Section 3.4.2 in order to obtain the convergence of a subsequence of approximated solutions. Section 3.4.3 is then devoted to the identification of the limit as a weak solution.

To enlighten the notations, we will remove the subscript m as soon as it is not necessary for understanding.

3.4.1 Reconstruction operators

In order to carry out the analysis of convergence, we introduce some reconstruction operators following the methodology proposed in [64].

The operators $\pi_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}} \to L^\infty(\Omega)$ and $\pi_{\mathcal{T},\boldsymbol{\Delta t}} : \mathbb{R}^{\mathcal{T}\times N_T} \to L^\infty((0,T)\times\Omega)$ are defined respectively by

$$\pi_{\mathcal{T}}\boldsymbol{u}(x) = u_K \quad \text{if } x \in K, \qquad \forall \boldsymbol{u} = (u_K)_{K\in\mathcal{T}},$$

and

$$\pi_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}(t,x) = u_K^n \quad \text{if } (t,x) \in (t_{n-1}, t_n]\times K, \qquad \forall \boldsymbol{u} = (u_K^n)_{K\in\mathcal{T}, 1\leq n\leq N_T}.$$

These operators allow passing from the discrete solution $(\boldsymbol{C}^n, \boldsymbol{\Phi}^n)_{1\leq n\leq N_T}$ to the approximate solution since

$$\Phi_{\mathcal{T},\boldsymbol{\Delta t}} = \pi_{\mathcal{T},\boldsymbol{\Delta t}}(\boldsymbol{\Phi}), \qquad c_{i,\mathcal{T},\boldsymbol{\Delta t}} = \pi_{\mathcal{T},\boldsymbol{\Delta t}}(\boldsymbol{c}_i), \qquad \forall i \in [\![1, N]\!].$$

To carry out the analysis, we further need to introduce an approximate gradient reconstruction. Since the boundary conditions play a crucial role in the definition of the gradient, we need to enrich the discrete solution by face values $(C_\sigma^n)_{\sigma\in\mathcal{E}_{\text{ext}}, 1\leq n\leq N}$ and $(\Phi_\sigma^n)_{\sigma\in\mathcal{E}_{\text{ext}}, 1\leq n\leq N}$ defined by $C_\sigma^n = C_{K\sigma}^n$ and $\Phi_\sigma^n = \Phi_{K\sigma}^n$

for $\sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K$. With a slight abuse of notations, we still denote by $\boldsymbol{C}^n = ((C_K^n)_{K\in\mathcal{T}}, (C_\sigma^n)_{\sigma\in\mathcal{E}_{\rm ext}})$ and $\boldsymbol{\Phi}^n = ((\Phi_K^n)_{K\in\mathcal{T}}, (\Phi_\sigma^n)_{\sigma\in\mathcal{E}_{\rm ext}})$ the elements of $\mathcal{A}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}$ and $\mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}$ containing both the cell values and the exterior faces values of the concentration and the potential respectively.

For $\sigma = K|L \in \mathcal{E}_{\rm int}$, we denote by Δ_σ the diamond cell corresponding to σ, that is the interior of the convex hull of $\sigma \cup \{x_K, x_L\}$. For $\sigma \in \mathcal{E}_{\rm ext}$, the diamond cell Δ_σ is defined as the interior of the convex hull of $\sigma \cup \{x_K\}$. The approximate gradient $\nabla_\mathcal{T} : \mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}} \to L^2(\Omega)^d$ is piecewise constant on the diamond cells Δ_σ, and it is defined as follows:

$$\nabla_\mathcal{T}\boldsymbol{u}(x) = d\frac{D_{K\sigma}\boldsymbol{u}}{d_\sigma}\boldsymbol{n}_{K\sigma} \quad \text{if } x \in \Delta_\sigma, \qquad \forall \boldsymbol{u} \in \mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}.$$

We also define $\nabla_{\mathcal{T},\boldsymbol{\Delta t}} : \mathbb{R}^{(\mathcal{T}\cup\mathcal{E}_{\rm ext})\times N} \to L^2(Q_T)^d$ by setting

$$\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}(t,\cdot) = \nabla_\mathcal{T}\boldsymbol{u}^n \quad \text{if } t \in (t_{n-1}, t_n], \qquad \forall \boldsymbol{u} = (\boldsymbol{u}^n)_{1\le n\le N} \in \mathbb{R}^{(\mathcal{T}\cup\mathcal{E}_{\rm ext})\times N}.$$

This reconstruction is merely weakly consistent (unless $d = 1$) and takes its source in [48, 66]. More consistent reconstruction operators will be introduced in Section 3.4.3. Let us recall now some key properties to be used in the analysis. First, for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}$,

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\boldsymbol{u} D_{K\sigma}\boldsymbol{v} = \frac{1}{d}\int_\Omega \nabla_\mathcal{T}\boldsymbol{u}\cdot\nabla_\mathcal{T}\boldsymbol{v}\mathrm{d}x.$$

This implies in particular that

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma |D_\sigma\boldsymbol{u}|^2 = \frac{1}{d}\int_\Omega |\nabla_\mathcal{T}\boldsymbol{u}|^2\mathrm{d}x, \qquad \forall \boldsymbol{u} \in \mathbb{R}^{\mathcal{T}\cup\mathcal{E}_{\rm ext}}. \tag{3.4.1}$$

3.4.2 Compactness

In this section we intend to prove a discrete H^1 estimate on the concentrations using the bound on the free-energy dissipation (3.3.14). To that extend we will introduce a chemical dissipation $\mathcal{D}_{\rm chem}$ as a discrete equivalent to $\sum c_i|\nabla h_i(c)|^2$ and compare it both with the usual distance and the total dissipation $\mathcal{D}$.

As the identification of the limit is only possible for

$$\inf_{Q_T} c_0 > \epsilon > 0,$$

the results of this section are proved under this assumption and complemented with remarks indicating whether the hypothesis is necessary or not. In order to apply

chain rules for the convergence, we need to change the face concentration $\mathcal{C}$ from the one defined by the numerical scheme through Lemma 3.3.1 to the logarithmic average:

$$\widetilde{\mathcal{C}}_i(C_K, C_L) = \frac{c_{K,i} - c_{L,i}}{\log(c_{K,i}) - \log(c_{L,i})} \qquad \forall i \in [\![0, N]\!]. \tag{3.4.2}$$

This choice of edge concentration will also be used in the definition of $\mathcal{D}_{\text{chem}}$ to avoid a dependency on the potential. The following lemma provides an estimate the numerical-flux based averages using this logarithmic average.

Lemma 3.4.1. *For all $\epsilon > 0$ there exists $\alpha_\epsilon > 0$ depending only on ϵ, M_Φ such that, for all $(C_K, C_L, \Phi_K, \Phi_L) \in \mathcal{A} \times \mathcal{A} \times [-M_\Phi, M_\Phi] \times [-M_\Phi, M_\Phi]$, and for all $i \in [\![1, N]\!]$:*

$$c_{K,0}, c_{L,0} > \epsilon \implies \alpha_\epsilon \widetilde{\mathcal{C}}_i(C_K, C_L) \leq \mathcal{C}_i(C_K, C_L, \Phi_K, \Phi_L). \tag{3.4.3}$$

Proof. For the centered scheme, this inequality is known with $\alpha_\epsilon = 1$ without assumption on c_0 [109]. For the Sedan scheme the proof is more intricate and uses the hypothesis on c_0. Equation (3.4.3) is equivalent to the boundedness of

$$R_i(C_K, C_L, \Phi_K, \Phi_L) := \frac{\widetilde{\mathcal{C}}_i(C_K, C_L)}{\mathcal{C}_i(C_K, C_L, \Phi_K, \Phi_L)},$$

for $c_{K,0}, c_{L,0} > \epsilon$. Introduce $x_i = \log \frac{c_{K,i}}{c_{L,i}}$, and $y_i = z_i\Phi_L + \nu_i(C_L) - z_i\Phi_K - \nu_i(C_K)$ as in the proof of lemma 3.3.1. By symmetry, one can assume $x_i \geq 0$ and thanks to our assumption on the solvent and the potential, y_i is bounded by some K. Moreover, we notice that by definition of x_i, (3.3.4) yields:

$$\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) = c_{L,i}\big(B(y_i)e^{x_i} - B(-y_i)\big),$$

so that we have:

$$R_i(C_K, C_L, \Phi_K, \Phi_L) = \frac{e^{x_i} - 1}{x_i} \frac{x_i - y_i}{B(y_i)e^{x_i} - B(-y_i)}.$$

The right-hand side can be seen as a continuous function of x_i, y_i. Is is bounded on the boundary of its definition domain $[0, +\infty) \times [-K, K]$ and admits a finite limit $\frac{1}{B(\mu)}$ for $x_i \to \infty, y_i \to \mu$, thus R_i is bounded. □

Then we try to take advantage of Proposition 3.3.1. As Lemma 3.3.4 already provides satisfying estimates on $\mathbf{\Phi}$, we introduce

$$\mathcal{D}_{\text{chem,i}}(C_K, C_L) : \begin{array}{ccc} \mathcal{A}^2 & \to & \mathbb{R} \\ (C_K, C_L) & \mapsto & \widetilde{\mathcal{C}}_i(C_K, C_L)(h_i(C_K) - h_i(C_L))^2, \end{array}$$

and

$$\mathcal{D}_{\text{chem}} = \sum_{i=1}^{N} \mathcal{D}_{\text{chem,i}}.$$

A first interesting result is that $\mathcal{D}_{\text{chem}}$ is a semimetric on $\mathcal{A}$. The non-negativity and symmetry properties are trivially satisfied, the last property is the subject of the following lemma.

Lemma 3.4.2. *We have $\mathcal{D}_{\text{chem}}(C_K, C_L) = 0$ if and only if $C_K = C_L$*

Proof. If $C_K = C_L$, we obviously have $\mathcal{D}_{\text{chem}}(C_K, C_L) = 0$, we will then focus on the other implication. Assume that we dispose of C_K, C_L in $\mathcal{A}$ such that $\mathcal{D}_{\text{chem}}(C_K, C_L) = 0$. We let for $i \in [\![0, N]\!]$:

$$a_{K,i} = \log \frac{c_{K,i}}{\bar{c}_K} \qquad a_{L,i} = \log \frac{c_{L,i}}{\bar{c}_L}, \tag{3.4.4}$$

such that $h_i(C_K) = a_{K,i} - k_i a_{K,0}$. We have $\widetilde{\mathcal{C}}_i(C_K, C_L) \geq \min(c_{K,i}, c_{L,i}) > 0$, thus $\mathcal{D}_{\text{chem}}$ is the sum of nonnegative terms. As we have $\mathcal{D}_{\text{chem}}(C_K, C_L) = 0$, we have:

$$a_{K,i} - k_i a_{K,0} = a_{L,i} - k_i a_{L,0} \qquad \forall i \in [\![1, N]\!].$$

Assume that $a_{K,0} = a_{L,0}$, then $A_K = A_L$, where $A = (a_0, ..., a_N)$. Using $\sum_{i=0}^{N} k_i e^{a_i} = \frac{1}{v_0 \bar{c}}$, $A_K = A_L$ implies $C_K = C_L$.

The other case is absurd: using the symmetry of $\mathcal{D}_{\text{chem}}$, one can freely assume that $a_{K,0} > a_{L,0}$. Using $k_i > 0$, we have $a_{K,i} > a_{L,i} \forall i \in [\![1, N]\!]$ hence:

$$1 = \sum_{i=0}^{N} e^{a_{K,i}} > \sum_{i=0}^{N} e^{a_{L,i}} = 1.$$

□

The function $\mathcal{D}_{\text{chem}}$ cannot be extended by continuity onto $\overline{\mathcal{A}}^2$. Some information for near zero concentrations is can be inferred from lemma 3.3.2. The following sequential result means that the semi-metric property is preserved near the boundary $\partial\mathcal{A}^2$.

Lemma 3.4.3. *Let (C_K^l, C_L^l) be a sequence of $\mathcal{A}^2$. If we have $\mathcal{D}_{\text{chem}}(C_K^l, C_L^l) \to 0$ then $C_K^l - C_L^l \to 0$*

Proof. For the sake of simplicity, as this result will only be used with a lower bound on c_0, we keep the proof to this simpler case and assume that $\inf(c_{K,0}^l, c_{L,0}^l) > 0$. To prove the limit, we will show that from any sub-sequence, we can extract a sub-sub-sequence such that $C_K^l - C_L^l \to 0$. Considering any sub-sequence, thanks

to the boundedness of $\mathcal{A}$ we can extract a sub-sub-sequence such that C_K^l and C_L^l converge. If we dispose of $i \in [\![1, N]\!]$ such that $c_{K,i}^l \to c^* > 0$ while $c_{L,i}^l \to 0$ (or the symmetric situation), then we have:

$$\widetilde{\mathcal{C}}_{\sigma,i}(C_K^l, C_L^l) \underset{l\to\infty}{\sim} \frac{c^*}{-\log(c_{L,i}^l)} \qquad \text{and} \qquad \left(h_i(C_K^l) - h_i(C_L^l)\right)^2 \underset{l\to\infty}{\sim} \log(c_{L,i}^l)^2$$

so that $\mathcal{D}_{\text{chem,i}}(C_K^l, C_L^l) \sim_{l\to\infty} -c^* \log(c_{L,i}^l) \to \infty$, which is absurd. Necessary, we have $c_{L,i}^l \to 0$ if and only of $c_{K,i}^l \to 0$. As we have $\mathcal{D}_{\text{chem,i}}(C_K^l, C_L^l) \to 0$, we also have:

$$a_{K,i}^l - k_i a_{K,0}^l = a_{L,i}^l - k_i a_{L,0}^l + o(1) \qquad \forall i \in [\![1, N]\!], \inf_l(c_{L,i}^l) > 0,$$

where $a_{K,i}, a_{L,i}$ are defined by (3.4.4). As both $c_{K,i}^l$ and $c_{L,i}^l$ are bounded away from zero, $a_{K,i}^l$ and $a_{L,i}^l$ are convergent. Using the symmetry and up to a subsequence we have either $a_{K,0}^l - a_{L_0}^l$ convergent of limit zero or bounded away from zero. We conclude using the same ideas as the proof of the previous lemma. □

This semi-metric is however not commonly used and the following lemma intends to compare it with the usual distance.

Proposition 3.4.1. *For all $i \in [\![0, N]\!]$, there exist M such that:*

$$\frac{(c_{K,i} - c_{L,i})^2}{\mathcal{D}_{\text{chem}}(C_K, C_L)} \leq M, \qquad \forall\, C_K, C_L \in \mathcal{A}^2. \tag{3.4.5}$$

Proof. We will prove the result for $i \in [\![1, N]\!]$ using *reductio ad absurdum* and case exhaustion.

Let $(C_K^n, C_L^n) \in (\mathcal{A}^2)^{\mathbb{N}}$ be such that $\frac{(c_{K,i}^n - c_{L,i}^n)^2}{\mathcal{D}_{\text{chem}}(C_K^n, C_L^n)} \to \infty$. We let $\epsilon^n := C_L^n - C_K^n$ and use the boundedness of $\mathcal{A}$ to extract a convergent sub-sequence of (C_K^n, ϵ^n) and denote (C^*, ϵ^*) its limit. As c_i is bounded, we have $\mathcal{D}_{\text{chem}}(C_K^n, C_L^n) \to 0$. Thanks to Lemma 3.4.3, we have $\epsilon^* = 0$ so that we will consider first order development in ϵ^n. We notice that the blow-up of the ratio implies that:

$$\mathcal{D}_{\text{chem,j}}(C_K^n, C_L^n) = o(|\epsilon^n|^2) \qquad \forall j \in [\![1, N]\!]. \tag{3.4.6}$$

For the sake of readability, we will drop from now on the superscript ^{n}. We have to consider three cases:

1. $c_j^* = 0$ implies $\epsilon_j = o(|\epsilon|)$;
2. we dispose of species such that $\epsilon_j \neq o(|\epsilon|)$ and $c_j^* = 0$, but for all of them $\log \frac{\epsilon_j + c_j}{c_j}$ remains bounded;

3. we dispose of a specie such that $\epsilon_j \neq o(|\epsilon|)$, $c_j^* = 0$, and up to a subsection, $\log \frac{\epsilon_j + c_j}{c_j}$ blows-up.

Preliminary remark about the solvent

We consider $j \in [\![1, N]\!]$ such that $c_j^* > 0$ and let $\overline{\epsilon}_j = \sum_{i=0}^{N} \epsilon_i$. We have, thanks to (3.4.6):

$$\log \frac{c_j + \epsilon_j}{c_j} = O(|\epsilon|) \quad \text{and} \quad \log \frac{\overline{c}_j + \overline{\epsilon}_j}{\overline{c}_j} = O(|\epsilon|) \quad \text{and} \quad h_j(C_K) - h_j(C_L) = O(|\epsilon|),$$

so that:

$$\log \frac{c_0 + \epsilon_0}{c_0} = O(|\epsilon|),$$

thus:

$$\frac{\epsilon_0}{c_0} = O(|\epsilon|) \qquad \text{and} \qquad \log \frac{c_0 + \epsilon_0}{c_0} = \frac{\epsilon_0}{c_0} + o(|\epsilon|).$$

Conclusion of the proof in case 1

The proof of this first case is by far the most intricate of the three. It is done in two step: first we use our hypothesis on $\mathcal{D}_{\text{chem}}$, c, and ϵ to obtain a estimate where the species are coupled through an ersatz of $\overline{\epsilon}$ and $\overline{c}$. Then we show an improved version of the Cauchy-Schwarz inequality to improve the estimate into decoupled estimates which are incompatible with our hypothesis.

First order development of h_j gives:

$$h_j(C_K) - h_j(C_L) = \frac{\epsilon_j}{c_j} - k_j \frac{\epsilon_0}{c_0} + (k_j - 1)\frac{\overline{\epsilon}}{\overline{c}} + o(\epsilon), \qquad \forall j \in [\![1, N]\!], c_j^* > 0.$$

Thanks to (3.4.6) we have the estimation:

$$\frac{\epsilon_j}{c_j} - k_j \frac{\epsilon_0}{c_0} + (k_j - 1)\frac{\overline{\epsilon}}{\overline{c}} = o(|\epsilon|), \qquad \forall j \in [\![1, N]\!], c_j^* > 0. \tag{3.4.7}$$

To correct the effect of the species with negligible concentrations, we let:

$$\widetilde{\epsilon_0} = -\sum_{\substack{c_j^* > 0 \\ j \neq 0}} k_j \epsilon_j \qquad \widetilde{\epsilon} = \widetilde{\epsilon_0} + \sum_{\substack{c_j^* > 0 \\ j \neq 0}} \epsilon_j \qquad \text{and} \qquad \widetilde{c} = \sum_{c_j^* > 0} c_j$$

By construction, we have $\widetilde{c} = \overline{c} + o(1)$. Using the hypothesis (1) we have $\widetilde{\epsilon} =$

$\overline{\epsilon} + o(|\epsilon|)$ and $\widetilde{\epsilon_0} = \epsilon_0 + o(|\epsilon|)$. These three results and equation (3.4.7) yields:

$$\frac{\epsilon_j}{c_j} - k_j \frac{\widetilde{\epsilon_0}}{c_0} + (k_j - 1)\frac{\widetilde{\epsilon}}{\widetilde{c}} = o(|\epsilon|), \qquad \forall j \in [\![1, N]\!], c_j^* > 0.$$

We let $\xi_j = \frac{\epsilon_j}{c_j} - \frac{\widetilde{\epsilon}}{\widetilde{c}}$ for $j \neq 0$ and $\xi_0 = \frac{\widetilde{\epsilon_0}}{c_0} - \frac{\widetilde{\epsilon}}{\widetilde{c}}$. Previous equation yields:

$$\xi_j = k_j \xi_0 + o(|\epsilon|), \qquad \forall j \in [\![1, N]\!], c_j^* > 0.$$

Considering $\sum_{c_j^*>0} c_j \xi_j$, we have:

$$0 = \widetilde{\epsilon} - \widetilde{\epsilon} = \sum_{c_j^*>0} c_j \xi_j = \sum c_j k_j \xi_0 + o(|\epsilon|) = \xi_0 \left(\frac{1}{v_0} + o(1) \right) + o(|\epsilon|),$$

so that:

$$\xi_0 = o(|\epsilon|).$$

We conclude the first part of the proof with the following estimate that follows from (3.4.6):

$$\sum_{c_j^*>0 \cup \{0\}} c_j \xi_j^2 = o(|\epsilon|^2). \tag{3.4.8}$$

For the sake of readability, we will drop the $\widetilde{}$ over ϵ_0 in the second part of the proof, use "$c_j^* > 0$" instead of "$c_j^* > 0$ or $j = 0$", and assume by symmetry that $\widetilde{\epsilon} \geq 0$. We have :

$$\sum_{c_j^*>0} c_j \xi_j^2 = \sum_{c_j^*>0} \frac{\epsilon_j^2}{c_j} - \frac{\widetilde{\epsilon}^2}{\widetilde{c}}. \tag{3.4.9}$$

Let $x_j = \sqrt{c_j}$, $y_j = \frac{\epsilon_j}{\sqrt{c_j}}$. We have:

$$\widetilde{\epsilon} = \sum_{c_j^*>0} x_j y_j, \qquad |x|^2 = \widetilde{c}, \qquad |y|^2 = \sum_{c_j^*>0} \frac{\epsilon_j^2}{c_j},$$

Thus the Cauchy-Schwarz inequality yields

$$\widetilde{\epsilon}^2 \leq \widetilde{c} \sum_{c_j^*>0} \frac{\epsilon_j^2}{c_j},$$

hence another proof of the non-negativity of the right-hand side of (3.4.9). We intend to use ideas presented in [6] to improve the estimation of $\widetilde{\epsilon}$. More precisely,

the stability version of the Cauchy-Schwarz presented in [5] gives:

$$\widetilde{\epsilon} = |x||y| \left(1 - \frac{1}{2}\left|\frac{x}{|x|} - \frac{y}{|y|}\right|^2\right).$$

We intend to show that $\left|\frac{x}{|x|} - \frac{y}{|y|}\right|$ is bounded away from zero. To show this bound we let:

$$K : (C, \epsilon) \mapsto \left|\frac{x}{|x|} - \frac{y}{|y|}\right|$$

and consider a minimizing sequence of K under the conditions

$$c_j > 0, \qquad \epsilon_0 = -\sum_{\substack{c_j^*>0 \\ j\neq 0}} k_j \epsilon_j.$$

As K is invariant by scaling, we can assume that we have a convergent minimizing sequence $C^l_{\text{inf}}, \epsilon^l_{\text{inf}}$ of limit $C^*_{\text{inf}}, \epsilon^*_{\text{inf}}$ and of norm equal to 1. Note that we do not assume $C \in \mathcal{A}$, nor $C^*_{\text{inf}} > 0$ thus we consider broader options than necessary for use in (3.4.9) to ensure existence of the minimum. Finally, we notice that K is non negative, its infimum is either zero or positive. We will prove the positivity by contradiction.

Assume that the limit of $K(C^l_{\text{inf}}, \epsilon^l_{\text{inf}})$ is zero, we show that $|y^l_{\text{inf}}|$ is convergent up to a subsequence. We consider j such that, up to a subsequence, $\frac{|y^l_j|}{|y^l_{\text{inf}}|}$ is bounded away from zero. If $c^*_{\text{inf},j} \neq 0$, $|y^l_j|$ is bounded thus $|y^l_{\text{inf}}|$ is too, and up to another subsequence, it is convergent. If $c^*_{\text{inf},j} = 0$ we notice that $x^l_{\text{inf},j} \to 0$ and $|x^l_{\text{inf}}|$ is bounded away from zero, so that $\frac{|y^l_j|}{|y^l_{\text{inf}}|} \to 0$, which is absurd. We let γ be the limit of $|y^l_{\text{inf}}|^2$.

As we have assumed the infimum to be zero, we have:

$$\epsilon^*_{\text{inf},j} = c^*_{\text{inf},j} \frac{\gamma}{\widetilde{c^*_{\text{inf}}}}, \qquad \forall j \text{ s.t. } c^*_j > 0.$$

This would imply that ϵ^*_{inf} is nonnegative, however, we have by definition $\epsilon^*_{\text{inf},0} = -\sum_{j=1}^N k_j \epsilon^*_{\text{inf},j}$ and ϵ^*_{inf} is of norm 1. This is absurd, hence the infimum cannot be zero. Thus we dispose of $0 < \alpha$ depending only on $k_1, \dots k_N$ and the subset $\{c^*_j > 0\}$ of $[\![0, N]\!]$ such that:

$$\widetilde{\epsilon} \leq |x||y| (1 - \alpha).$$

As we have assumed (using symmetry) $\widetilde{\epsilon} \geq 0$, we also have $\alpha \leq 1$. So that we

have:

$$\sum_{c_j^*>0} \frac{\epsilon_j^2}{c_j} - \frac{\widetilde{\epsilon}^2}{\widetilde{c}} = |y|^2 - \frac{\widetilde{\epsilon}^2}{|x|^2} \geq |y|^2(1-(1-\alpha)^2) = \sum_{c_j^*>0} \frac{\epsilon_j^2}{c_j}(1-(1-\alpha)^2).$$

Thanks to equations (3.4.8) and (3.4.9), we have:

$$\sum_{c_j^*>0} \frac{\epsilon_j^2}{c_j} = o(|\epsilon|^2),$$

thus, thanks to (1), $\epsilon_j = o(|\epsilon|)$ for all $j \in [\![0, N]\!]$, which is absurd.

Conclusion of the proof in case 2

We dispose of j such that $c_j \to 0$ and $\epsilon_j \neq o(|\epsilon|)$, thus have up to a subsequence:

$$|\epsilon| = O(\epsilon_j) \qquad \text{and} \qquad \mathcal{D}_{\text{chem,j}} = \widetilde{\mathcal{C}}_{\sigma,j}\left(\log\frac{c_j+\epsilon_j}{c_j} + O(|\epsilon|)\right)^2.$$

The assumed boundedness of $\log\frac{c_j+\epsilon_j}{c_j}$ implies that $\epsilon_j = O(c_j)$ thus, $c_j \neq o(|\epsilon|)$. Moreover, we also dispose of $\alpha = \min(1, \inf_n \frac{c_j^n+\epsilon_j^n}{c_j^n}) > 0$ such that:

$$\widetilde{\mathcal{C}}_{\sigma,j} \geq \alpha c_j$$

Necessary, we have $\log\frac{c_j+\epsilon_j}{c_j} \to 0$ thus:

$$\mathcal{D}_{\text{chem,j}} \geq \alpha\frac{\epsilon_j^2}{c_j} + o(\frac{\epsilon_j^2}{c_j}),$$

which is bigger than $|\epsilon|^2$ and thus contradicts (3.4.6).

Conclusion of the proof in case 3

Let j be such that $\epsilon_j \neq o(|\epsilon|)$, $c_j^* = 0$, and $\log\frac{\epsilon_j+c_j}{c_j}$ blows-up.
We have:

$$h_j(C+\epsilon) - h_j(C) = \log\frac{\epsilon_j+c_j}{c_j} + o(1),$$

and:

$$\widetilde{\mathcal{C}_{\sigma,j}}(C+\epsilon, C) = \frac{\epsilon_j}{\log\frac{\epsilon_j+c_j}{c_j}},$$

so that:

$$\mathcal{D}_{\text{chem,j}} \sim \epsilon_j \log \frac{\epsilon_j + c_j}{c_j},$$

which contradicts (3.4.6) since $\epsilon_j \neq o(|\epsilon|)$.

Global conclusion

As each of the cases lead to a contradiction, we have the desired inequality for $i \in [\![1, N]\!]$. For the solvent, we see that:

$$c_{K,0} - c_{L,0} = -\sum_{i=1}^{N} k_i (c_{K,i} - c_{L,i}),$$

thus the announced result up to the choice of a bigger constant M. □

Using these tools, we may now prove the following necessary compactness inequality:

Proposition 3.4.2. *For all $\epsilon > 0$, there exist M such that :*

$$\inf_{\substack{mesh\ m \\ n \in [\![1, N_{T,m}]\!] \\ K \in \mathcal{T}_m}} c^n_{m,K,0} > \epsilon \implies \|\nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{c}_i\|^2_{L^2(Q_T)} \leq M, \ \forall i \in [\![0, N]\!], \forall m.$$

Proof. We will show the result for $i \in [\![1, N]\!]$ and use the definition of $\mathcal{A}$ to extend it the solvent. For improved readability, we will drop the subscript m. By definition, we have:

$$|\nabla_{\mathcal{T}, \boldsymbol{\Delta t}} \boldsymbol{c}_i|^2_{L^2(Q_T)} = \sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma \in \mathcal{E}_{\text{int}}} \tau_\sigma (D_\sigma \boldsymbol{c}_i^n)^2$$

Thanks to Proposition 3.4.1 and Lemma 3.4.1, we have:

$$\sum_{\sigma \in \mathcal{E}_{\text{int}}} \tau_\sigma (D_\sigma \boldsymbol{c}_i^n)^2 \leq M \sum_{\sigma = K|L \in \mathcal{E}_{K,\text{int}}} \tau_\sigma \mathcal{D}_{\text{chem}}(C_K^n, C_L^n) \leq \frac{M}{\alpha_\epsilon} \sum_{\sigma \in \mathcal{E}_{\text{int}}} \sum_{j=1}^{N} \tau_\sigma \mathcal{C}^n_{\sigma,j} (D_\sigma h_j(\boldsymbol{C}^n))^2.$$

It is sufficient to bound $\sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma \in \mathcal{E}_{\text{int}}} \tau_\sigma \mathcal{C}^n_{\sigma,j} (D_\sigma h_j(\boldsymbol{C}^n))^2$, for all $j \in [\![1, N]\!]$ to get the desired result. We have:

$$(D_\sigma h_j(\boldsymbol{C}^n))^2 \leq 2(D_\sigma(h_j(\boldsymbol{C}^n) - z_j \boldsymbol{\Phi}^n))^2 + 2(z_j D_\sigma \boldsymbol{\Phi}^n)^2$$

Thanks to equation (3.3.14) of Lemma 3.3.4, we dispose of M such that :

$$\sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma \in \mathcal{E}_{\text{int}}} \tau_\sigma \mathcal{C}_{\sigma,j}^n (D_\sigma(h_j(\boldsymbol{C}^n) - z_j \boldsymbol{\Phi}^n))^2 < M.$$

Moreover, $\sum_{\sigma \in \mathcal{E}_{\text{int}}} \tau_\sigma \mathcal{C}_{\sigma,j}^n (z_j D_\sigma \boldsymbol{\Phi}^n)^2$ is also bounded thanks to (3.3.12) and the L^∞ bound on $\mathcal{C}_i$ and z_i. Thus we have:

$$\sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma \in \mathcal{E}_{\text{int}}} \tau_\sigma \mathcal{C}_{\sigma,j}^n (D_\sigma h_j(\boldsymbol{C}^n))^2 \leq M, \tag{3.4.10}$$

which in turn yields the desired result.

For the solvent we notice that:

$$\nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{c}_0 = -\sum_{i=1}^{N} k_i \nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{c}_i,$$

so that the bound on all $\nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{c}_i$ transfers into a bound on $\nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{c}_0$. □

Using this discrete $L^2(H^1)$ estimate, we use a discrete Aubin-Lions lemma to get the compactness of the sequence of solutions, as stated in following proposition:

Proposition 3.4.3. *Let $(\boldsymbol{C}_m, \boldsymbol{\Phi}_m)$ be the family of discrete solutions defined either by the centered scheme or by the Sedan scheme. In both cases, there exists $\Phi \in L^\infty(Q_T; \mathbb{R}) \cap L^2((0,T); H^1(\Omega))$, $C \in L^\infty(Q_T; \overline{\mathcal{A}})$ such that, up to a subsequence,*

$$\pi_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{C}_m \underset{m \to \infty}{\longrightarrow} C \quad \textit{strongly in } L^2(Q_T)^{N+1}, \tag{3.4.11}$$

$$\nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{C}_m \underset{m \to \infty}{\longrightarrow} \nabla C \quad \textit{weakly in } L^2(Q_T), \tag{3.4.12}$$

$$\pi_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{\Phi}_m \underset{m \to \infty}{\longrightarrow} \Phi \quad \textit{in the } L^\infty(Q_T) \textit{ weak-}\star \textit{ sense}, \tag{3.4.13}$$

$$\nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{\Phi}_m \underset{m \to \infty}{\longrightarrow} \nabla \Phi \quad \textit{in the } L^\infty([0,T], L^2(\Omega)^d) \textit{ weak-}\star \textit{ sense}. \tag{3.4.14}$$

Proof. For improved readability we drop again the subscripts m. The proof of the first two result relies on a discrete Aubin-Lions lemma [78, Lemma 3.4]. We intend to use it in the setting described in [31, Lemma 9]. Proposition 3.4.2 provides a first property, but we still have to prove that there exist C independent of the mesh such that $\sum_n \|\boldsymbol{c}_i^n - \boldsymbol{c}_i^{n-1}\|_{\mathcal{T},-1} \leq C$, where $\|\cdot\|_{\mathcal{T},-1}$ is defined by duality:

$$\|\boldsymbol{c}\|_{\mathcal{T},-1} = \sup_{\boldsymbol{\varphi}} \left(\int_\Omega \pi_{\mathcal{T}} \boldsymbol{c} \pi_{\mathcal{T}} \boldsymbol{\varphi},\ \|\pi_{\mathcal{T}} \boldsymbol{\varphi}\|_{L^2}^2 + \|\nabla_{\mathcal{T}} \boldsymbol{\varphi}\|_{L^2}^2 = 1 \right).$$

Let $\boldsymbol{\varphi} \in \mathbb{R}^{\mathcal{T}}$. Tanks to (3.2.4b), we have:

$$\int_\Omega \pi_{\mathcal{T}}(\boldsymbol{c}_i^n - \boldsymbol{c}_i^{n-1})\pi_{\mathcal{T}}\boldsymbol{\varphi} = -\Delta t_n \sum_{K\in\mathcal{T}} \varphi_K \sum_{\sigma\in\mathcal{E}_K} F_{K\sigma,i}^n.$$

Using the definition of $F_{K\sigma,i}^n$ along with the definition of $\mathcal{C}_{\sigma,i}$ respectively equations (3.2.5) and (3.3.1), we have:

$$\int_\Omega \pi_{\mathcal{T}}(\boldsymbol{c}_i^n - \boldsymbol{c}_i^{n-1})\pi_{\mathcal{T}}\boldsymbol{\varphi} = \Delta t_n \sum_{\sigma=K|L\in\mathcal{E}_{K,\mathrm{int}}} D_i \tau_\sigma \mathcal{C}_{\sigma,i}(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) D_{K\sigma}\left(h_i(\boldsymbol{C}^n) + z_i\boldsymbol{\Phi}^n\right) D_{K\sigma}\boldsymbol{\varphi}.$$

Thanks to the Cauchy-Schwarz inequality, we have:

$$\int_\Omega \pi_{\mathcal{T}}(\boldsymbol{c}_i^n - \boldsymbol{c}_i^{n-1})\pi_{\mathcal{T}}\boldsymbol{\varphi} \le \Delta t_n D_i \left(\sum_{\sigma=K|L\in\mathcal{E}^{\mathrm{int}}} \tau_\sigma \mathcal{C}_{\sigma,i}(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n)\left(D_{K\sigma}h_i(\boldsymbol{C}^n) + z_i\boldsymbol{\Phi}^n\right)^2 \right)^{\frac{1}{2}} \left(\sum_{\sigma=K|L\in\mathcal{E}^{\mathrm{int}}} \tau_\sigma \mathcal{C}_{\sigma,i}(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n)\left(D_{K\sigma}\boldsymbol{\varphi}\right)^2 \right)^{\frac{1}{2}}.$$

Using the definition of $\mathcal{D}_{\sigma,i}$, $\mathcal{C}_{\sigma,i} \le \frac{1}{k_i v_0}$ and $\|\nabla_{\mathcal{T}}\boldsymbol{\varphi}\|_{L^2}^2 \le 1$, we have:

$$\|\boldsymbol{c}_i^n - \boldsymbol{c}_i^{n-1}\|_{\mathcal{T},-1} \le \frac{\Delta t_n}{k_i D_i v_0} \left(\sum_{\sigma=K|L\in\mathcal{E}^{\mathrm{int}}} \tau_\sigma \mathcal{D}_{\sigma,i}(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) \right)^{\frac{1}{2}}.$$

Using the Cauchy-Schwarz inequality and Lemma 3.3.4 equation (3.3.14), we have :

$$\sum_n \|\boldsymbol{c}_i^n - \boldsymbol{c}_i^{n-1}\|_{\mathcal{T},-1} \le \left(\sum_n \frac{\Delta t_n}{k_i^2 D_i^2 v_0^2} \right)^{\frac{1}{2}} \left(\sum_n \Delta t_n \sum_{\sigma=K|L\in\mathcal{E}_{K,\mathrm{int}}} \tau_\sigma \mathcal{D}_{\sigma,i}(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) \right)^{\frac{1}{2}} \le C.$$

This concludes the proof of equations (3.4.11) and (3.4.12).

We may now focus on the convergence of the potential. The existence of Φ satisfying (3.4.13) is a straightforward consequence of (3.3.11). Similarly, (3.3.12) implies the existence of a vector field u such that $\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\Phi}_m \underset{m\to\infty}{\longrightarrow} u$ in the $L^\infty([0,T], L^2(\Omega)^d)$ weak-$\star$ sense.

We have to identify u with $\nabla\Phi$. We let $w \in C_c^\infty(Q_T, R^d)$ and define:

$$\boldsymbol{w}_\sigma^n = \fint_\sigma w(t_n, x)dx \qquad \forall \sigma \in \mathcal{E}, n \in [\![1, N_T]\!],$$

and the associated diamond-cell reconstruction:

$$w_{\mathcal{E},\boldsymbol{\Delta t}}(t, x) = \boldsymbol{w}_\sigma^n \qquad \text{if } x \in \Delta_\sigma \text{ and } t \in (t_{n-1}, t_n].$$

Thanks to the smoothness of w, we have convergence of $w_{\mathcal{E},\boldsymbol{\Delta t}}$ toward w and:

$$\iint_{Q_T} w_{\mathcal{E},\boldsymbol{\Delta t}} \cdot \nabla_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{\Phi} \to \iint_{Q_T} w \cdot u.$$

Using the geometric relation $d_\sigma m_\sigma = dm_{\Delta_\sigma}$ and the definition of $\boldsymbol{w}_\sigma^n$, we have:

$$\iint_{Q_T} w_{\mathcal{E},\boldsymbol{\Delta t}} \cdot \nabla_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{\Phi} = -\sum_{i=1}^{N_T} \Delta t_n \sum_{K \in \mathcal{T}} \Phi_K^n \int_K \operatorname{div}(w(t_n, x))dx.$$

Thanks to the smoothness of w and the convergence of $\boldsymbol{\Phi}$, we have:

$$\iint_{Q_T} w_{\mathcal{E},\boldsymbol{\Delta t}} \cdot \nabla_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{\Phi} \to -\iint_{Q_T} \Phi \operatorname{div}(w) = \iint_{Q_T} \nabla\Phi \cdot w$$

This concludes the identification of u and the proof of (3.4.14). □

These convergence topologies are sub-optimal and will be improved later in Lemma 3.4.4. First, we notice that for the concentrations, we also dispose of edge values defined by $\mathcal{C}_\sigma$ and $\widetilde{\mathcal{C}_\sigma}$ in equations (3.3.1) and (3.4.2). Using these face values, we introduce another reconstruction. For i in $[\![1, N]\!]$, we let:

$$c_{\mathcal{E},\boldsymbol{\Delta t},i}(x, t) = \begin{cases} \mathcal{C}_{\sigma,i}(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) & \text{if } x \in \Delta_{K|L} \text{ and } t \in (t_{n-1}, t_n], \\ c_K^n & \text{if } x \in \Delta_\sigma, \sigma \in \mathcal{E}_K \cap \mathcal{E}^{\text{ext}} \text{ and } t \in (t_{n-1}, t_n]. \end{cases}$$

Similarly, we introduce $\widetilde{c}_{\mathcal{E},\boldsymbol{\Delta t},i}$. As we expect, these reconstructions are convergent and share their limit with $\pi_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{c}_i$. This is the main purpose of the following lemma.

Lemma 3.4.4. *Let C be as in Proposition 3.4.3. We have:*

$$\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{c}_{m,i} \to c_i \qquad \text{strongly in } L^p, p\in[1,\infty) \qquad \forall i \in [\![0,N]\!], \tag{3.4.15}$$

$$\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\Phi} \to \Phi \qquad \text{strongly in } L^p, p\in[1,\infty), \tag{3.4.16}$$

$$c_{\mathcal{E}_m,\boldsymbol{\Delta t}_m,i} \to c_i \qquad \text{strongly in } L^p, p\in[1,\infty) \qquad \forall i \in [\![1,N]\!], \tag{3.4.17}$$

$$\widetilde{c}_{\mathcal{E}_m,\boldsymbol{\Delta t}_m,i} \to c_i \qquad \text{strongly in } L^p, p\in[1,\infty) \qquad \forall i \in [\![0,N]\!]. \tag{3.4.18}$$

Proof. Equation (3.4.15) is a straightforward consequence of (3.4.11) and the boundedness of $\mathcal{A}$. The proof of (3.4.17) and (3.4.18) rely on the Lemma 3.D.2. Thanks to Proposition 3.4.2, the hypothesis is satisfied with $p, \widetilde{p} = 2$, using (3.4.15), we have the L^1 convergence of the diamond reconstructions. Thanks to the L^∞ bound on the edge concentrations, this result translate in the desired equations. The enhanced convergence of the potential relies on the same ideas as the ones given in the previous proof ([78, Lemma 3.4] and [31, Lemma 9]) to get strong L^2 convergence. This is done following the lines of [29, Proposition 4.5]. □

Finally, we show a weak-convergence property on the gradients of the logarithms:

Lemma 3.4.5. *Let C be as in Proposition 3.4.3. We have:*

$$\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \log(\overline{\boldsymbol{c}}) \to \nabla \log(\overline{c}_m) \qquad \text{weakly in } L^2(Q_T)^d. \tag{3.4.19}$$

Moreover, assuming $\inf c_0 > 0$*, we have :*

$$\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \log(\boldsymbol{c}_{m,0}) \to \nabla \log(c_0) \qquad \text{weakly in } L^2(Q_T)^d. \tag{3.4.20}$$

Proof. Let us start with the proof on equation (3.4.20). By definition (3.4.2), we have:

$$\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \log(\boldsymbol{c}_{m,0}) = \frac{1}{\widetilde{c}_{\mathcal{E}_m,\boldsymbol{\Delta t}_m,0}} \nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \boldsymbol{c}_{m,0},$$

so that, using (3.4.18), (3.4.12), and the assumed bound on c_0 we have:

$$\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \log(\boldsymbol{c}_{m,0}) \to \frac{1}{c_0} \nabla c_0.$$

We conclude using the bound on c_0 again to use the continuous chain-rule and get the announced result.

For (3.4.19), we proceed similarly. Notice that since $\overline{c} \geq \frac{1}{v_0 \max k_i} > 0$ the bound does not need to be assumed. we only need the strong L^2 convergence of the reconstruction using the logarithmic average on the diamond cells. This is an application of Lemma 3.D.2, as in the proof of Lemma 3.4.4. □

3.4.3 Identification

In this section we will identify the limits obtained in Proposition 3.4.3 as weak solutions in the sense of Definition 5. First we improve the convergence topology on the potential and identify it as a weak solution of the Poisson equation.

Proposition 3.4.4. *The function $\Phi \in L^\infty((0,T), H^1(\Omega))$ defined in Proposition 3.4.3 satisfies: $\Phi - \Phi^D \in L^\infty((0,T), H_{\Gamma^D})$ and for all $\psi \in \mathcal{H}_{\Gamma^D}$ and almost all $t \in (0,T)$ equation (3.1.14) holds:*

$$\lambda^2 \int_\Omega \nabla\Phi(t,x)\cdot\nabla\psi(x)\mathrm{d}x = \int_\Omega \psi(x)\sum_{i=1}^N z_i c_i(t,x)\mathrm{d}x.$$

Proof. Let $\psi \in C_c^\infty([0,T]\times\left\{\Omega\cup\Gamma^N\right\})$, then define $\psi_K^n = \psi(x_K,t_n)$ and $\psi_\sigma^n = \psi(x_\sigma,t_n)$ for $1\le n\le N$, $K\in\mathcal{T}$ and $\sigma\in\mathcal{E}_{\text{ext}}$. As for [29, Proposition 4.5], we introduce an other reconstruction of the gradient following [63] (see [52] for a practical example). Let $\widehat{\nabla}_\mathcal{T}:\mathbb{R}^\mathcal{T}\to L^\infty(\Omega)^d$ be strongly consistent i.e.,

$$\widehat{\nabla}_\mathcal{T}\boldsymbol{\psi}^n \underset{h_\mathcal{T}\to 0}{\longrightarrow} \nabla\psi(\cdot,t_n) \text{ uniformly in } \overline{\Omega}, \quad \forall n\in\{1,\dots,N\}, \tag{3.4.21}$$

thanks to the smoothness of ψ. The operator $\widehat{\nabla}$ is also such that

$$\int_\Omega \nabla_\mathcal{T}\boldsymbol{u}\cdot\widehat{\nabla}_\mathcal{T}\boldsymbol{v}\mathrm{d}x = \sum_{\sigma\in\mathcal{E}}\tau_\sigma D_{K\sigma}\boldsymbol{u}D_{K\sigma}\boldsymbol{v}, \qquad \forall \boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^\mathcal{T}.$$

The scheme (3.2.4a) then reduces to

$$\lambda^2\int_\Omega \nabla_\mathcal{T}\boldsymbol{\Phi}^n\cdot\widehat{\nabla}_\mathcal{T}\boldsymbol{\psi}^n\mathrm{d}x = \int_\Omega \pi_\mathcal{T}\boldsymbol{\psi}^n\sum_{i=1}^N z_i\pi_\mathcal{T}\boldsymbol{c}_i^n\mathrm{d}x, \quad \forall n\in[\![1,N]\!],\ \forall\boldsymbol{\psi}\in\mathbb{R}^{(\mathcal{T}\cup\mathcal{E}_{\text{ext}})\times N}.$$

Integrating with respect to time over $(0,T)$ and passing to the limit $h_\mathcal{T}, h_{\boldsymbol{\Delta t}}\to 0$ thanks to Proposition 3.4.3 equations (3.4.11) and (3.4.14) and equation (3.4.21) we have:

$$\lambda^2\iint_{Q_T}\nabla\Phi\cdot\nabla\psi\mathrm{d}x\mathrm{d}t = \iint_{Q_T}\psi\sum_{i=1}^N z_i c_i\mathrm{d}x\mathrm{d}t, \qquad \forall\psi\in C_c^\infty([0,T]\times\Omega\cup\Gamma^N).$$

By density of $C_c^\infty([0,T]\times\Omega\cup\Gamma^N)$ in $L^\infty([0,T],H_{\Gamma^D})$ and continuity of the linear

application, we have:

$$\lambda^2 \iint_{Q_T} \nabla\Phi \cdot \nabla\psi \mathrm{d}x\mathrm{d}t = \iint_{Q_T} \psi \sum_{i=1}^{N} z_i c_i \mathrm{d}x\mathrm{d}t, \qquad \forall \psi \in L^\infty([0,T], H_{\Gamma^D}).$$

In particular, (3.1.14) holds for almost every $t \in (0,T)$.

Concerning the boundary conditions for Φ, the fact that $\Phi = \Phi^D$ on $(0,T)\times\Gamma^D$ can be proved for instance following the lines of [23, Section 4]. □

The following theorem focuses on the identification of C as a weak solution satisfying (3.1.13). As announced in Theorem 3.2.2 this can only be done with an assumption on the solvent. Remark 3.C.1 is a first clue of the validity of this assumption. For positive initial condition, this assumption is valid in all the numerical test. In the 1D setting and under a CFL condition, it might be possible to prove it through improvements of Lemmas 3.3.2 and 3.3.5. This could be the topic of further research.

Theorem 3.4.1. *Let C and Φ be as in Propositions 3.4.3. If one has $\inf c_0 > 0$, they are weak solutions of (3.1.3)–(3.1.8) in the sense of Definition 5.*

Proof. Let $i \in [\![1, N]\!]$, $\varphi \in C_c^\infty([0,T) \times \overline{\Omega})$, then define $\varphi_K^n = \varphi(x_K, t_n)$ for all $n \in \{0, \dots, N_T\}$ and $K \in \mathcal{T}$. Multiplying (3.2.4b) by $\Delta t_n \varphi_K^{n-1}$, then summing over $K \in \mathcal{T}$ and $n \in \{1, \dots, N_T\}$ leads to

$$T_1 + T_2 + T_3 = 0, \tag{3.4.22}$$

where we have set

$$\begin{aligned}
T_1 &= \sum_{n=1}^{N_T} \sum_{K\in\mathcal{T}} m_K (c_{K,i}^n - c_{K,i}^{n-1}) \varphi_K^{n-1}, \\
T_2 &= \sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \mathcal{C}_{\sigma,i}^n D_{K\sigma} h_i(\boldsymbol{c}^n) D_{K\sigma} \boldsymbol{\varphi}^{n-1}, \\
T_3 &= z_i \sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \mathcal{C}_{\sigma_i}^n D_{K\sigma} \boldsymbol{\Phi}^n D_{K\sigma} \boldsymbol{\varphi}^{n-1},
\end{aligned}$$

where $\varphi_{K\sigma}^{n-1} = 0$ for $\sigma \in \mathcal{E}_{\mathrm{ext}}$ and $\mathcal{C}_\sigma$ is defined by Lemma 3.3.1. The treatment of

terms T_1 and T_3 is exactly the same as in [29, Proposition 4.7] and we have:

$$T_1 \underset{m\to\infty}{\longrightarrow} -\iint_{Q_T} c_i \partial_t \varphi \mathrm{d}x\mathrm{d}t - \int_\Omega c_i^0 \varphi(0,\cdot)\mathrm{d}x, \tag{3.4.23}$$

$$T_3 \underset{m\to\infty}{\longrightarrow} \iint_{Q_T} z_i c_i \nabla\Phi\cdot\nabla\varphi \mathrm{d}x\mathrm{d}t. \tag{3.4.24}$$

The treatment of the term T_2 is more intricate. First we let $\widetilde{T}_2$ be the same term with a different edge concentration:

$$\widetilde{T}_2 = \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \widetilde{\mathcal{C}}_{\sigma,i}^n D_{K\sigma} h_i(\boldsymbol{c}^n) D_{K\sigma}\boldsymbol{\varphi}^{n-1},$$

where $\widetilde{\mathcal{C}}$ is the logarithmic mean introduced in (3.4.2). We will first prove the convergence of $\widetilde{T}_2$ then identify its limit. To this end, we set:

$$\begin{aligned}
\widetilde{T}_{2,1} =& \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \widetilde{\mathcal{C}}_{\sigma,i}^n D_{K\sigma} \log(\boldsymbol{c}_i^n) D_{K\sigma}\boldsymbol{\varphi}^{n-1},\\
\widetilde{T}_{2,2} =& -k_i \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \widetilde{\mathcal{C}}_{\sigma,i}^n D_{K\sigma} \log(\boldsymbol{c}_0^n) D_{K\sigma}\boldsymbol{\varphi}^{n-1},\\
\widetilde{T}_{2,3} =& (k_i-1) \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \widetilde{\mathcal{C}}_{\sigma,i}^n D_{K\sigma} \log(\overline{\boldsymbol{c}}^n) D_{K\sigma}\boldsymbol{\varphi}^{n-1}.
\end{aligned}$$

For term $\widetilde{T}_{2,1}$ we use the chain rule $\widetilde{\mathcal{C}}_{\sigma,i}^n D_{K\sigma}\log(\boldsymbol{c}_i^n) = D_{K\sigma}\boldsymbol{c}_i^n$ and get :

$$\widetilde{T}_{2,1} = \sum_{n=1}^{N} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma}\boldsymbol{c}_i^n D_{K\sigma}\boldsymbol{\varphi}^{n-1} = \iint_{Q_T} \nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\cdot \boldsymbol{c}_i \widehat{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\varphi}\mathrm{d}c\mathrm{d}t.$$

Thanks to the weak convergence of $\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{c}_i$ and the strong convergence of $\widehat{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{\varphi}$, we have:

$$\widetilde{T}_{2,1} \to \iint_{Q_T} \nabla c_i\cdot\nabla\varphi \mathrm{d}x\mathrm{d}t.$$

For the other terms, we need the enhanced convergence of gradients provided

by Lemma 3.4.5. So that the terms $\widetilde{T}_{2,2}$ and $\widetilde{T}_{2,3}$ have the following limits:

$$\widetilde{T}_{2,2} = -k_i \iint_{Q_T} \widetilde{c}_{\mathcal{E}_m,\boldsymbol{\Delta t}_m,i} \nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \log(\boldsymbol{c}_0) \widehat{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \boldsymbol{\varphi} \mathrm{d}x\mathrm{d}t$$
$$\to -k_i \iint_{Q_T} c_i \nabla \log(c_0) \nabla \varphi \mathrm{d}x\mathrm{d}t,$$

$$\widetilde{T}_{2,3} = (k_i - 1) \iint_{Q_T} \widetilde{c}_{\mathcal{E}_m,\boldsymbol{\Delta t}_m,i} \nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \log(\overline{\boldsymbol{c}}) \widehat{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \mathrm{d}x\mathrm{d}t \boldsymbol{\varphi}$$
$$\to (k_i - 1) \iint_{Q_T} c_i \nabla \log(\overline{c}) \nabla \varphi \mathrm{d}x\mathrm{d}t.$$

Let us now establish that T_2 and $\widetilde{T}_2$ share the same limit. Thanks to the triangle and Cauchy-Schwarz inequalities, one has

$$|T_2 - \widetilde{T}_2| \leq \sum_{n=1}^{N} \Delta t_n \sum_{\sigma \in \mathcal{E}} \tau_\sigma \left| \mathcal{C}_{\sigma,i}^n - \widetilde{\mathcal{C}}_{\sigma,i}^n \right| |D_\sigma h_i(\boldsymbol{c}^n)| \left| D_\sigma \boldsymbol{\varphi}^{n-1} \right|$$
$$\leq \left(\sum_{n=1}^{N} \Delta t_n \sum_{\sigma \in \mathcal{E}} \tau_\sigma \mathcal{C}_{\sigma,i}^n |D_\sigma h(\boldsymbol{c}^n)|^2 \right)^{1/2} \left(\sum_{n=1}^{N} \Delta t_n \sum_{\sigma \in \mathcal{E}} \tau_\sigma \frac{(\mathcal{C}_{\sigma,i}^n - \widetilde{\mathcal{C}}_{\sigma,i}^n)^2}{\mathcal{C}_{\sigma,i}^n} |D_\sigma \boldsymbol{\varphi}^{n-1}|^2 \right)^{1/2}.$$

The first term in the right-hand side is uniformly bounded thanks to (3.4.10). Thus our problem amounts to show that

$$\mathcal{R} := \sum_{n=1}^{N} \Delta t_n \sum_{\sigma \in \mathcal{E}} \tau_\sigma \frac{(\mathcal{C}_{\sigma,i}^n - \widetilde{\mathcal{C}}_{\sigma,i}^n)^2}{\mathcal{C}_{\sigma,i}^n} |D_\sigma \boldsymbol{\varphi}^{n-1}|^2 \underset{m \to \infty}{\longrightarrow} 0. \tag{3.4.25}$$

Let us reformulate $\mathcal{R}$ as

$$\mathcal{R} = \sum_{n=1}^{N} \Delta t_n \sum_{\sigma \in \mathcal{E}} \tau_\sigma |\mathcal{C}_{\sigma,i}^n - \widetilde{\mathcal{C}}_{\sigma,i}^n| \left| 1 - \frac{\widetilde{\mathcal{C}}_{\sigma,i}^n}{\mathcal{C}_{\sigma,i}^n} \right| |D_\sigma \boldsymbol{\varphi}^{n-1}|^2.$$

Thanks to Lemma 3.4.1, the quantity $\left| 1 - \frac{\widetilde{\mathcal{C}}_{\sigma,i}^n}{\mathcal{C}_{\sigma,i}^n} \right|$ is uniformly bounded, whereas the regularity of φ implies that $D_\sigma \boldsymbol{\varphi}^{n-1} \leq \|\nabla \varphi\|_\infty d_\sigma$. Putting this in the above expression of $\mathcal{R}$, we obtain that

$$0 \leq \mathcal{R} \leq C \| c_{\mathcal{E}_m,\boldsymbol{\Delta t}_m,i} - \widetilde{c}_{\mathcal{E}_m,\boldsymbol{\Delta t}_m,i} \|_{L^1(Q_T)} \underset{m \to \infty}{\longrightarrow} 0,$$

thanks to Lemma 3.4.4. Thus T_2 and $\widetilde{T}_2$ share the same limit, which gives the announced result. □

3.5 Numerical Examples

The numerical examples have been implemented in the Julia language [17] based on the package `VoronoiFVM.jl` [73] which realizes the implicit Euler Voronoi finite volume method for nonlinear diffusion-convection-reaction systems on simplicial grids. The resulting nonlinear systems of equations are solved using Newton's method with optional parameter embedding. An advantage of the implementation in Julia is the availability of `ForwardDiff.jl` [115], an automatic differentiation package. This package allows the assembly of analytical Jacobians based on a generic implementation of nonlinear parameter functions without the need to write source code for derivatives.

3.5.1 Species redistribution in a one-dimensional cell filled with binary electrolyte

Let $\Omega = (0, L)$ with $L = 20$. As an initial state, assume a binary electrolyte with two ionic species with opposite charges and a solvent. At moment $t = 0$, we assume a spatially constant, electroneutral distribution of the ions. We apply a potential difference via Dirichlet boundary conditions $\Phi|_{x=0} = -10$ and $\Phi_{x=L} = 10$ and solve the Poisson equation with these data as initial value. We set homogeneous Neumann boundary conditions for both ionic species. With starting time step size $\Delta t = 10^{-3}$ we start the evolution until the species distribution reaches its equilibrium under the applied potential difference. As discussed in [30], the time step sizes are controlled such that the energy dissipation per time step is limited: $E(t_i) - E(t_{i+1}) \leq 10^{-1}$.

Fig. 3.2 shows the evolution in the case $v_0 = v_1 = v_2 = 1$, $z_0 = 0, z_1 = 1, z_2 = -1$. At the end of the time evolution, most of the ions are accumulated in their respective polarization boundary layers, almost completely displacing the solvent. As predicted, the ion concentration is bounded by 1. The computation used the flux (S).

Fig. 3.3 shows the evolution in the case $v_0 = v_1 = v_2 = 1$ and $z_0 = 0, z_1 = 2, z_2 = -1$. Once again, at the end of the evolution, anions and cations pile up in the corresponding boundary layers. Ion concentrations are bounded by 1, but due to the larger charge of the cation, the corresponding boundary layer becomes smaller.

Fig. 3.4 shows the evolution in the case $v_0 = v_2 = 1$, $v_1 = 2$ and $z_0 = 0, z_1 = 1, z_2 = -1$. Once again, at the end of the evolution, anions and cations pile up in

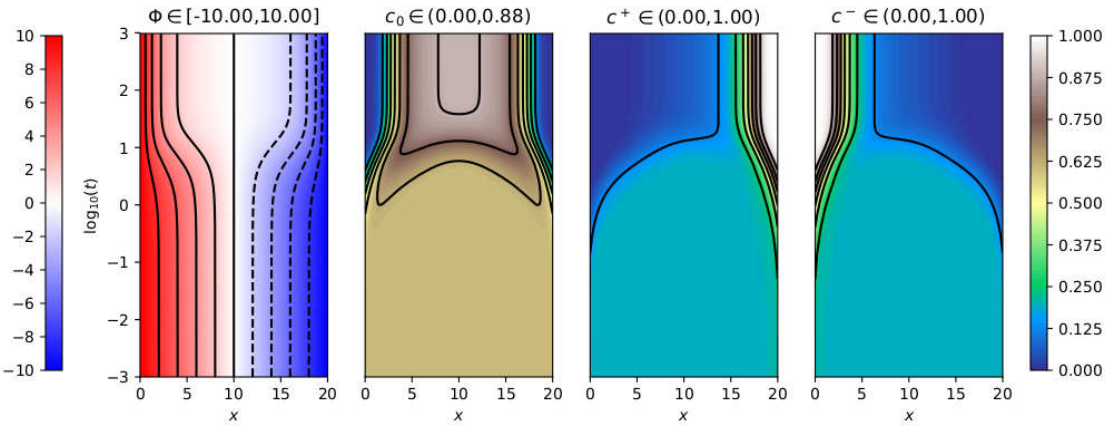


Figure 3.2 – Evolution of electrostatic potential Φ, solvent concentration c_0, anion concentration c^- and cation concentration c^+ for a symmetric binary electrolyte with equal sizes of solvent molecules, anions and cations.

the corresponding boundary layers but now, the cation concentration is bounded by $\frac{1}{2}$. The corresponding evolution of the relative free energy $E(t) - E_\infty$ is shown in Fig. 3.5. We observe an exponential decay and almost equal behavior for both variants of the flux approximation (S) and (C). Moreover, the time step control algorithm keeps the dissipation per timestep below the intended limit.

3.5.2 1D stationary convergence test

In the same domain as above, we set $v_0 = 1, v_1 = 2, v_2 = 1$, and $z_1 = 1, z_2 = -1$. This time, we look for the stationary solution with homogeneous Dirichlet boundary conditions for Φ, and Dirichlet boundary conditions for the concentrations. These boundary conditions are for $x = 0$, $c_1 v_1 = 1.0 - 3\epsilon, c_2 v_2 = \epsilon$ and for $x = L$, $c_1 v_1 = \epsilon, c_2 v_2 = 1 - 3\epsilon$, where $\epsilon = 10^{-2}$. Implicitely, this sets $c_0 = 2\epsilon$ at both boundaries. The result of the numerical convergence tests (comparison to fine grid solution with 40960 grid points) for both types of fluxes suggest $O(h^2)$ convergence in the L^2 norm and $O(h)$ convergence in the H^1 seminorm.

3.5.3 An electrolytic diode

The second example regards a domain $\Omega = (0, W) \times (0, L)$ with $W = 2$ and $L = 10$. We assume $z_0 = 0, z_1 = 1, z_2 = -1$ and $v_0 = 1, v_1 = 4, v_2 = 4$. At $y = 0$ and $y = L$ we fix concentrations to a value $c_1 = c_2 = 0.01$ We set $\Phi|_{y=0} = 0$ and apply a changing value Φ_{bias} at $y = L$. At $x = 0$ we apply symmetry (homogeneous Neumann) boundary conditions for Φ, c_1, c_2. Homogeneous Neumann

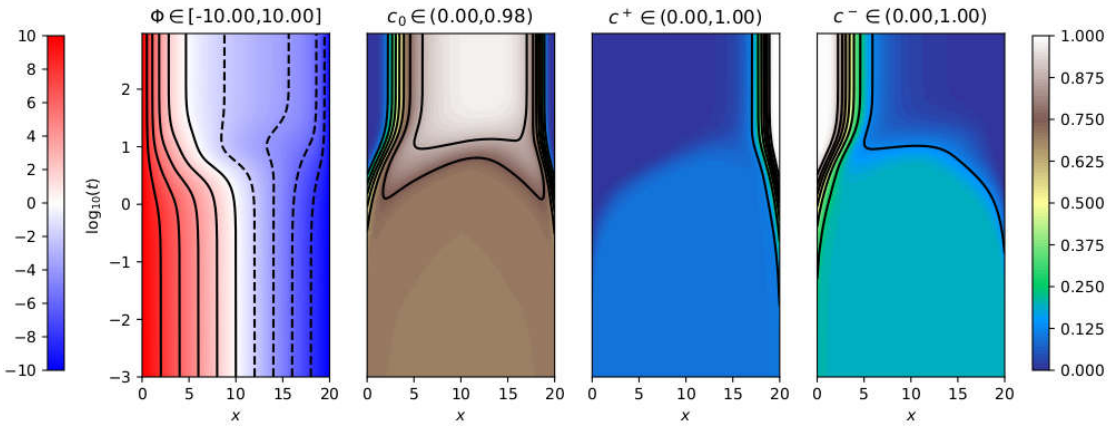


Figure 3.3 – Evolution of electrostatic potential Φ, solvent concentration c_0, anion concentration c^- and cation concentration c^+ for an asymmetric binary electrolyte with equal sizes of solvent molecules, cations and anions.

boundary conditions are also applied for c_1, c_2 at $x = W$. We set Neumann boundary conditions $\lambda \nabla \Phi \cdot n = q(y)$ at $x = W$, where

$$q(y) = \begin{cases} \sigma, & y \in (\frac{1}{2}L, \frac{3}{4}L) \\ -\sigma, & y \in (\frac{1}{4}L, \frac{1}{2}L) \\ 0, & \text{else} \end{cases}$$

with $\sigma = 5$.

Fig. 3.7 shows three different states of the electrolytic diode. Fig. 3.8 (left) shows the corresponding current-voltage curve. We see a well developed rectification effect: At reverse bias, ion concentrations under the charged surface are rather low, resulting in low conductance and low ionic current. Whereas at forward bias, larger ion concentrations lead to a larger ionic current.

Fig. 3.8 (right) shows the estimated error of the IV curve in dependence of the grid refinement. Reference was a calculation on a grid with the quarter of the stepsize of the finest grid result shown. From this experiment, we postulate a convergence rate for the ionic current calculation of $O(h^2)$.

Acknowledgement

The authors are grateful to C. Chainais-Hilliaret and C. Cances for productive discussions and careful reviewing of this manuscript. Their attention to the size-interest-readability trade-offs contributed a lot to the readability of this paper.

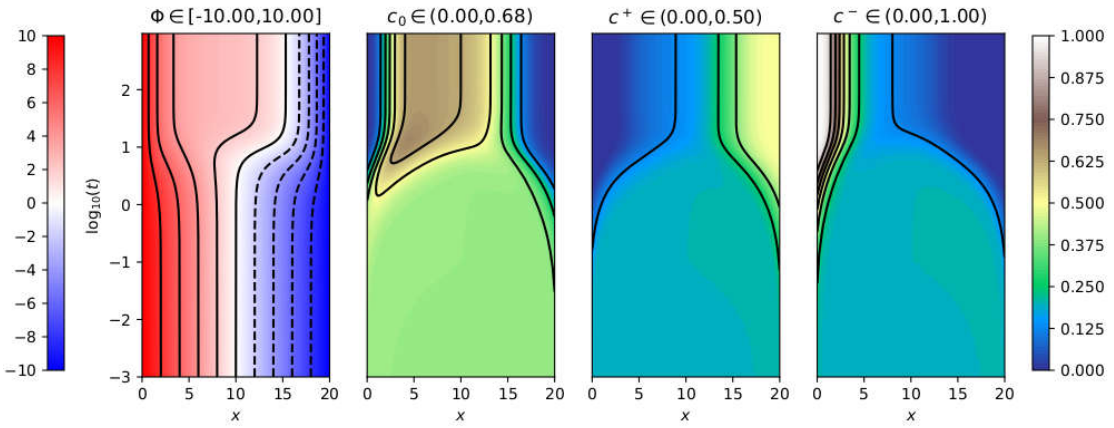


Figure 3.4 – Evolution of electrostatic potential Φ, solvent concentration c_0, anion concentration c^- and cation concentration c^+ for a symmetric binary electrolyte with equal sizes of solvent molecules and anions, but larger cations.

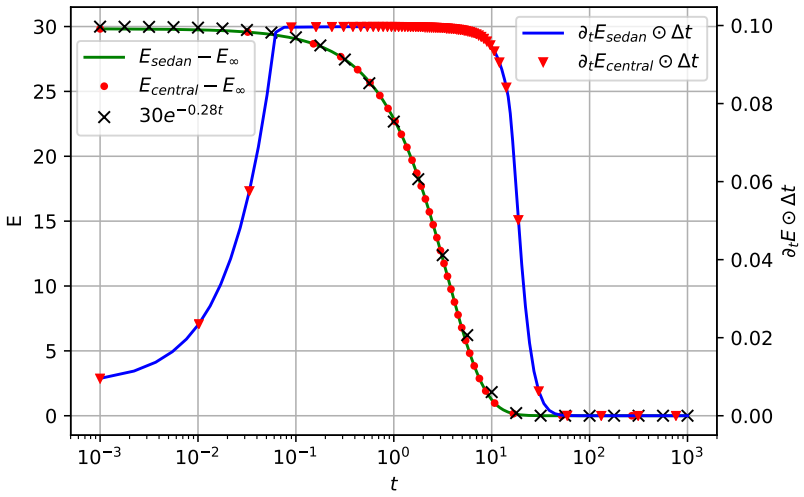


Figure 3.5 – Evolution of relative free energy and energy dissipation per time step for symmetric binary electrolyte with equal sizes of solvent molecules and anions, but larger cations.

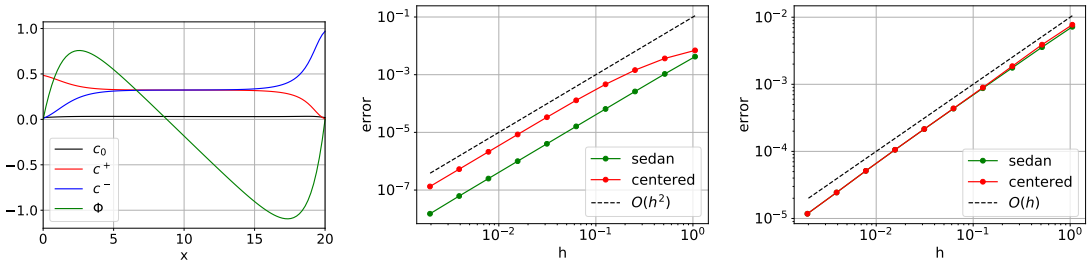


Figure 3.6 – Left: stationary solution of Dirichlet problem. Center and right: results of numerical convergence test.

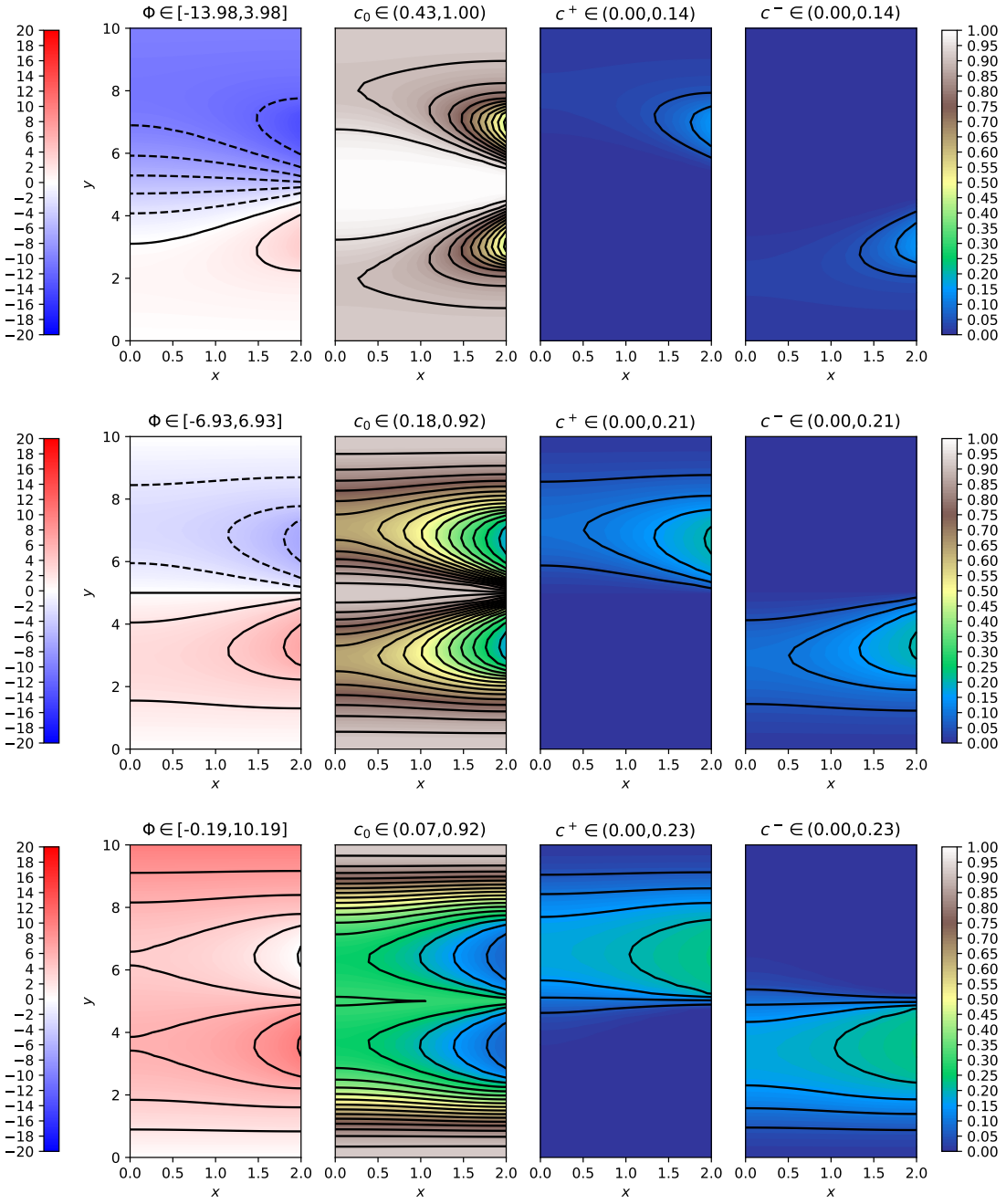


Figure 3.7 – Electrostatic potential Φ, solvent concentration c_0, anion concentration c^- and cation concentration c^+ in an electrolytic diode filled with a symmetric binary electrolyte with equal sizes of solvent molecules at reverse bias $\Phi_{bias} = -10$ (top), zero bias $\Phi_{bias} = 0$ (center) and forward bias $\Phi_{bias} = 10$ (bottom)

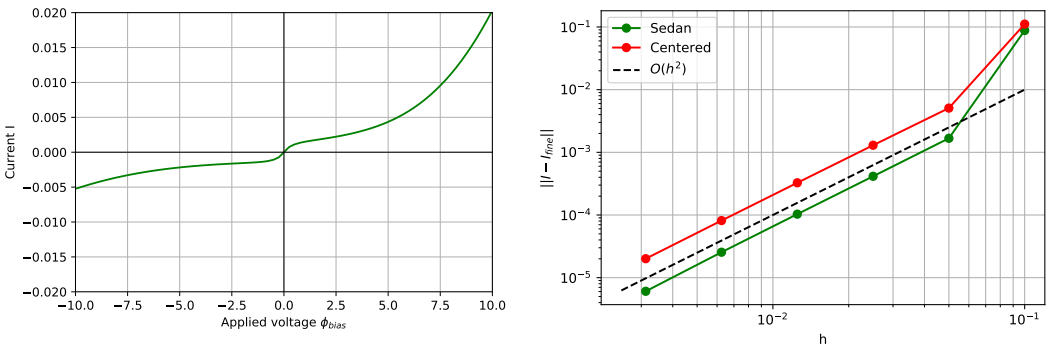


Figure 3.8 – Left: Current-voltage curve for electrolytic diode, calculated using the scheme (S). Right: Convergence of calculated IV curve.

3.A Chemical free energy density and chemical potentials

In this appendix, we aim to prove (3.1.9), (3.1.10), and some convexity of:

$$H(c_1, ...c_N) = -\overline{c}\log(\overline{c}) + \sum_{i=0}^{N} c_i \log(c_i),$$

where c_0 and $\overline{c}$ are functions of $c_1, ..., c_N$. This is summarized in the following lemma:

Lemma 3.A.1. *The N-variables function H is convex, moreover we have:*

$$\partial_{c_i} H(c_1, ..., c_N) = h_i(C), \qquad \forall i \in [\![1, N]\!],\ C = (c_1, ..., c_N) \in \mathcal{A}, \tag{3.1.9}$$

$$\frac{-\log(N+1)}{v_0 \min k_i} \leq H(C) \leq 0 \qquad \forall C \in \mathcal{A}. \tag{3.1.10}$$

Elementary computation shows that :

$$\partial_{c_i} H(C) = \log\frac{c_i}{\overline{c}} - k_i \log\frac{c_0}{\overline{c}} \qquad \forall i \in [\![1, N]\!].$$

Hence the announced relation (3.1.9).

We now focus on the proof of the convexity of H over $\mathcal{A}$. Let $C, C^* \in \mathcal{A}$, we have:

$$(\nabla_{\mathbb{R}^N} H(C) - \nabla_{\mathbb{R}^N} H(C^*)|C - C^*) = \sum_{i=0}^{N} \left(\log\frac{c_i}{\overline{c}} - \log\frac{c_i^*}{\overline{c}^*}\right)(c_i - c_i^*). \tag{3.A.1}$$

To prove the convexity of H, it is sufficient to show that this is non-negative. To that extend, we introduce $\mathcal{A}_{N+1}$ the natural extension of $\mathcal{A}$ in $\mathbb{R}^{N+1}$ and consider the right-hand side of (3.A.1) as a function of $C_{N+1} = (c_0(c_1, ...c_N), c_1..., c_N) \in \mathcal{A}_{N+1}$ parameterized by C^*:

$$G_{C^*}(c_0, \ldots, c_N) = \sum_{i=0}^{N} \left(\log\frac{c_i}{\overline{c}} - \log\frac{c_i^*}{\overline{c}^*}\right)(c_i - c_i^*),$$

and show that $\min_{C_{N+1} \in \mathcal{A}_{N+1}}(G_{C^*}(C)) = 0$. To do so we compute the derivatives of G_{C^*} as a function of $\mathbb{R}^{N+1}$ and use the Lagrange multiplier theorem. After some

simplifications, we have for all $i \in [\![0, N]\!]$:

$$\partial_{c_i} G_{C^*}(c_0, ...c_N) = \frac{\overline{c}^*}{c_i}\left(\frac{c_i}{\overline{c}} - \frac{c_i^*}{\overline{c}^*}\right) + \left(\log\frac{c_i}{\overline{c}} - \log\frac{c_i^*}{\overline{c}^*}\right)$$

Notice that both terms have the sign of $\frac{c_i}{\overline{c}} - \frac{c_i^*}{\overline{c}^*}$. The Lagrange multiplier theorem states that any extremum satisfies:

$$\exists \alpha \in \mathbb{R}, \forall i \in [\![0, N]\!], \qquad \partial_{c_i} G_{C^*} = \alpha k_i$$

Hence, all the partial derivatives of G_{C^*} should have the same sign. Moreover, we notice that the sum of $\frac{c_i}{\overline{c}} - \frac{c_i^*}{\overline{c}^*}$ is zero. This is only possible the sign of the derivatives is constantly zero, i.e. : $\frac{c_i}{\overline{c}} = \frac{c_i^*}{\overline{c}^*}$. At such a point, we have $G_{C^*} = 0$. As the coercitivity and continuity of G_{C^*} grants the existence of a minimum, we have the desired result:

$$0 \leq (\nabla_{\mathbb{R}^N} H(C) - \nabla_{\mathbb{R}^N} H(C^*) | C - C^*),$$

which yields the convexity of H.

We still have to establish the bounds (3.1.10). To that end, we notice that:

$$H(C) = \overline{c} \sum_{i=0}^{N} \frac{c_i}{\overline{c}} \log\frac{c_i}{\overline{c}} \qquad \forall C \in \mathcal{A}.$$

As $\overline{c}$ is non-negative and $0 \leq \frac{c_i}{\overline{c}} \leq 1$, we have $H(C) \leq 0$. For the lower bound, we notice that $-\sum_{i=0}^{N} \frac{c_i}{\overline{c}} \log\frac{c_i}{\overline{c}}$ can be interpreted as the entropy of a random variable over a set of $N+1$ elements. It is common knowledge that it is maximal for $\frac{c_i}{\overline{c}} = \frac{1}{N+1}$ thus:

$$-\overline{c}\log(N+1) \leq H(C)$$

Finally, notice that $\frac{1}{v_0 \max k_i} \leq \overline{c} \leq \frac{1}{v_0 \min k_i}$ yields

$$\frac{-\log(N+1)}{v_0 \min k_i} \leq H(C),$$

which is the desired bound.

3.B Proof of Lemma 3.3.2

This appendix is devoted to the proof of Lemma 3.3.2 stating the blow-up of the diffusion for extreme concentrations. More precisely, we recall:

Lemma 3.B.1. *Let for $\delta, \epsilon, M, c > 0$, $i \in [\![1, N]\!]$:*

$$\Psi_{\delta,\epsilon,M,i}(c) := \inf_{\substack{(C_K,C_L)\in\mathcal{A}^2,\ (\Phi_K,\Phi_L)\in[-M,M]^2 \\ c_{K,0},c_{L,0}>\epsilon,\ c_{K,i}\geq\min(\delta,\frac{0.5}{k_i v_0}),\ c_{L,i}<c}} \mathcal{D}_i(C_K, C_L, \Phi_K, \Phi_L),$$

$$\Upsilon_{\delta,M}(c) := \inf_{\substack{(C_K,C_L)\in\mathcal{A}^2,\ (\Phi_K,\Phi_L)\in[-M,M]^2 \\ c_{K,0}\geq\min(\delta,\frac{0.5}{v_0}),\ c_{L,0}<c}} \mathcal{D}(C_K, C_L, \Phi_K, \Phi_L).$$

We have, for all $\delta, \epsilon, M > 0$:

$$\lim_{c\to 0^+} \Upsilon_{\delta,M}(c) = +\infty \qquad \lim_{c\to 0^+} \Psi_{\delta,\epsilon,M,i}(c) = +\infty \qquad \forall i \in [\![1, N]\!].$$

We will prove the result for $\Psi_{\delta,\epsilon,M,i}$ first, then use this property to show the bound on the solvent.

3.B.1 Limit of $\Psi_{\delta,\epsilon,M,i}$

In this section we intend to prove the limit:

$$\lim_{\substack{c\to 0 \\ c>0}} \Psi_{\delta,\epsilon,M,i}(c) = +\infty \qquad \forall i \in [\![1, N]\!],\ \delta, \epsilon, M > 0.$$

The proof for the centered scheme relies on expression (3.3.5):

$$\mathcal{D}_i(C_K, C_L, \Phi_K, \Phi_L) = \mathcal{C}_i(C_K, C_L, \Phi_K, \Phi_L)\left(h_i(C_K) + z_i\Phi_K - h_i(C_L) - z_i\Phi_L\right)^2.$$

We notice that $h_i(C_K) + z_i\Phi_K - h_i(C_L) - z_i\Phi_L$ blows up and that $\mathcal{C}_i \geq \frac{c_{K,i}}{2}$, hence the blow-up of the limit.

For the Sedan scheme, it is more intricate. We bound $\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L)$ away from zero to take advantage of the blow-up of $(h_i(C_K) + z_i\Phi_K - h_i(C_L) - z_i\Phi_L)$. The positivity of the product, ensures that the limit will have the right sign. Let $\delta, \epsilon, M > 0$, $i \in [\![1, N]\!]$. We denote by O_c the set:

$$O_c = \left\{(C_K, C_L) \in \mathcal{A}^2,\ (\Phi_K, \Phi_L) \in [-M, M]^2 \mid c_{K,0}, c_{L,0} > \epsilon,\ c_{K,i} \in [\delta, 1),\ c_{L,i} < c\right\}.$$

We notice that the hypothesis $c_0 > \epsilon$ yields a bound on ν_i. Moreover, this bound is uniform in c. We intend to use this bound to prove that the flux function defined by (S) is bounded away from zero. We let:

$$M' = \sup_{c\in\mathbb{R}_{+,*}} \left(\sup_{(C_K,C_L,\Phi_K,\Phi_L)\in O_c} z_i\Phi_L + \nu(C_{L,i}) - z_i\Phi_K - \nu(C_{K,i}) \right) < \infty.$$

We have, for all $(C_K, C_L, \Phi_K, \Phi_L) \in O_c$:

$$\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) \geq B(M')\delta - B(-M')c,$$

hence $\mathcal{F}_i$ is bounded away from zero for c small enough and the desired result.

3.B.2 Limit of $\Upsilon_{\delta,M}$

In this section, we prove the remaining limit:

$$\lim_{c\to 0} \Upsilon_{\delta,M}(c) = +\infty \qquad \forall \delta, M > 0$$

To reuse the ideas of previous section, we would like to dispose of a specie i such that $c_{L,i} > \epsilon$. We start by building one artificially. Let $\delta, M > 0$, and:

$$O(c) = \left\{(C_K, C_L) \in \mathcal{A}^2,\ (\Phi_K, \Phi_L) \in [-M, M]^2 \mid c_{K,0} \in [\delta, 1),\ c_{L,0} < c\right\}.$$

Notice that for all $(C_K, C_L, \Phi_K, \Phi_L) \in O_c$, we dispose of $i \in [\![1, N]\!]$ such that $c_{L,i} \geq \frac{1-v_0 c}{N k_i v_0}$. Notice also that $\Upsilon_{\delta,M}$ is increasing, it is then sufficient to prove the limit for a given sequence. Let c^n be sequence that steadily decreases to zero such that for all $n \in \mathcal{N}$, $c^n \leq \frac{1}{2v_0}$ and there exist $i \in [\![1, N]\!]$, $(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) \in O_{c^n}$ satisfying:

$$\mathcal{D}(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) \leq \Upsilon_{\delta,M}(c^n) + \frac{1}{n} \qquad \text{and} \qquad c_{L,i}^n \geq \frac{1}{2Nk_i v_0}.$$

We have, using $c_{L,0}^n \leq c^n$, $c_{K,i}^n \leq \frac{1}{k_i v_0}$, $\frac{1}{v_0 \max k_j} \leq \overline{c} \leq \frac{1}{v_0 \min k_j}$, $c_{K,0}^n \geq \delta$, the bounds on Φ, and $c_{L,i}^n \geq \frac{1}{2Nk_i v_0}$:

$$h_i(C_K^n) + z_i\Phi_K^n - h_i(C_L^n) - z_i\Phi_L^n \leq k_i \log\frac{c^n}{\delta} + |k_i - 1| \log\frac{\max k_j}{\min k_j} + \log(2N) + 2M|z_i|.$$

As all the terms are bounded except $\log(c^n)$ which goes to $-\infty$, we have blow-up of $h_i(C_K^n) + z_i\Phi_K^n - h_i(C_L^n) - z_i\Phi_L^n$.

For the centered scheme, we use $\mathcal{C}_i(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) \geq \frac{1}{4Nv_i}$, and we have :

$$\Upsilon_{\delta,M}(c^n) \geq \mathcal{D}_i(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) - \frac{1}{n} \geq \frac{1}{4Nv_i}\left(h_i(C_K^n) + z_i\Phi_K^n - h_i(C_L^n) - z_i\Phi_L^n\right)^2 - \frac{1}{n},$$

hence the desired result.

For the Sedan scheme, we will also only consider $\mathcal{D}_i$, but we need a more precise

approach : as in previous section, we bound the flux away from zero. We let:

$$M' = \sup_{c\in(0,\frac{1}{2v_0}]} \left(\sup_{(C_K,C_L,\Phi_K,\Phi_L)\in O_{c^0}, c_{L,i}\geq \frac{1}{2Nv_i}} z_i\Phi_L + (k_i-1)\log \overline{c_L} - z_i\Phi_K - \nu(C_{K,i}) \right).$$

We have:

$$\mathcal{F}_i(C_K^n, C_L^n, \Phi_K^n, \Phi_L^n) \leq B\left(-k_i \log c^n - M'\right)\frac{1}{v_i} - B\left(k_i \log c^n + M'\right)\frac{1}{2Nv_i}.$$

As the right-hand side tends to $-\infty$, the left-hand side is bounded away from zero. Using the previously detailed arguments, we have the desired limit.

3.C Study of a numerical scheme for $h_i = \log(c_i) - \alpha \log(c_0)$

To prove the existence of solutions to the Sedan and centered scheme, we introduce this simplified cross diffusion system where the coupling occurs only through the solvent using the chemical potential defined above. This system is discretized using the ideas of the centered scheme and [34]. In detail, we use equation (3.2.4b), (3.2.4c) with $k_i, D_i = 1$, $z_i = 0$, and :

$$\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) = \widetilde{C}_j(C_K, C_L)\left(h_i(C_K) - h_i(C_L)\right),$$
$$\widetilde{C}_j(C_K, C_L) = \frac{c_{K,i} - c_{L,i}}{\log(c_{K,i}) - \log(c_{L,i})}$$

where $h_i(C)$ is: $\log(c_i) - \alpha \log(c_0)$. We want to bound the concentrations away from zero uniformly in α. This is the meaning of the following lemma, which is highly inspired by Lemma 3.3.5.

Lemma 3.C.1. *There exist $\epsilon = \min(\epsilon_0, \epsilon_1, \ldots, \epsilon_N) > 0$ depending on, among other things, C^0 and decreasing with $h_{\boldsymbol{\Delta t}}$ and $\min_{\sigma\in\mathcal{E}} \tau_\sigma$ such that for all $\boldsymbol{C}^{n-1} \in \mathcal{A}^{\mathcal{T}}$ satisfying Lemma 3.3.3, $\alpha \in [0,1]$, we have:*

$$c_{K,i}^n \geq \epsilon_i \qquad \forall K, i$$

The proof follows the same reasoning as for the full system and is only sketched

here. Using (3.2.4c) we have:

$$\frac{c_{K,0}^n - c_{K,0}^{n-1}}{\Delta t_n} m_K = -\sum_{\sigma=K|L\in\mathcal{E}_{\text{int}}} \tau_\sigma \left(c_{K,0}^n - c_{L,0}^n\right) - \alpha \sum_{\sigma=K|L\in\mathcal{E}_{\text{int}}} \tau_\sigma \mathcal{R}(C_K^n, C_L^n) \left(\log(c_{K,0}^n) - \log(c_{L,0}^n)\right) \tag{3.C.1}$$

where we have set:

$$\mathcal{R}(C_K, C_L) = \sum_{i=1}^N \frac{c_{K,i} - c_{L,i}}{\log(c_{K,i}) - \log(c_{L,i})}.$$

Remark 3.C.1. Noticing that $\mathcal{R}(C_K, C_L) \geq 0$ yields a maximum principle on c^0. As we did not assume that c_0^0 is uniformly positive, we have to compute further.

Multiplying (3.C.1) by $\log(c_{K,0}^n)$ and summing over $K \in \mathcal{T}$ yields:

$$\sum_{K\in\mathcal{T}} \frac{c_{K,0}^n - c_{K,0}^{n-1}}{\Delta t_n} m_K \log(c_{K,0}^n) + \sum_{\sigma=K|L\in\mathcal{E}_K} \tau_\sigma D_{KL}\boldsymbol{c}_0^n D_{KL} \log \boldsymbol{c}_0^n + \alpha \sum_{\sigma=K|L\in\mathcal{E}_K} \tau_\sigma \mathcal{R}(C_K, C_L) \left(D_{KL} \log \boldsymbol{c}_0^n\right)^2 = 0$$

using the convexity of $u \log u$, we have:

$$\sum_{K\in\mathcal{T}} \frac{m_k}{\Delta t_n} (c_{K,0}^n \log(c_{K,0}^n) - c_{K,0}^{n-1} \log(c_{K,0}^{n-1})) \leq -\sum_{\sigma=K|L\in\mathcal{E}_K} \tau_\sigma D_{KL}\boldsymbol{c}_0^n D_{KL} \log \boldsymbol{c}_0^n - \alpha \sum_{\sigma=K|L\in\mathcal{E}_K} \mathcal{R}(C_K, C_L) \left(D_{KL} \log \boldsymbol{c}_0^n\right)^2.$$

We may now use the decay of this entropy to prove the desired result for $i = 0$. To that extent we proceed as in Lemma 3.3.5 and see that $D_{KL}\boldsymbol{c}_0^n D_{KL} \log \boldsymbol{c}_0^n$ is clearly coercive in the sense of lemma 3.3.2 while the part in α is non-negative. This yields the uniform bound for $\boldsymbol{c}_0$.

The bound for $\boldsymbol{c}_i$ relies on the entropy $\widetilde{H} = \sum_{j=1}^N c_j \log(c_j) + \alpha c_0 \log(c_0)$. As for Lemma 3.A.1, this entropy restricted to $\mathcal{A}$ is convex and its derivatives as a function of $\mathbb{R}^N$ are the chemical potentials, Thus multiplying the conservation equation by h yields:

$$\sum_{K\in\mathcal{T}} m_K \left(\widetilde{H}(C_K^n) - \widetilde{H}(C_K^{n-1})\right) \leq \sum_{\sigma\in\mathcal{E}_{\text{int}}} \sum_{j=1}^N \widetilde{C}_{\sigma,j} \left(D_\sigma h_j(\boldsymbol{C}^n)\right)^2.$$

This new dissipation is also coercive in the sense of Lemma 3.3.2, thus we can proceed as in Lemma 3.3.5 to get the announced bounds.

3.D A simple convergence lemma

In this section, we express the results of Chapter 2 lemma 4.2 and [34, lemma 4.2] in a more generic fashion. We let $\mathcal{T}_m$ be a sequence of admissible meshes of Ω such that $h_{\mathcal{T}_m} \to 0$, $\boldsymbol{u}_m \in \mathbb{R}^{\mathcal{T}_m}$, and $\widetilde{\boldsymbol{u}}_m \in \mathbb{R}^{\mathcal{E}_m^{\rm int}}$ such that for all $\sigma = K|L \in \mathcal{E}_m^{\rm int}$:

$$\min(u_K, u_L) \leq \widetilde{u}_{\sigma,m} \leq \max(u_K, u_L).$$

Lemma 3.D.1. *If we dispose of $p \in [1, \infty)$ such that*

$$h_{\mathcal{T}_m} \|\nabla_{\mathcal{T}_m} \boldsymbol{u}_m\|_{L^p(\Omega)} \to 0$$

The L^1 convergence of the natural and diamond reconstructions are equivalent, moreover if one of them is convergent, they share the same limit.

Proof. This result is equivalent to:

$$\|\pi_{\mathcal{T}_m} \boldsymbol{u}_m - \widetilde{\boldsymbol{u}}_{m,\mathcal{E}_m}\|_{L^1(\Omega)} \to 0.$$

For the sake of simplicity, we drop the subscript m for the rest of the proof. We let $\Delta_{K\sigma}$ be the half diamond cell $\Delta_\sigma \cap K$, and notice that $m(\Delta_{K\sigma}) = \frac{1}{d} m_\sigma d(x_K, \sigma) \leq \frac{h_\mathcal{T} m_\sigma}{d}$. Elementary calculations yield:

$$\|\pi_\mathcal{T} \boldsymbol{u} - \widetilde{\boldsymbol{u}}_\mathcal{E}\|_{L^1(\Omega)} \leq \frac{h_\mathcal{T}}{d} \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} m_\sigma |u_K - u_\sigma|.$$

Thanks to our average assumption, we have $|u_K - u_\sigma| \leq D_\sigma \boldsymbol{u}$ thus:

$$\|\pi_\mathcal{T} \boldsymbol{u} - \widetilde{\boldsymbol{u}}_\mathcal{E}\|_{L^1(\Omega)} \leq \frac{2h_\mathcal{T}}{d} \sum_{\sigma \in \mathcal{E}^{\rm int}} m_\sigma D_\sigma \boldsymbol{u}.$$

Let p be as in the lemma and q its Hölder conjugate. We have:

$$\sum_{\sigma \in \mathcal{E}^{\rm int}} m_\sigma D_\sigma \boldsymbol{u} \leq \left(\sum_{\sigma \in \mathcal{E}^{\rm int}} m_\sigma d_\sigma \left(\frac{D_\sigma \boldsymbol{u}}{d_\sigma} \right)^p \right)^{\frac{1}{p}} \left(\sum_{\sigma \in \mathcal{E}^{\rm int}} m_\sigma d_\sigma \right)^{\frac{1}{q}} \leq d m(\Omega)^{\frac{1}{q}} \|\nabla_\mathcal{T} \boldsymbol{u}\|_{L^p(\Omega)},$$

hence:

$$\|\pi_{\mathcal{T}}\boldsymbol{u} - \widetilde{\boldsymbol{u}}_{\mathcal{E}}\|_{L^1(\Omega)} \leq 2m(\Omega)^{\frac{1}{q}} h_{\mathcal{T}} \|\nabla_{\mathcal{T}}\boldsymbol{u}\|_{L^p(\Omega)} \to 0. \tag{3.D.1}$$

This concludes the proof of the lemma. □

For reconstructions in $\Omega \times [0,T]$, we consider $\boldsymbol{u}_m \in \mathbb{R}^{\mathcal{T}_m \times \boldsymbol{\Delta t}_m}$, $\widetilde{\boldsymbol{u}}_m \in \mathbb{R}^{\mathcal{E}_m^{\mathrm{int}} \times \boldsymbol{\Delta t}_m}$ satisfying the same average property, and we have the same result.

Lemma 3.D.2. *If we dispose of $p \in [1,\infty)$, $\widetilde{p} \in [1,\infty)$ such that*

$$h_{\mathcal{T}_m} \|\nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{u}_m\|_{L^{\widetilde{p}}([0,T], L^p(\Omega))} \to 0$$

The $L^1(\Omega \times [0,T])$ convergence of the natural and diamond reconstructions are equivalent, moreover if one of them is convergent, they share the same limit.

Proof. This result is equivalent to:

$$\|\pi_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{u}_m - \widetilde{\boldsymbol{u}}_{m, \mathcal{E}_m, \boldsymbol{\Delta t}_m}\|_{L^1(\Omega \times [0,T])} \to 0.$$

We make use of the computations for the previous lemma, namely (3.D.1) yields for all $n \in [\![1, N_{T,m}]\!]$:

$$\|\pi_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{u}_m^n - \widetilde{\boldsymbol{u}}_{m, \mathcal{E}_m, \boldsymbol{\Delta t}_m}^n\|_{L^1(\Omega)} \leq 2m(\Omega)^{\frac{1}{q}} h_{\mathcal{T}_m} \|\nabla_{\mathcal{T}_m} \boldsymbol{u}^n\|_{L^p(\Omega)}.$$

Thus:

$$\|\pi_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{u}_m - \widetilde{\boldsymbol{u}}_{m, \mathcal{E}_m, \boldsymbol{\Delta t}_m}\|_{L^1(\Omega \times [0,T])} \leq 2m(\Omega)^{\frac{1}{q}} h_{\mathcal{T}_m} \sum_{n=1}^{N_{T,m}} \Delta t_n \|\nabla_{\mathcal{T}_m} \boldsymbol{u}^n\|_{L^p(\Omega)}.$$

Hölder's inequality yields:

$$\|\pi_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{u}_m - \widetilde{\boldsymbol{u}}_{m, \mathcal{E}_m, \boldsymbol{\Delta t}_m}\|_{L^1(\Omega \times [0,T])} \leq 2m(\Omega)^{\frac{1}{q}} T^{\frac{1}{\widetilde{q}}} h_{\mathcal{T}_m} \|\nabla_{\mathcal{T}_m, \boldsymbol{\Delta t}_m} \boldsymbol{u}\|_{L^{\widetilde{p}}([0,T], L^p(\Omega))},$$

where $\widetilde{q}$ is the Hölder conjugate of $\widetilde{p}$. Using the assumed estimation of the gradient, we have the announced result. □

Chapter 4

Variationnal modeling

Ce chapitre est un travail en cours. Il pourrait donner lieu à un article en collaboration avec Clément Cancès.

Outline of the current chapter

4.1 Introduction **142**
4.1.1 The recipe of Variational modelling for generalized gradient flows . . . 143
4.2 Two preliminary corrections to the Poisson Nernst Planck system **145**
4.2.1 Common settings . . . 146
4.2.2 Two ways of enforcing volume conservation . . . 147
4.3 Several possible choices for Ψ **151**
4.3.1 Fick-like dissipation . . . 151
4.3.2 Solvent-free dissipation . . . 152
4.3.3 Variation of the solvent-free dissipation . . . 153
4.3.4 Dissipation with macroscopic velocity . . . 154
4.3.5 Stefan-Maxwell dissipation . . . 156
4.4 Construction of schemes **158**
4.4.1 Discretisation of $\Omega \times [0, T]$ . . . 158
4.4.2 Formulation of the schemes . . . 160
4.5 Numerical comparison of the models **164**

4.1 Introduction

The proper coupling of individually well studied physical phenomenon is a mathematical and experimental challenge. A coupling between the diffusive and electric forces has been proposed by Walther Nernst and Max Planck in the late 19^{th} century [100]. This model is easy to compute (see introduction), and is widely used for industrial and modelling purposes. However, this coupling fails to account for the behaviour at high concentrations especially the boundary layers. Many different approaches have been made to correct these effects. Several of them are presented in [60]. In this article they also propose another correction. In this chapter we intend to build on variationnal considerations [113] to propose thermodynamically consistent coupled models. By construction, these models will be generalized gradient flows.

A general recipe for such models [104, 105, 41, 114] will be illustrated in Section 4.1.1 with the Nernst-Planck-Poisson system. In Section 4.2.1, we extend this framework to situations allowing for large concentrations by modifying the expression of the activities from concentrations to molar fractions. Along with this modification, we incorporate incompressibility in Section 4.2.2. Two approaches are investigated there, one enforcing a local balance of the fluxes (refered to as strong incompressibility), the other one only requiring a global balance (and referred to

as mild). We then apply these models to six different dissipation mechanisms inspired by the literature in Section 4.3. Numerical schemes for these models are provided in Section 4.4 and the behaviour of the models are showcased in a realistic salt-water settings in the last section thanks to preliminary numerical results.

4.1.1 The recipe of Variational modelling for generalized gradient flows

The variational modeling is a model-building art in which one chooses $\mathcal{Z}$ a set of admissible states, $E : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ an energy associated to each of these states, $\mathcal{J}$ a set of transformations of the admissible states, and $\Psi : \mathcal{J} \to \overline{\mathbb{R}}_+$ a cost function associated to a transformation. Ψ usually also depends on the admissible state. Two different examples of these sets and functions has been given in Section 1.1.4 of the general introduction for the particle diffusion equation derived by Adolf Fick in 1855. The current section focuses on the Nernst-Planck-Poisson system:

$$\partial_t c_i - \operatorname{div} D_i \left(\nabla c_i + c_i z_i \nabla \Phi\right) = 0, \quad \forall i \in [\![1, N]\!], \tag{4.1.1a}$$

$$-\lambda^2 \Delta \Phi = \sum_{i=1}^{N} z_i c_i. \tag{4.1.1b}$$

In this chapter, it will be supplemented with an initial condition and no flux boundary conditions:

$$\left(\nabla c_i + c_i z_i \nabla \Phi\right).n = 0 \qquad \text{on} \quad \partial\Omega.$$

We suppose that the electrostatic potential Φ is supplemented with a Dirichlet boundary condition on a non-negligible part of the boundary $\Gamma_{\mathrm{D}} \subset \partial\Omega$ and homogeneous Neumann boundary condition on $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$:

$$\Phi = \Phi^D \qquad \text{on } (0,T) \times \Gamma_D, \qquad\qquad \nabla\Phi \cdot n = 0 \qquad \text{on } (0,T) \times \Gamma_N, \tag{4.1.2}$$

where Φ^D is assumed to be constant in time and in $H^1(\Omega) \cap L^\infty(\Omega)$. Tanks to (4.1.1b) and (4.1.2), Φ can be seen as an implicit function of c, hence we let:

$$E(c) = \int_\Omega \sum_{i=1}^{N} c_i \log c_i + \lambda^2 \frac{|\nabla\Phi|^2}{2} \mathrm{d}x - \lambda^2 \int_{\Gamma_D} \Phi_D \nabla\Phi \cdot n. \tag{4.1.3}$$

One can notice that the first term of the energy is used in the introduction for the second derivation of Fick's law (see Section 1.1.4.b). Similarly, we let

$$\mathcal{Z} = \Omega \to [0, +\infty)^N$$

be the set of admissible concentrations. These concentrations follow a conservation equation:

$$\partial_t c_i + \operatorname{div} J_i = 0, \tag{4.1.4}$$

where J_i is an admissible transformation to be defined. We let

$$\mathcal{J} = \left\{ (J_1, \dots, J_N) \in H_{\operatorname{div}}(\Omega, \mathbb{R}^d)^N \mid \forall i, J_i \cdot n = 0 \text{ on } \partial\Omega \right\}$$

be the set of admissible transformations and

$$\Psi : J \in \mathcal{J} \mapsto \int_\Omega \sum_{i=1}^N \frac{|J_i|^2}{2 D_i c_i}$$

be the dissipation potential, i.e. the cost associated to a transformation.

The key idea to define the fluxes is to choose:

$$J_c \in \operatorname*{argmin}_{J \in \mathcal{J}} \Psi(J) + \langle \frac{\delta E}{\delta c}, -\operatorname{div} J \rangle. \tag{4.1.5}$$

The first term means that the system should not evolve too much, the second that it should decrease the energy as much as possible. Obviously, both terms are contradictory, but given that Ψ is coercive, the system will not jump form one state to another. Since J_c is a minimizer in (4.1.5) , we have:

$$\Psi(J_c) + \langle E'(c), -\operatorname{div} J_c \rangle - \Psi(0) \leq 0. \tag{4.1.6}$$

Thanks to the minimality of Ψ for $J = 0$, we have

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \langle E'(c), \partial_t c \rangle \overset{(4.1.4)}{=} \langle E'(c), -\operatorname{div} J_c \rangle \overset{(4.1.6)}{\leq} 0,$$

hence the decay of the energy E. As Ψ is also convex, we only have one minimum. To find it, we look for critical points, i.e. points where the derivative of

$$\Psi(J) + \langle \frac{\delta E}{\delta c}, -\operatorname{div} J \rangle$$

is zero. Letting $\mu_i = \log c_i + z_i \Phi + 1$ be the i-th component of the derivative of E, we have using integration by part:

$$\frac{\delta \Psi}{\delta J_i} = \frac{J_i}{D_i c_i}, \qquad \frac{\delta \langle E'(c), -\operatorname{div} J \rangle}{\delta J} = \nabla \mu,$$

so that:

$$\frac{J_i}{D_i c_i} + \nabla \mu_i = 0 \qquad \forall i \in [\![1, N]\!].$$

In other words, we have:

$$\begin{aligned} \partial_t c_i + \operatorname{div} J_i = 0, \qquad & \forall i \in [\![1, N]\!], \\ J_i = -D_i c_i \nabla \left(\log c_i + z_i \Phi \right), \qquad & \forall i \in [\![1, N]\!]. \end{aligned}$$

Up to the chain rule $c \nabla \log c = \nabla c$, we have the announced system.

In a more general setting, given A symmetric semi definite positive (see [107, 108] and the general remarks made in the general introduction, section 1.1.4.c), we consider an energy E, a state space $\mathcal{Z}$, an admissible transformation set $\mathcal{J}$ regular enough and

$$\Psi(J) = \int_\Omega \frac{J^T A J}{2}, \tag{4.1.7}$$

where $J^T A J$ is a short notation for:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} A_{i,j} J_i \cdot J_j.$$

If A is not definite, then either we have discontinuities in time or we have multiple choices of fluxes for the PDE systems. Form now on we will assume for simplicity that A is invertible so that

$$B = A^{-1} \tag{4.1.8}$$

is well defined [1]. Using shortened notations, we have:

$$J_c = B \nabla \mu, \tag{4.1.9}$$

where μ is the derivative of the energy. From now on, for the sake of simplicity, we will drop the subscript c on the fluxes.

4.2 Two preliminary corrections to the Poisson Nernst Planck system

The Poisson Nernst Planck system suffers form two flaws, first the mixing entropy $\sum_{i=1}^{N} c_i \log c_i$ is consistant with the statistical physics and the experiments

1. or that you agree to work with general inverses with infinite eigenvalues

only when the total concentration $\sum_{i=1}^{N} c_i$ is uniform in space, i.e the molar fraction and the concentrations are proportional and the proportionality coefficient does not depend on the composition of the mixture. This flaw will be tackled in the following section. The second flow, to be treated in Section 4.2.2, is the lack of mechanical coupling wich implies huge concentrations near the ohmic contacts. For reasons that will appear clearly at the end of next section, the consistence with the thermodynamics also requires mechanical coupling.

4.2.1 Common settings

In this chapter, we focus on the modeling of charged particles, therefore we have to model the electric field. We will use a electrostatic setting. As in previous section, we let:

$$-\lambda^2 \Delta \Phi = \sum_{i=1}^{N} z_i c_i,$$

supplemented with a Dirichlet boundary condition for the potential on a non-negligible part of the boundary $\Gamma_{\mathrm{D}} \subset \partial\Omega$ and homogeneous Neumann boundary condition on $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$:

$$\Phi = \Phi^D \qquad \text{on } (0,T) \times \Gamma_D, \qquad\qquad \nabla\Phi \cdot n = 0 \qquad \text{on } (0,T) \times \Gamma_N. \tag{4.2.1}$$

Here Φ^D is assumed to be constant in time and in $H^1(\Omega) \cap L^\infty(\Omega)$. As previously Φ can be seen as a function of the concentrations. Thus, we let the admissible state describe the concentrations only:

$$\mathcal{Z} = \left\{ (c_1, \dots, c_N) \in \Omega \to \mathbb{R}^N \middle| \forall i \in [\![1, N]\!], \int_\Omega c_i(t) = \int_\Omega c_i^0 \right\},$$

where c_i^0 is the initial condition of the specie i. For the transformation process, we keep:

$$\mathcal{J} = \left\{ (J_1, \dots, J_N) \in H^1(\Omega, \mathbb{R}^d)^N \mid \forall i, J_i \cdot n = 0 \text{ on } \partial\Omega \right\}.$$

We still have to define E. A free-energy for this problem has been proposed in [60]. Once rephrased in the setting of Chapter 3 :

$$E(c) = \int_\Omega \sum_{i=1}^{N} c_i \log \frac{c_i}{\overline{c}} + \lambda^2 \frac{|\nabla\Phi|^2}{2} \mathrm{d}x - \lambda^2 \int_{\Gamma_D} \Phi_D \nabla\Phi \cdot n, \tag{4.2.2}$$

where $\overline{c} = \sum_{i=1}^{N} c_i$ is the total concentration and $\frac{c_i}{\overline{c}}$ the molar fraction of the i-th specie. We have:

$$\langle \frac{\delta E}{\delta c}, -\operatorname{div} J \rangle = -\sum_{i=1}^{N} \int_{\Omega} \left(\log \frac{c_i}{\overline{c}} + z_i \phi \right) \operatorname{div} J_i.$$

We let $\mu_i = \log \frac{c_i}{\overline{c}} + z_i \phi$ the electrochemical potential and $\mu = (\mu_1, ..., \mu_N)$ so that:

$$\langle \frac{\delta E}{\delta c}, -\operatorname{div} J \rangle = \langle \nabla \mu, J \rangle.$$

For $z_i = 0$, we expect to find a reasonnable diffusive setting. Unfortunately this is not the case. Indeed as the chemical potential depends on the concentrations only through the molar fractions, for two neutral species the fully compressible model (4.1.9) would consider $c_1 = \alpha c_2$ at equilibrium for all $\alpha > 0$ and c_1 measurable. Such non-smooth solutions are obviously not physical.

4.2.2 Two ways of enforcing volume conservation

One of the key feature introduced by [60, 59] is the incompressibility of the fluid. In other words, given $v = (v_1, ..., v_N)$ the molar volumes of each species we should have:

$$\sum_{i=1}^{N} c_i v_i = 1. \tag{4.2.3}$$

This assumption is capital for reliable diffusion.

Form now on we assume that the initial condition satisfies this constraint. To transfer this constraint to the following times, we have two possible choices.

4.2.2.a Mild incompressibility constraint

The first natural solution is to place the constraint on the energy as in [60, 59]:

$$E^{\text{incomp}}(c) = E(c) + \chi(1 - \sum_{i=1}^{N} c_i v_i), \qquad \chi(x) = \begin{cases} 0 & \text{if } x = 0, \\ +\infty & \text{else.} \end{cases}$$

With this constraint, the function $\Psi(J) + \langle E'(c), -\operatorname{div} J \rangle$ is no longer derivable, thus we use the sub-differential theory[50]. In this framework, we have:

$$\partial \chi(0) = \mathbb{R},$$

so that:

$$\mu^{\text{incomp}} = \{\mu + v_i\pi, \pi \in \mathbb{R}\}, \qquad \text{for } \sum c_i v_i = 1.$$

Critical points are now points where 0 belongs to the subdifferential. The fluxes are then:

$$J^{\text{incomp}} = B\nabla(\mu + v\pi), \qquad \sum c_i v_i = 1.$$

The second half of the equation yields $\partial_t \sum c_i v_i = 0$ hence

$$\operatorname{div} \sum_{i=1}^{N} J_i v_i = 0 \tag{4.2.4}$$

and the equivalent expression of the fluxes:

$$J^{\text{incomp}} = B\nabla(\mu + v\pi), \qquad \operatorname{div} \sum_{i=1}^{N} J_i v_i = 0.$$

The equivalent constraint $\operatorname{div} \sum_{i=1}^{N} J_i v_i = 0$ can be expressed on the dissipation potential, and will be referred to as "mild" constraint in opposition to the strong one of next section. Thus we let:

$$\Psi^{\text{mild}}(J) = \Psi(J) + \chi(\operatorname{div} \sum_{i=1}^{N} J_i v_i), \tag{4.2.5}$$

and the associated fluxes:

$$J^{\text{mild}} = -B(\nabla\mu + v\pi), \qquad \operatorname{div} \sum_{i=1}^{N} J_{c,i}^{\text{mild}} v_i = 0, \tag{4.2.6}$$

4.2.2.b Strong incompressibility constraint

When the evolution of the system occurs from exchanges of particles, as in a cristaline network, the condition

$$\sum_{i=1}^{N} J_i v_i = 0$$

arises more naturally (see the system of Chapter 5). This constraint is stronger than (4.2.4) and we let:

$$\Psi^{\text{strong}}(J) = \Psi(J) + \chi(\sum_{i=1}^{N} J_i v_i), \tag{4.2.7}$$

with a slight abuse of notation if the space dimension d is greater than 1. With this choice, we have:

$$\partial\Psi^{\text{strong}}(J) = \{AJ + vF, F \in \Omega \to \mathbb{R}^N\} \qquad \text{for} \quad \sum_{i=1}^{N} J_i v_i = 0,$$

hence

$$J^{\text{strong}} = -B\left(vF + \nabla\mu\right), \qquad \sum_{i=1}^{N} J_i^{\text{strong}} v_i = 0. \tag{4.2.8}$$

It is often difficult to know which condition should be used, for exemple the relaxation from strong condition to mild condition has been proposed by Otto and E [111] for a polymer model proposed by de Gennes [54].

4.2.2.c Reformulation of the fluxes

The use of d Lagrange multipliers in (4.2.8) would be computationally intensive and can be avoided. Indeed, in (4.2.8) F can be explicitly computed:

$$J^{\text{strong}} = -B\left(vF + \nabla\mu\right), \qquad F = -\frac{v^T B \nabla\mu}{v^T B v}, \tag{4.2.9}$$

or

$$J^{\text{strong}} = -\widetilde{B}\nabla\mu, \quad \widetilde{B} = \left(I_N - \frac{Bvv^T}{v^T B v}\right) B. \tag{4.2.10}$$

Using the symmetry of A (and thus B), we have:

$$\widetilde{B} = B - \frac{bb^T}{v^T b}, \quad b = Bv.$$

To enhance the difference between the mild and strong enforcement of the incompressibility, we can notice that in (4.2.6) we can use F defined in (4.2.9) and a divergence free flux u_{conv}. More precisely, we let:

$$u_{\text{conv}} = -\sum_{i=1}^{N} J_i^{\text{mild}} v_i = v^T B v \nabla\pi + v^T B \nabla\mu,$$

so that:

$$\nabla\pi = \frac{-v^T B\nabla\mu}{v^T Bv} + \frac{u_{\text{conv}}}{v^T Bv}.$$

Finally (4.2.6) becomes

$$J^{\text{mild}} = -\widetilde{B}\nabla\mu - \frac{Bv}{v^T Bv}u_{\text{conv}}, \quad u_{\text{conv}} = v^T Bv\nabla\pi + v^T B\nabla\mu, \quad \text{div}\sum_{i=1}^{N} J_i^{\text{mild}} v_i = 0, \tag{4.2.11}$$

where $\widetilde{B}$ is as in (4.2.10). This formulation is used in [43]. The addition of this convective term shows that the mild enforcement of $\sum_{i=1}^{N} c_i v_i = 1$ allows for more flexibility in the fluxes, and thus a faster free-energy dissipation (see numerical experiments, namely Figure 4.3). Let's emphasize this aspect.

In the strong setting, the evolution free energy E can be computed using the fluxes (4.2.10) and the relation $\frac{dE}{dt} = -\langle\nabla\mu, J\rangle$. This yields :

$$\frac{dE^{\text{strong}}}{dt} = -\int_\Omega \nabla\mu^T \tilde{B}\nabla\mu, \tag{4.2.12}$$

where $\tilde{B}$ is semi-definite positive.

In the mild setting, we notice that:

$$\nabla\mu + v\pi = AJ^{\text{mild}} = A\widetilde{B}\nabla\mu + \frac{vu_{\text{conv}}}{v^T Bv}. \tag{4.2.13}$$

Moreover, thanks to $v^T \,\text{div}\, J^{\text{mild}} = 0$ we have:

$$\langle\mu, \text{div}\, J^{\text{mild}}\rangle = \langle\mu + v\pi, \text{div}\, J^{\text{mild}}\rangle. \tag{4.2.14}$$

Using

$$\frac{dE^{\text{mild}}}{dt} = \langle\mu, \text{div}\, J^{\text{mild}}\rangle,$$

an integration by part and equations (4.2.13), (4.2.14) and (4.2.6) we have:

$$-\frac{dE^{\text{mild}}}{dt} = \langle A\widetilde{B}\nabla\mu + \frac{vu_{\text{conv}}}{v^T Bv}, B\nabla(\mu + v\pi)\rangle. \tag{4.2.15}$$

By symmetry of A and $\widetilde{B}$ and using $\widetilde{B}v = 0$, we have:

$$\langle A\widetilde{B}\nabla\mu, B\nabla(\mu + v\pi)\rangle = \langle\nabla\mu, \widetilde{B}\nabla\mu + v\pi\rangle = \langle\nabla\mu, \widetilde{B}\nabla\mu\rangle.$$

For the second contribution in (4.2.15), we recall that $u_{\rm conv} = -v^T J^{\rm mild}$ so that:

$$\langle \frac{v u_{\rm conv}}{v^T B v}, B\nabla(\mu + v\pi)\rangle = \int_\Omega \frac{|u_{\rm conv}|^2}{v^T B v}.$$

Finally we have:

$$-\frac{{\rm d}E^{\rm mild}}{{\rm d}t} = \langle \nabla\mu, \widetilde{B}\nabla\mu\rangle + \int_\Omega \frac{|u_{\rm conv}|^2}{v^T B v} \geq -\frac{{\rm d}E^{\rm strong}}{{\rm d}t}. \tag{4.2.16}$$

4.3 Several possible choices for Ψ

In this section, we propose different dissipation mechanisms. We do not try to be exhaustive [22] as fairly general formulas have been given in the previous sections (4.2.6) or (4.2.11) and (4.2.9) or (4.2.10), but we aim at explicit expressions of the Onsager matrices $A(c)$ and $B(c)$ defined in (4.1.7) and (4.1.8) as clear as possible. The models are given by increasing order of complexity. In many cases the simplest expression of the PDE system uses the fluxes and the conservation equation (4.1.4). In these cases, this conservation equation will not be repeated and we will only give expressions of the fluxes.

4.3.1 Fick-like dissipation

First, we intend to use the same dissipation potential as in the second Fick's exemple of the general introduction and the Poisson Nernst Planck model of section 4.1.1:

$$\Psi(J) = \int_\Omega \sum_{i=1}^N \frac{|J_i|^2}{2D_i c_i}. \tag{4.3.1}$$

This choice of dissipation will fix the homogeneity of the diffusion constants D_i along the chapter.

In this case, we have $A = {\rm diag}(\frac{1}{D_i c_i})$ and $B = {\rm diag}(D_i c_i)$ so that (4.2.6) (with mild incompressibility constraint) becomes:

$$J_i^{\rm mild} = -D_i c_i \nabla(\mu_i + v_i\pi), \quad \forall i \in [\![1, N]\!] \qquad \text{and} \qquad {\rm div}\left(\sum v_i J_i^{\rm mild}\right) = 0. \tag{4.3.2}$$

For the strong incompressibility constraint, the easiest form is (4.2.10) with

$$\widetilde{B} = {\rm diag}(D_i c_i) - \frac{bb^T}{\sum_{i=1}^N D_i c_i v_i^2}, \quad b = \begin{pmatrix} D_1 c_1 v_1 \\ \vdots \\ D_N c_N v_N \end{pmatrix} = Bv.$$

This matrix is also usefull for the convective formulation of the mild fluxes (4.2.11).

4.3.2 Solvent-free dissipation

In the Fick-like setting we gave no special role for the solvent. In some setting the solvent is much more mobile than the other species. Let the solvent be the last specie, we set:

$$\Psi(J) = \int_\Omega \sum_{i=1}^{N-1} \frac{|J_i|^2}{2D_i c_i}. \tag{4.3.3}$$

This dissipation is not coercive as the matrix A has a kernel of dimension one. To recover coervicity we have to enforce the strong volume-filling condition (4.2.7). In that setting, the first-order optimality condition gives:

$$\frac{J_i^{\text{strong}}}{D_i c_i} = -v_i F - \nabla \mu_i, \quad \forall i \in [\![1, N-1]\!] \tag{4.3.4}$$

$$0 = -v_N F - \nabla \mu_N \tag{4.3.5}$$

$$\sum v_i J_i^{\text{strong}} = 0. \tag{4.3.6}$$

Using (4.3.5) in (4.3.4), we set

$$J_i^{\text{strong}} = -D_i c_i \nabla \left(\mu_i - \frac{v_i}{v_N} \mu_N \right), \quad \forall i \in [\![1, N-1]\!].$$

The equation (4.3.6) provides

$$J_N^{\text{strong}} = -\sum_{i=1}^{N-1} \frac{v_i}{v_N} J_i^{\text{strong}}$$

or simply

$$c_N = \frac{1}{v_N} - \sum_{i=1}^{N} \frac{v_i}{v_N} c_i.$$

For the mild constraint we no longer have uniqueness of the optimal J. Indeed, we have $\nabla \pi = \frac{-1}{v_N} \nabla \mu_N$, thus

$$J_i^{\text{mild}} = -D_i c_i \nabla \left(\mu_i - \frac{v_i}{v_N} \mu_N \right), \quad \forall i \in [\![1, N-1]\!],$$

$$J_N^{\text{mild}} = -\sum_{i=1}^{N-1} \frac{v_i}{v_N} J_i + u_{\text{conv}},$$

where $\mathrm{div}(u_{\mathrm{conv}}) = 0$. The degree of freedom of a divergence free additional term for J_N^{mild} is does not impact the concentration profiles. Therefore the strong and mild constraint yield the same model:

$$\partial_t c_i + \mathrm{div}(J_i^{\mathrm{incomp}}) = 0, \quad \forall i \in [\![1, N-1]\!], \tag{4.3.7a}$$

$$J_i^{\mathrm{incomp}} = -D_i c_i \nabla \left(\mu_i - \frac{v_i}{v_N} \mu_N \right), \quad \forall i \in [\![1, N-1]\!], \tag{4.3.7b}$$

$$c_N = \frac{1}{v_N} - \sum_{i=1}^{N} \frac{v_i}{v_N} c_i. \tag{4.3.7c}$$

This setting is the one of Chapter 3.

Remark 4.3.1. The model proposed in [60, 59] even-though similar came from a different modeling technique. Among other differences, the molar volumes ratio in the flux was a ratio of the molar masses and they added a convective speed following the Euler equation. An intermediary model proposed in [71] and derived in the general introduction can be obtained with:

$$\Psi^{\mathrm{static}}(J) = \sum_{i=1}^{N-1} \frac{|J_i|^2}{2D_i c_i} + \chi(\sum_{i=1}^{N} M_i J_i) + \chi(\mathrm{div} \sum_{i=1}^{N} v_i J_i),$$

thus:

$$J_i^{\mathrm{incomp}} = -D_i c_i \nabla \left((\mu_i + v_i \pi) - \frac{M_i}{M_N} (\mu_N + v_N \pi) \right), \quad \forall i \in [\![1, N-1]\!].$$

This two Lagrange-multipliers model will not be considered in this chapter.

4.3.3 Variation of the solvent-free dissipation

In several cases, the model comes from permutation of solvant molecules and solvated ions, therefore we propose a modification of the dissipation (4.3.3) symetrizing the roles of the solvant and the ions:

$$\Psi(J) = \int_\Omega \sum_{i=1}^{N-1} \frac{|J_i|^2}{2D_i \frac{c_i c_N}{\bar{c}}}.$$

The division by $\bar{c}$ is necessary for the homogeneity. As in the previous model, we need the strong condition on the fluxes (4.2.7) to enforce coercivity. Computations

are highly similar and we obtain:

$$\partial_t c_i + \operatorname{div} J_i^{\text{strong}} = 0, \qquad \forall i \in [\![1, N-1]\!], \tag{4.3.8a}$$

$$J_i^{\text{strong}} = D_i \frac{c_i c_N}{\overline{c}} \nabla \left(\mu_i - \frac{v_i}{v_N} \mu_N \right), \qquad \forall i \in [\![1, N-1]\!], \tag{4.3.8b}$$

$$c_N = \frac{1}{v_N} - \sum_{i=1}^{N-1} \frac{c_i v_i}{v_N}. \tag{4.3.8c}$$

As previously, the mild condition allows for several choices for the fluxes but all of them yields the PDE system defined above. This system is studied in [26] (for $N = 2$).

4.3.4 Dissipation with macroscopic velocity

In this section, we consider that friction stems from the differences of the velocities. Therefore, we consider $u_i = J_i/c_i$ the speed of each species and $u(J, c)$ a macroscopic speed. For the sake of simplicity, we restrict ourselves to macroscopic speeds defined as convex combinations of the species speeds, i.e.:

$$u(J, c) = \sum_{i=1}^{N} \alpha_i(c) u_i, \tag{4.3.9}$$

where $\alpha_i : \mathbb{R}^N \to \mathbb{R}_+$ are such that $\sum_{i=1}^N \alpha_i(c) = 1$. For example, we could have $\alpha_1, ..., \alpha_{N-1} = 0$ and $\alpha_N = 1$ so that u is the "solvent speed". An other example is the "barycentric speed" $\alpha_i(c) = M_i c_i/\rho$, where M_i is the molar mass of the specie i and $\rho = \sum_{i=1}^N M_i c_i$. For the dissipation process, we let:

$$\Psi(J) = \int_\Omega \eta \frac{|u(J,c)|^2}{2} + \int_\Omega \sum_{i=1}^{N} \frac{c_i |u_i - u(J,c)|^2}{2D_i}$$

or equivalently:

$$\Psi(J) = \int_\Omega \eta \frac{|u(J,c)|^2}{2} + \int_\Omega \sum_{i=1}^{N} \frac{|J_i - c_i u(J,c)|^2}{2D_i c_i}.$$

The first term can be thought of as a microscopic friction with the surrounding environment. Its introduction is required to ensure the coercivity of Ψ. Indeed, if $\eta = 0$, then for $u_1 = u_2 = ... = u_N$ we would have $\Psi(J) = 0$. To find the PDE

system hidden behind this dissipation, we compute its first variation. We have:

$$\frac{\delta\Psi(J)}{\delta J}\cdot H = \int_\Omega \eta u(J,c)\cdot\sum_{i=1}^N \frac{\alpha_i(c)}{c_i}H_i + \sum_{i=1}^N \frac{(J_i - c_i u(J,c))\cdot\left(H_i - c_i\sum_{j=1}^N \frac{\alpha_j(c)}{c_j}H_j\right)}{D_i c_i}.$$

To ease the practical computation of the PDE, we let A such that:

$$\frac{\partial\Psi(J)}{\partial J}\cdot H = \int_\Omega \sum_{i=1}^N\sum_{j=1}^N A_{i,j}J_j\cdot H_i.$$

This matrix can be identified to the one defined in (4.1.7). A computation whose length is the only difficulty yields:

$$A = \left(\eta + \sum_{k=1}^N \frac{c_k}{D_k}\right)aa^T + \mathrm{diag}(\frac{1}{D_i c_i}) - ad^T - da^T,$$

with $a = (\frac{\alpha_1}{c_1},...,\frac{\alpha_N}{c_N})^T$ and $d = (\frac{1}{D_1},...,\frac{1}{D_N})^T$. This matrix is symmetric definite positive thus nonsingular, however the formulation of a general inverse B is not obvious and beyond the scope of this chapter. The practical models are thus only available on their most general formulations (4.2.6) or (4.2.11) and (4.2.10).

When u is the "solvent speed" ($\alpha_i = \delta_{i,N}$) the computations are more palatable and we have

$$A = \begin{pmatrix} \frac{1}{D_1c_1} & & & \frac{-1}{D_1c_N} \\ & \ddots & & \vdots \\ & & \frac{1}{D_{N-1}c_{N-1}} & \frac{-1}{D_{N-1}c_N} \\ \frac{-1}{D_1c_N} & \cdots & \frac{-1}{D_{N-1}c_N} & \frac{\eta+\sum_{k=1}^{N-1}\frac{c_k}{D_k}}{c_N^2} \end{pmatrix}.$$

The inversion of an arrowhead matrix is explicit, and in this setting, we have:

$$B = \begin{pmatrix} D_1c_1 & & & \\ & \ddots & & \\ & & D_{N-1}c_{N-1} & \\ & & & 0 \end{pmatrix} + \frac{1}{\eta}cc^T. \tag{4.3.10}$$

We adapt the models (4.2.6), and (4.2.9) to the particular choice for B (4.3.10).

With the mild incompressibility constraint, the resulting system is:

$$J_i^{\text{mild}} = -c_i \left(D_i \nabla(\mu_i + v_i \pi) + \frac{1}{\eta} \left(\nabla \pi + \sum_{j=1}^{N} c_j \nabla \mu_j \right) \right), \qquad \forall i \in [\![1, N-1]\!], \tag{4.3.11a}$$

$$J_N^{\text{mild}} = -\frac{c_N}{\eta} \left(\nabla \pi + \sum_{j=1}^{N} c_j \nabla \mu_j \right), \tag{4.3.11b}$$

$$\operatorname{div} \sum_{i=1}^{N} v_i J_i^{\text{mild}} = 0. \tag{4.3.11c}$$

And finally the strong incompressibility constraint yields:

$$J_i^{\text{strong}} = -c_i \left(D_i \left(v_i F + \nabla \mu_i \right) + \frac{1}{\eta} \left(F + \sum_{j=1}^{N} c_j \nabla \mu_j \right) \right), \qquad \forall i \in [\![1, N-1]\!], \tag{4.3.12a}$$

$$J_N^{\text{strong}} = -\frac{c_N}{\eta} \left(F + \sum_{j=1}^{N} c_j \nabla \mu_j \right), \tag{4.3.12b}$$

$$F = \frac{-\sum_{i=1}^{N-1} D_i c_i v_i \nabla \mu_i - \frac{1}{\eta} \sum_{i=1}^{N} c_i \nabla \mu_i}{\frac{1}{\eta} + \sum_{i=1}^{N-1} D_i c_i v_i^2}. \tag{4.3.12c}$$

4.3.5 Stefan-Maxwell dissipation

A popular alternative approach to model multi-component diffusion has been proposed independently by James Clerk Maxwell [102] and Josef Stefan [121]. The corresponding model can be derived from the multi-component Boltzmann equation [19]. The most generic presentation of Maxwell-Stefan dissipation potential is the following :

$$\Psi(J) = \int_\Omega \sum_{1 \le i,j \le N} \frac{|u_i - u_j|^2}{4 K_{i,j}(c)},$$

where $K(c) = (K_{i,j}(c))_{1 \le i,j \le N}$ is symmetric with positive coefficients and $u_i = \frac{J_i}{c_i}$ is the speed of specie i. We have:

$$\frac{\delta \Psi}{\delta J} \cdot H = \int_\Omega \sum_{1 \le i,j \le N} \frac{(u_i - u_j) \cdot h_i}{K_{i,j}(c)}, \qquad h_i = \frac{H_i}{c_i} \quad \forall i \in [\![1, N]\!].$$

As the modelling setting developped in the previous sections uses J_i rather than u_i, we have:

$$\Psi(J) = \int_\Omega \sum_{1\le i,j\le N} \frac{|c_j J_i - c_i J_j|^2}{4K_{i,j}(c)(c_i c_j)^2}.$$

Unfortunately, $K_{i,j}$ and the diffusion coeficients of the previous sections D_i are not homogeneous. To correct that flaw, we let:

$$D_{i,j}(c) = \frac{K_{i,j}(c)c_i c_j}{\overline{c}},$$

where $D_{i,j} = D_{j,i} > 0$ has the same dimension as D_i the Fickian diffusion coefficients leading to

$$\Psi(J) = \int_\Omega \sum_{1\le i,j\le N} \frac{|c_j J_i - c_i J_j|^2}{4D_{i,j}(c)c_i c_j \overline{c}}.$$

From now on, we will assume that $D_{i,j}$ does not depend on c for the sake of simplicity, so that:

$$\Psi(J) = \int_\Omega \sum_{1\le i,j\le N} \frac{c_i c_j |u_i - u_j|^2}{4D_{i,j}\overline{c}} = \int_\Omega \sum_{1\le i,j\le N} \frac{|c_j J_i - c_i J_j|^2}{4D_{i,j}c_i c_j \overline{c}}. \tag{4.3.13}$$

As in the previous section, to identify the matrix A of (4.1.7), we compute the first variation of Ψ:

$$\frac{\delta\Psi}{\delta J}\cdot H = \int_\Omega \sum_{1\le i,j\le N} \frac{(c_j J_i - c_i J_j)\cdot H_i}{D_{i,j}c_i\overline{c}}. \tag{4.3.14}$$

One can easily check that the matrix of coefficients

$$a_{i,i} = -\sum_{j\neq i} \frac{c_j}{D_{i,j}\overline{c}c_i} \qquad a_{i,j} = \frac{1}{D_{i,j}\overline{c}}, \quad i\neq j$$

satisfies $\frac{\delta\Psi}{\delta J} = AJ$. Unfortunately, this matrix is singular. Eventhough this is not clear on A, one can see that for (c, J) satisfying $u_1 = u_2 = ... = u_N$, we have $\Psi(J) = 0$, so that $AJ = 0$ reciprocally, this is the only isotropic line of Ψ hence $\ker(A) = \text{span}(c)$ as shown for instance in [94]. To get rid of this non trivial kernel, we add $\kappa|\sum_{i=1}^N v_i c_i u_i|^2$ to Ψ and let:

$$\Psi_\kappa(J) = \Psi(J) + \kappa\int_\Omega |\sum_{i=1}^N v_i J_i|^2. \tag{4.3.15}$$

This choice ensures that with the constraint $\sum_{i=1}^{N} v_i J_i = 0$ the solutions of the PDE system are the same as the original for all κ. In [94, 21], the kernel of A is treated with the elimination of a specie to get existence results. The approach performed here is closer to the perturbation technique used in [40]. With this addition, we have:

$$\Psi_\kappa(J) = \int_\Omega \frac{J^T A_\kappa J}{2} \qquad A_\kappa = A + 2\kappa v v^T.$$

The matrix A_κ cannot be inverted explicitly in general, so that the model is generally kept in its primal form. However letting $B_\kappa = A_\kappa^{-1}$, we have:

$$J_\kappa^{\text{mild}} = B_\kappa \nabla(\mu + v\pi), \qquad \operatorname{div} v^T J_\kappa^{\text{mild}} = 0,$$
$$J^{\text{strong}} = \widetilde{B}\nabla\mu,$$

where $\widetilde{B}$ is computed using (4.2.10) and does not depend on κ.

4.4 Construction of schemes

Now that we dispose of several models, we want to be able to simulate them. For each model we use the two point flux approximation finite volumes in space and the implicit Euler method in time. The notation choices for these schemes are consistent with those done in introduction and the previous chapters. For the sake of concision of this thesis, we will keep this section minimal, more details can be found on section 3.2.1 of Chapter 3. Then in Section 4.4.2 we detail the numerical schemes used. Different formulations of the fluxes are needed depending on the knowledge of an explicit formulation of B and the elimination of an unknown.

4.4.1 Discretisation of $\Omega \times [0, T]$

We recall here the definition of an admissible mesh.

Definition 7. *An admissible mesh of Ω is a triplet $(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}})$ such that the following conditions are fulfilled.*

(i) The set $\mathcal{T}$ is finite and each control volume (or cell) $K \in \mathcal{T}$ is non-empty, open, polyhedral, and convex. We assume that

$$K \cap L = \emptyset \quad \text{if } K, L \in \mathcal{T} \text{ with } K \neq L, \qquad \text{while} \quad \bigcup_{K\in\mathcal{T}} \overline{K} = \overline{\Omega}.$$

(ii) *Each face $\sigma \in \mathcal{E}$ is closed and is contained in a hyperplane of $\mathbb{R}^d$, with positive $(d-1)$-dimensional Hausdorff (or Lebesgue) measure denoted by $m_\sigma = \mathcal{H}^{d-1}(\sigma) > 0$. We assume that $\mathcal{H}^{d-1}(\sigma \cap \sigma') = 0$ for $\sigma, \sigma' \in \mathcal{E}$ unless $\sigma' = \sigma$. For all $K \in \mathcal{T}$, we assume that there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \bigcup_{\sigma \in \mathcal{E}_K} \sigma$. Moreover, we suppose that $\bigcup_{K \in \mathcal{T}} \mathcal{E}_K = \mathcal{E}$. Given two distinct control volumes $K, L \in \mathcal{T}$, the intersection $\overline{K} \cap \overline{L}$ either reduces to a single face $\sigma \in \mathcal{E}$ denoted by $K|L$, or its $(d-1)$-dimensional Hausdorff measure is 0.*

(iii) *The cell centers $(x_K)_{K \in \mathcal{T}}$ belong to their cell: $x_K \in K$, and are such that, if $K, L \in \mathcal{T}$ share a face $K|L$, then the vector $x_L - x_K$ is orthogonal to $K|L$.*

(iv) *For the boundary faces $\sigma \subset \partial\Omega$, we assume that either $\sigma \subset \overline{\Gamma}_D$ or $\sigma \subset \overline{\Gamma}_N$. For $\sigma \subset \partial\Omega$ with $\sigma \in \mathcal{E}_K$ for some $K \in \mathcal{T}$, we assume additionally that there exists $x_\sigma \in \sigma$ such that $x_\sigma - x_K$ is orthogonal to σ.*

We denote by m_K the d-dimensional Lebesgue measure of the control volume K. The set of the faces is partitioned into two subsets: the set $\mathcal{E}_{\rm int}$ of the interior faces defined by $\mathcal{E}_{\rm int} = \{\sigma \in \mathcal{E} \mid \sigma = K|L \text{ for some } K, L \in \mathcal{T}\}$, and the set $\mathcal{E}_{\rm ext}$ of the exterior faces defined by $\mathcal{E}_{\rm ext} = \{\sigma \in \mathcal{E} \mid \sigma \subset \partial\Omega\}$, which can also be partitioned into $\mathcal{E}^{\rm D} = \{\sigma \subset \overline{\Gamma}_{\rm D}\}$ and $\mathcal{E}^{\rm N} = \{\sigma \subset \overline{\Gamma}_{\rm N}\}$.

Given $\sigma \in \mathcal{E}$, we let

$$d_\sigma = \begin{cases} |x_K - x_L| & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ |x_K - x_\sigma| & \text{if } \sigma \in \mathcal{E}_{\rm ext}, \end{cases} \qquad \text{and} \quad \tau_\sigma = \frac{m_\sigma}{d_\sigma}.$$

Concerning the time discretization of $(0,T)$, we consider an increasing finite family of times $0 = t_0 < t_1 < \ldots, < t_{N_T} = T$. We denote by $\Delta t_n = t_n - t_{n-1}$ for $1 \le n \le N_T$, by $\boldsymbol{\Delta t} = (\Delta t_n)_{1 \le n \le N_T}$. We will use boldface notations for vectors whose number of components is dependent on the mesh.

For all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_K$, we define the mirror values $c^n_{K\sigma}$, $\Phi^n_{K\sigma}$, and $\pi^n_{K\sigma}$ (when in the mild setting) of c^n_K, Φ^n_K and $\pi^n_{K\sigma}$ respectively across σ by setting

$$c^n_{K\sigma} = \begin{cases} c^n_L & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ c^n_K & \text{if } \sigma \in \mathcal{E}_{\rm ext}, \end{cases} \qquad \Phi^n_{K\sigma} = \begin{cases} \Phi^n_L & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ \Phi^n_K & \text{if } \sigma \in \mathcal{E}^N, \\ \Phi^D_\sigma = \fint_\sigma \Phi^D \mathrm{d}\boldsymbol{\gamma} & \text{if } \sigma \in \mathcal{E}^D. \end{cases}$$

$$\pi^n_{K\sigma} = \begin{cases} \pi^n_L & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ \pi^n_K & \text{if } \sigma \in \mathcal{E}_{\rm ext}, \end{cases} \tag{4.4.1}$$

Given $\boldsymbol{u} = (u_K)_{K \in \mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, we define the oriented and absolute jumps of $\boldsymbol{u}$ across

any edge by

$$D_{K\sigma}\boldsymbol{u} = u_{K\sigma} - u_K, \quad D_\sigma \boldsymbol{u} = |D_{K\sigma}\boldsymbol{u}|, \qquad \forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_K.$$

We may now use these operators to describe our scheme.

4.4.2 Formulation of the schemes

From Section 4.3, we have three kinds of dissipation dotentials Ψ. Those where B is known as in Section 4.3.1 and the solvent speed model of Section 4.3.4, those where we are not able to compute B as in Section 4.3.5 and the general form of Section 4.3.4, and those where the matrix is not well defined, like in Sections 4.3.2 and 4.3.3. After some general setting for our schemes, we propose expression of the fluxes for each of these cases in three separate subsections.

The conservation equation is approximated using a backward-Euler scheme in time :

$$m_K \frac{c_{K,i}^n - c_{K,i}^{n-1}}{\Delta t_n} + \sum_{\sigma \in \mathcal{E}_K^{\text{int}}} F_{K\sigma,i}^n = 0, \qquad \forall K \in \mathcal{T}, i \in [\![1, N]\!], \tag{4.4.2a}$$

where $F_{K\sigma,i}^n$ should be a conservative and consistent approximation of

$$-\frac{1}{\Delta t_n} \int_{t_{n-1}}^{t_n} \int_\sigma F_i \cdot n_{K\sigma}$$

where $n_{K\sigma}$ denotes the normal to σ outward K. For practical purposes, the potential Φ is considered as a full fledged variable, and discretized as $\boldsymbol{\Phi}$ using:

$$-\lambda^2 \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \boldsymbol{\Phi}^n = m_K \sum_{i=1}^N z_i c_{K,i}^n, \qquad \forall K \in \mathcal{T}. \tag{4.4.2b}$$

Similarly, in the mild setting, π is also considered as a variable and we discretize it into $\boldsymbol{\pi} = (\pi_K^n)_{K \in \mathcal{T}, n \in [\![1, N_T]\!]}$ using the equations:

$$\sum_{\sigma \in \mathcal{E}_K^{\text{int}}} \sum_{i=1}^N v_i F_{K\sigma,i}^n. = 0, \qquad \forall K \in \mathcal{T}. \tag{4.4.2c}$$

Because of the choice of no-flux boundary conditions, π is defined up to a constant.

To fix this degree of freedom, we add the constraint:

$$\sum_{K\in\mathcal{T}} |K| pi_K^n = 0, \qquad \forall n \in [\![1, N_T]\!]. \tag{4.4.2d}$$

Using the no-flux boundary condition, we can drop equation (4.4.2c) on one cell to enforce the previous constraint there, see for instance [44].

The only remaining degree of freedom is the expression of the numerical fluxes. For the sake of clarity the time index n will be dropped.

4.4.2.a When the Onsager matrix B is known

This setting is the most straightforward since we can set

$$B_\sigma = B\left(\frac{c_K + c_{K\sigma}}{2}\right) = (B_{\sigma,i,j})_{1\le i,j\le N}.$$

More intricate choices for this average could be defined as in Chapters 2 and 3. The numerical fluxes are then very naturally deduced from the formulations of the continuous fluxes:

$$F^{\text{mild}} = B\left(\nabla\mu + v\pi\right),$$
$$F^{\text{strong}} = \widetilde{B}\nabla\mu, \quad \widetilde{B} = B - \frac{bb^T}{v^T b}, \quad b = Bv.$$

After discretization, in the mild setting, we set:

$$F^{\text{mild}}_{K\sigma,i} = \tau_\sigma \sum_{j=1}^{N} B_{\sigma,i,j} D_{K\sigma}(\boldsymbol{\mu}_j + v_j\boldsymbol{\pi}), \qquad \forall i \in [\![1, N]\!],$$

or for short:

$$F^{\text{mild}}_{K\sigma} = \tau_\sigma B_\sigma D_{K\sigma}(\boldsymbol{\mu} + v\boldsymbol{\pi}).$$

Similarly, we have:

$$F^{\text{strong}}_{K\sigma} = \tau_\sigma \widetilde{B}_\sigma D_{K\sigma}\boldsymbol{\mu}.$$

In the mild setting, we can also build a scheme adapted from (4.2.11):

$$F_{K\sigma}^{\text{mild c-up}} = \tau_\sigma \widetilde{B}_\sigma D_{K\sigma}\boldsymbol{\mu} + \tau_\sigma \frac{B(c_\sigma^{\text{up}})v}{v^T B(c_\sigma^{\text{up}})v} u_{K\sigma},$$
$$u_{K\sigma} = v^T B_\sigma v D_{K\sigma}\boldsymbol{\pi} + v^T B_\sigma D_{K\sigma}\boldsymbol{\mu},$$
$$c_\sigma^{\text{up}} = \begin{cases} c_K & \text{if } u_{K\sigma} > 0, \\ c_{K\sigma} & \text{else.} \end{cases}$$

This scheme is based on ideas of [24].

4.4.2.b When the Onsager matrix B is not known explicitely

In this case, we know that the matrix A is invertible. However we do not dispose of an easy expression of its inverse. A first very natural idea is to compute the numerical inverse of A any time the matrix is needed. Eventhough this approach has the advantage of code reusability, general matrix inversion can be more costly and less accurate than system solving. Moreover, for several real-life applications, the number of species can be hig (say hundreds) so that the computational overcost may become prohibitive.

For the mildly incompressible fluxes we solve the numerical system:

$$A_\sigma F_{K\sigma}^{\text{mild}} = \tau_\sigma D_{K\sigma}(\boldsymbol{\mu} + v\boldsymbol{\pi}),$$

where $A_\sigma = A(\frac{c_K + c_{K\sigma}}{2})$ for all $\sigma \in \mathcal{E}_{\text{int}}$. For the strongly incompressible fluxes, we have one more system to solve:

$$A_\sigma F_{K\sigma}^{\text{strong}} = \tau_\sigma \left(I_N - \frac{v b_\sigma^T}{v^T b_\sigma} \right) \nabla\boldsymbol{\mu}, \qquad A_\sigma b_\sigma = v.$$

When using the splitted centered-upwind flux,the fluxes are computed as follows:

$$A_\sigma b_\sigma = v,$$
$$u_{K\sigma} = v^T b_\sigma D_{K\sigma}\boldsymbol{\pi} + b_\sigma^T D_{K\sigma}\boldsymbol{\mu},$$
$$c_\sigma^{\text{up}} = \begin{cases} c_K & \text{if } u_{K\sigma} > 0 \\ c_{K\sigma} & \text{otherwise} \end{cases}$$
$$A(c_\sigma^{\text{up}}) b_\sigma^{\text{up}} = v$$
$$A_\sigma F_{K\sigma}^{\text{mild c-up}} = \tau_\sigma \left(I_N - \frac{v b_\sigma^T}{v^T b_\sigma} \right) D_{K\sigma}\boldsymbol{\mu} + \tau_\sigma \frac{b_\sigma^{\text{up}}}{v^T b_\sigma^{\text{up}}} u_{K\sigma}.$$

This yields three system resolutions per edge and time step.

4.4.2.c The special case of Sections 4.3.2 and 4.3.3

The model described in Section 4.3.2 has already been studied in [81] with two schemes and in [71] with a third one. For the sake of simplicity, only the case of an electroneutral solvant ($z_N = 0$) is presented here. In these cases, due to the elimination of an unknown the system (4.4.2) is no longer formally valid. Thus, we set:

$$\frac{c_{K,i}^n - c_{K,i}^{n-1}}{\Delta t_n} + \sum_{\sigma=K|L} F_{K\sigma,i}^n = 0, \qquad \forall K \in \mathcal{T}, \forall i \in [\![1, N-1]\!], \tag{4.4.3}$$

$$-\lambda^2 \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \mathbf{\Phi}^n = m_K \sum_{i=1}^{N-1} z_i c_{K,i}^n, \qquad \forall K \in \mathcal{T}, \tag{4.4.4}$$

$$\boldsymbol{c}_N^n = \frac{1}{v_N} - \sum_{i=1}^{N-1} \frac{v_i}{v_N} \boldsymbol{c}_i^n. \tag{4.4.5}$$

We still have to define the fluxes. The centered flux is:

$$\begin{aligned} F_{K\sigma,i} &= \tau_\sigma D_i \frac{c_{K,i} + c_{K\sigma,i}}{2} D_{K\sigma} \left(\boldsymbol{\mu}_i - \frac{v_i}{v_N} \boldsymbol{\mu}_N \right) \\ &= \tau_\sigma D_i \frac{c_{K,i} + c_{K\sigma,i}}{2} D_{K\sigma} \left(\log \frac{\boldsymbol{c}_i}{\overline{\boldsymbol{c}}} - \frac{v_i}{v_N} \log \frac{\boldsymbol{c}_N}{\overline{\boldsymbol{c}}} + z_i \mathbf{\Phi} \right), \quad \overline{\boldsymbol{c}} = \sum_{i=1}^{N} \boldsymbol{c}_i. \end{aligned}$$

For the variation of this model presented in Section 4.3.3, we adapt the centered and flux by a simple multiplication by $\frac{c_{K,N}+c_{L,N}}{\overline{c_K}+\overline{c_L}}$:

$$F_{K\sigma,i} = \tau_\sigma D_i \frac{c_{K,i} + c_{K\sigma,i}}{2} \frac{c_{K,N} + c_{L,N}}{\overline{c_K} + \overline{c_L}} D_{K\sigma} \left(\boldsymbol{\mu}_i - \frac{v_i}{v_N} \boldsymbol{\mu}_N \right), \quad \overline{\boldsymbol{c}} = \sum_{i=1}^{N} \boldsymbol{c}_i.$$

Remark 4.4.1. The sedan and activity based fluxes studied in Chapter 2 could be expressed in the current formalism and extended to the new model using :

$$c_{KL,N} = \max \left(\frac{1}{v_N} - \sum_{i=1}^{N-1} \frac{v_i}{v_N} c_{KL,i}, \frac{c_{K,N} + c_{L,N}}{2} \right), \qquad \overline{c}_{KL} = \sum_{i=1}^{N} c_{KL,i}. \tag{4.4.6}$$

4.5 Numerical comparison of the models

To compare our models, we consider two solutions of sodium chloride (NaCl, or salt), the first one is saturated in solvated species with molars fractions of $6/67 \simeq 8.96\,10^{-2}$ for each of the two ions and $55/67 \simeq 8.21\,10^{-1}$ for the water, thus close to the $360g.L^{-1}$ mark. The second one is 10^5 times more diluted with molar fractions of $\frac{6E-5}{55+12E-5}$ for the ions, and $\frac{55}{55+12E-5}$ for the water. In both cases, the molar volumes are:

$$v_{\mathrm{Na+}} = 1.94 \qquad v_{\mathrm{Cl\text{-}}} = 23.24 \qquad v_{\mathrm{H_2O}} = 18.07,$$

according to measurements of [101]. The molar masses (used for the barycentric speed model) are given by:

$$M_{\mathrm{Na+}} = 22.99 \qquad M_{\mathrm{Cl\text{-}}} = 35.45 \qquad M_{\mathrm{H_2O}} = 18.02.$$

Finally the Fickian diffusion coefficients are set to:

$$D_{\mathrm{Na+}} = 1 \qquad D_{\mathrm{Cl\text{-}}} = 2 \qquad D_{\mathrm{H_2O}} = 3.$$

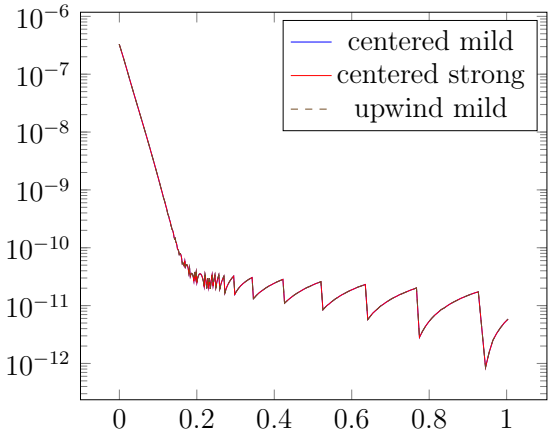
We consider units such that the Debye length λ is equal to 1.

The initial conditions are chosen homogeneous with respect to space. We consider a one dimensional case where an external electric potential $\Phi^D = 5$ is applied on the left boundary ($x = 0$) while $\Phi^D = -5$ is applied on the right boundary ($x = 1$), and a resembling two dimensional case, where we apply these conditions only on the upper half of the boundary ($y \in [0.5, 1)$) and set homogeneous Neumann boundary condition on the lower half and for the top and bottom boundaries ($y = 1$ and $y = 0$). Since u_{conv} is divergence free and $u_{\mathrm{conv}} \cdot n = 0$ on $\partial\Omega$, then $u_{\mathrm{conv}} = 0$ in the one dimensional setting. Therefore we deduce from (4.2.10) and (4.2.11) that mildly and strongly incompressible models coincide when $d = 1$. This is further confirmed by Figures 4.1 and 4.2.

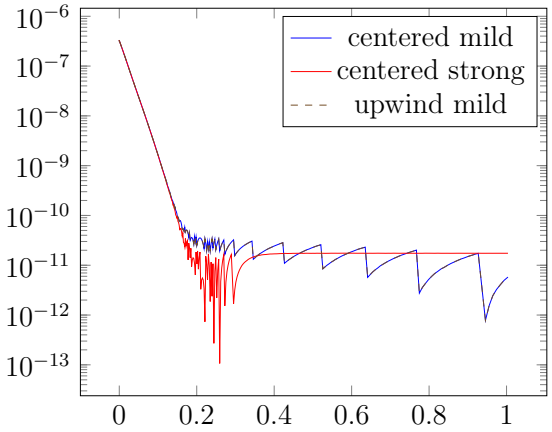
Another striking remark that can be done from figures 4.1 and 4.2 is the following: In the diluted regime, alll the dissipation mechanisms behave similarly to the usual Nernst-Planck-Poisson system, cf Figure 4.1. This is no longer true in the saturated regime, when depending on the choice of the dissipation potential Ψ, the dissipation rate of the free energy can change of one order of magnitude, cf Figure 4.2.

Finally, in the two dimensional setting, the convective velocity u_{conv} is no longer identically equal to 0 as depicted on Figure 4.7. Therefore, the free energy is dissipated faster for the mildly incompressible setting system than for the strongly incompressible ones. Figure 4.3 illustrates our finding (4.2.16).

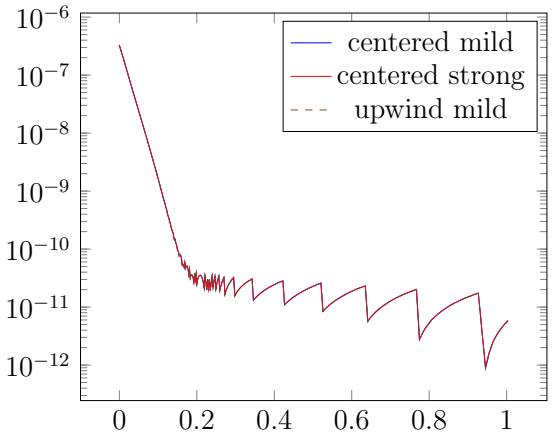
As the stationary solutions do not depend on the dissipation potential, the profiles presented here are obtained using the mild Fick-like dissipation potential which seems to have the fastest free energy decay. The stationary solutions are obtained at the first time step where the numerical error makes the free energy decrease. To prevent blow-ups in the numerical fluxes, the concentrations are capped at a minimum of 10^{-10}. This cut-off is visible in Figure 4.6.



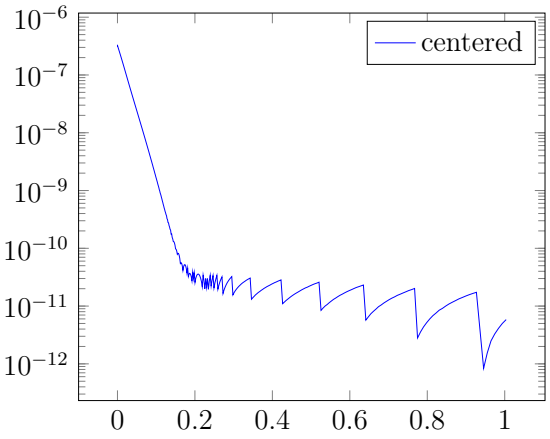
(a) Fick-like dissipation



(b) Dissipation using the barycentric speed as a macroscopic speed. $\eta = 2$

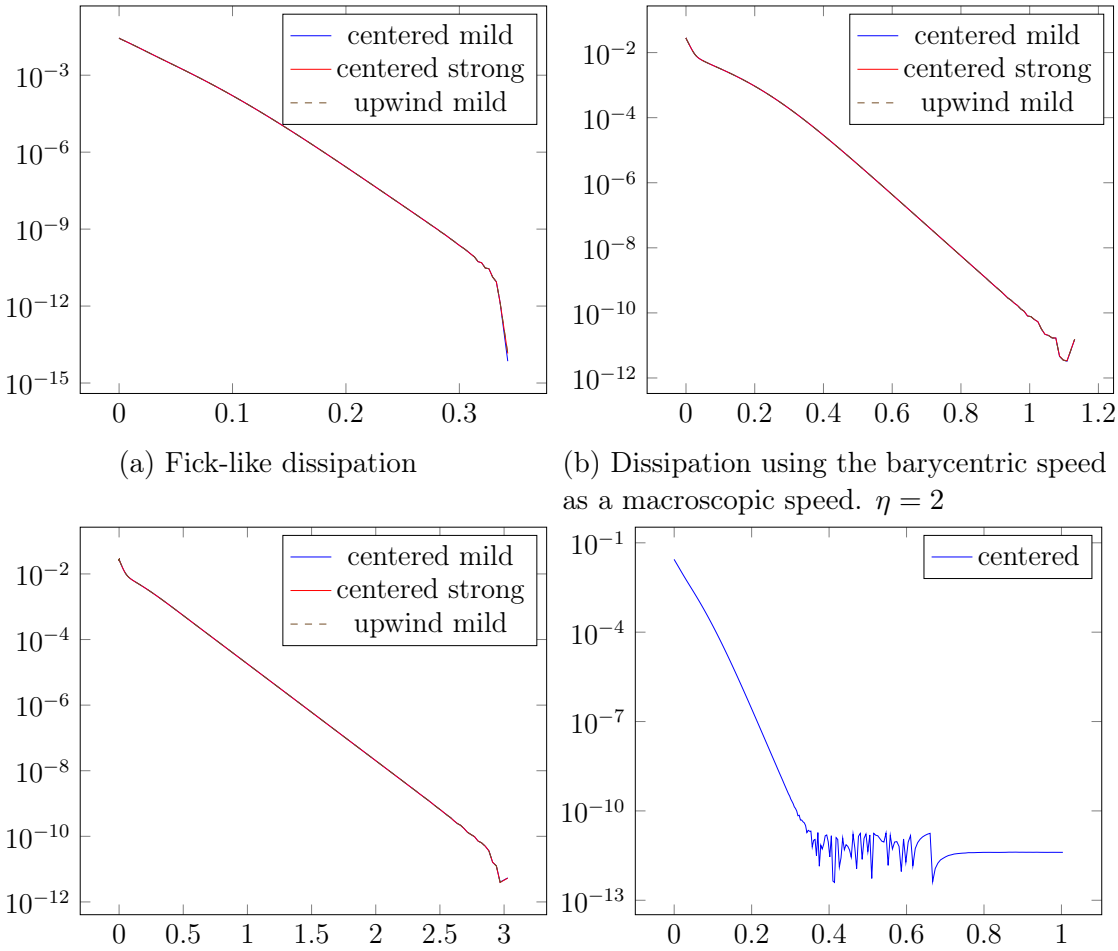


(c) Dissipation using the solvent speed as a macroscopic speed. $\eta = 2$



(d) Dissipation using the solvent free dissipation.

Figure 4.1 – Relative energy dissipation in the diluted one dimensional setting



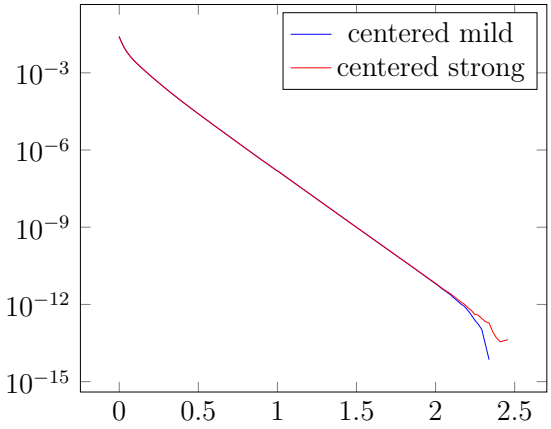
(a) Fick-like dissipation

(b) Dissipation using the barycentric speed as a macroscopic speed. $\eta = 2$

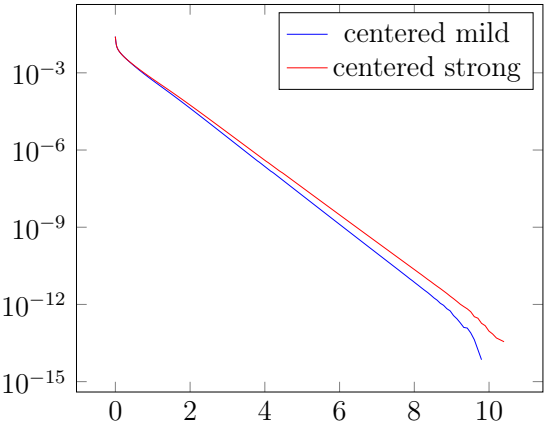
(c) Dissipation using the solvent speed as a macroscopic speed. $\eta = 2$

(d) Dissipation using the solvent free dissipation.

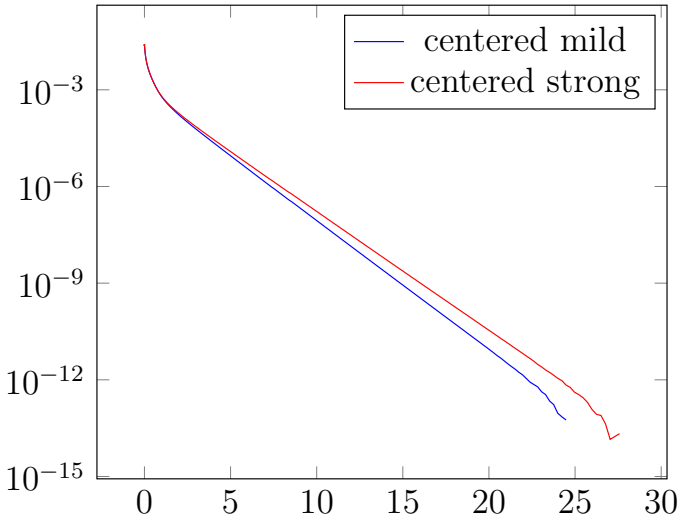
Figure 4.2 – Relative energy dissipation in the saturated one dimensional setting



(a) Fick-like dissipation

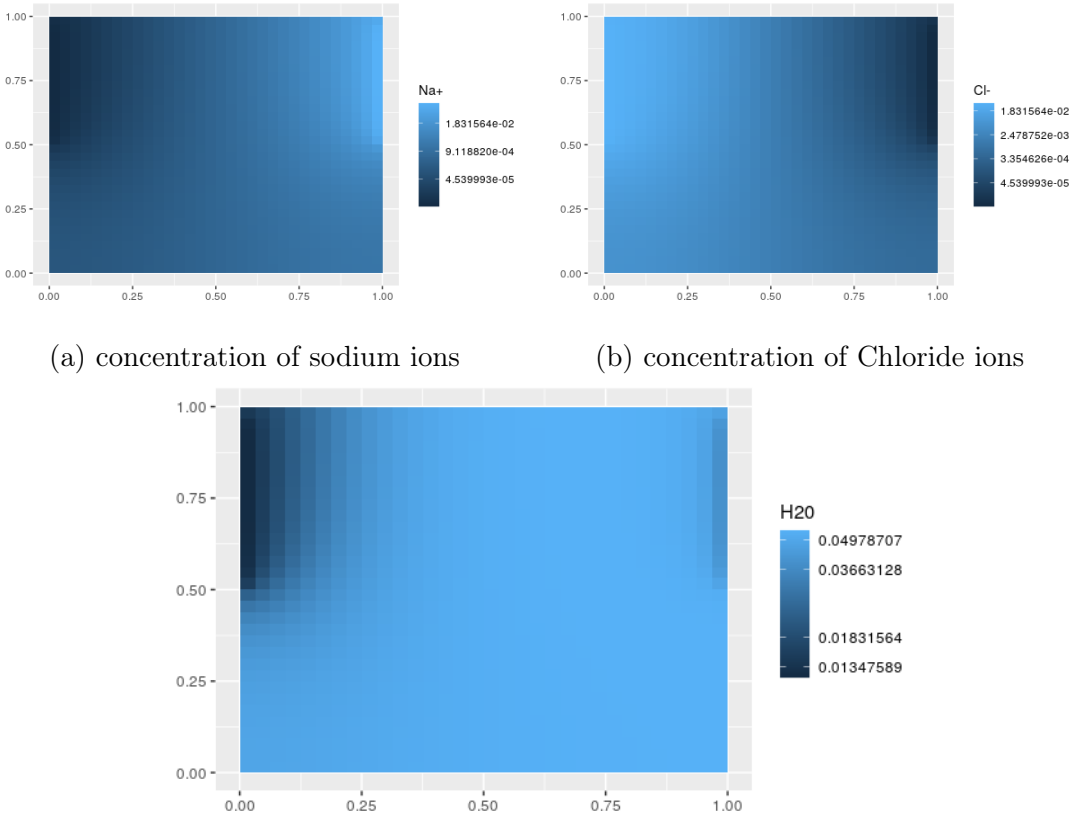


(b) Dissipation using the barycentric speed as a macroscopic speed. $\eta = 2$



(c) Dissipation using the solvent speed as a macroscopic speed. $\eta = 2$

Figure 4.3 – Relative energy dissipation in the saturated two dimensional setting



(a) concentration of sodium ions

(b) concentration of Chloride ions

(c) concentration of water

Figure 4.4 – Concentrations profiles of the saturated 2D stationary solution

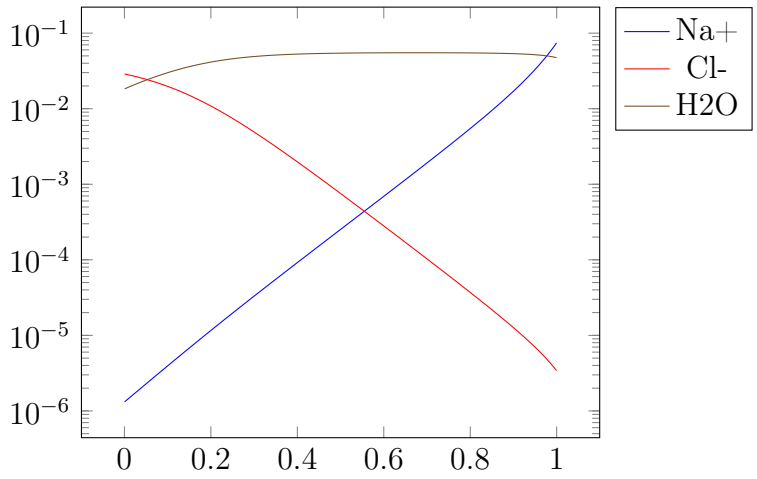


Figure 4.5 – concentration profile for the saturated 1D stationary solution

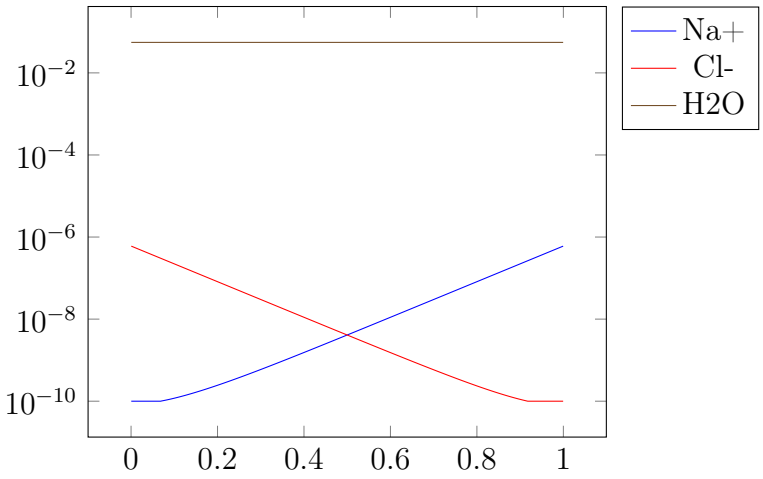
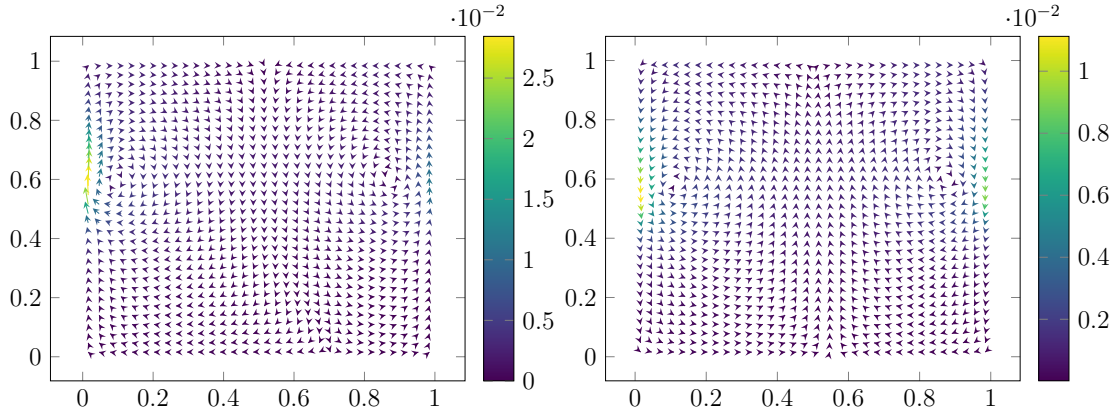
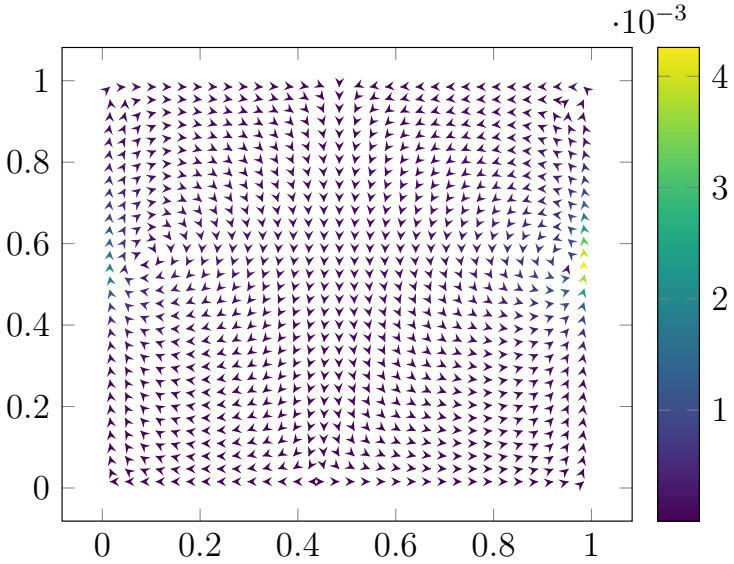


Figure 4.6 – concentration profile for the diluted 1D stationary solution



(a) Dissipation using the Fick-like dissipation potential

(b) Dissipation using the barycentric speed as a macroscopic speed. $\eta = 2$



(c) Dissipation using the solvent speed as a macroscopic speed. $\eta = 2$

Figure 4.7 – Convective speed profile for the saturated 1D solution at time 0.005

Chapter 5

A convergent entropy diminishing finite volume scheme for a cross-diffusion system

Ce chapitre est un travail en collaboration avec Clément Cances. Il a été publié dans SINUM [34].

We study a two-point flux approximation finite volume scheme for a cross-diffusion system. The scheme is shown to preserve the key properties of the continuous systems, among which the decay of the entropy. The convergence of the scheme is established thanks to compactness properties based on the discrete entropy - entropy dissipation estimate. Numerical results illustrate the behavior of our scheme.

Outline of the current chapter

5.1 Introduction **174**
5.1.1 The system under study 174
5.1.2 Formal structure 175
5.1.3 Objectives 179
5.2 Finite Volume approximation **181**
5.2.1 Discretization of $(0,T)\times\Omega$ 181
5.2.2 Numerical scheme 183
5.2.3 Main results and organization 184
5.3 Numerical analysis on a fixed mesh **186**
5.3.1 A priori estimates 186
5.3.2 Existence of solutions 188
5.3.3 Entropy dissipation 189
5.4 Convergence analysis **192**
5.4.1 Reconstruction operators 192
5.4.2 Compactness properties 193
5.4.3 Convergence towards a weak solution 197
5.5 Numerical results **199**
5.5.1 Convergence under grid refinement 199
5.5.2 On the influence of the parameter $a^\star$ 201
5.5.3 A 2D test case with reaction 201
5.6 Conclusion **203**

5.1 Introduction

5.1.1 The system under study

The system studied in this paper has been originally introduced by [11] to model the production of solar panels using vapor deposition. In this system, we study the diffusion of N species whose respective concentrations are $U=(u_1,\dots,u_N)$ in a (nonempty) connected bounded open domain Ω of $\mathbb{R}^d$ for a fixed time T. We denote by $Q_T=(0,T)\times\Omega$. The diffusion occurs through exchanges between different species which are quantified by the matrix $A=(a_{i,j})$ of cross-diffusion coefficients. It leads to the following system of partial differential equations:

$$\partial_t u_i - \operatorname{div}\left(\sum_{j=1}^{N} a_{i,j}\left(u_j\nabla u_i - u_i\nabla u_j\right)\right)=0 \quad \text{in } Q_T \text{ for } i\in[\![1,N]\!]. \tag{5.1.1}$$

The matrix A is assumed to be symmetric with nonnegative coefficients, i.e. $a_{i,j} = a_{j,i} \geq 0$. A does not depend on U and thus differs from the diffusion matrix $D(U) = (d_{i,j}(U))$ defined by

$$d_{i,j}(U) = \delta_{i,j} \sum_{k \neq i} a_{i,k} u_k - a_{i,j} u_i,$$

where $\delta_{i,j}$ stands for Kronecker symbol, such that the problem (5.1.1) rewrites

$$\partial_t U - \operatorname{div}\left(D(U)\nabla U\right) = 0. \tag{5.1.2}$$

System (5.1.2) enters the family of the nonlinear cross-diffusion systems since D depends on U and has nonzero off-diagonal entries. Challenges both from the analytical and numerical points of view come from the presence of off-diagonal zeros in A. In the previous contributions [27, 84, 31], the zeros are integrated through the assumption that the cross-diffusion occurs with and only with a solvent specie. Until Section 5.4 we will not make any assumption about the zeros of A. A non-degeneracy assumption will be further assumed in Section 5.4, but our convergence result could extend to the particular cross-diffusion matrices considered in [84, 31, 83].

We supplement system (5.1.1) with no-flux boundary conditions

$$\sum_{j=1}^{N} a_{i,j} \left(u_j \nabla u_i - u_i \nabla u_j\right) \cdot n = 0 \quad \text{on } (0,T) \times \partial\Omega, \quad i \in [\![1, N]\!]. \tag{5.1.3}$$

The initial concentration $U^0 = (u_1^0, \ldots, u_N^0)$ is supposed to be measurable and to map Ω into

$$\mathcal{A} = \left\{ U = (u_1, \ldots, u_N) \in \mathbb{R}_+^N \;\middle|\; \sum_{i=1}^{N} u_i = 1 \right\},$$

so we write in the condensed form $U^0 \in L^\infty(\Omega; \mathcal{A})$, which means that U^0 is measurable and takes its values in the bounded subset $\mathcal{A}$ of $\mathbb{R}^N$. Finally, we assume that all the chemical species under consideration are present:

$$\int_\Omega u_i^0 \mathrm{d}x > 0, \quad \forall i \in [\![1, N]\!]. \tag{5.1.4}$$

5.1.2 Formal structure

This system has several structural properties, the goal of this subsection is to exhibit them. The calculations presented in this section are formal: we assume

that the solutions to (5.1.1) enjoy enough regularity to justify the calculations below. Rigorous proofs at the continuous level for the system under consideration here can be found in [11, 14] (see also [84]). The properties listed here can also be obtained by passing to the limit in the numerical scheme. The first property we point out is the conservation of mass for all the species involved in System (5.1.1).

Lemma 5.1.1 (conservation of mass)**.** (5.1.1) *and* (5.1.3) *corresponding to an initial data* $U^0 \in L^\infty(\Omega;\mathcal{A})$*, then*

$$\int_\Omega u_i(t,x)\mathrm{d}x = \int_\Omega u_i^0(x)\mathrm{d}x, \quad \forall t \in [0,T],\ \forall i \in [\![1,N]\!].$$

Proof. Let U be a solution of (5.1.1), $t \in [0,T], i \in [\![1,N]\!]$, and let $\varphi(x,s) = 1_{[0,t]}(s)$. With this particular choice of φ, we have for all s that

$$\int_\Omega \mathrm{div}\left(\sum_{j=1}^N a_{i,j}\left(u_j\nabla u_i - u_i\nabla u_j\right)\right)\varphi(x,s)\mathrm{d}x$$
$$= -\int_\Omega \sum_{j=1}^N a_{i,j}\left(u_j\nabla u_i - u_i\nabla u_j\right)\nabla\varphi(x,s)\mathrm{d}x = 0.$$

Hence, using φ as a test function in (5.1.1), we have:

$$\int_0^t \frac{d}{ds}\left(\int_\Omega u_i(x,s)\mathrm{d}x\right)\mathrm{d}s = 0.$$

The fundamental theorem of calculus yields the desired lemma. □

The symmetry of the matrix $A = (a_{i,j})$ yields:

$$\sum_{i=1}^N\sum_{j=1}^N a_{i,j}\left(u_j\nabla u_i - u_i\nabla u_j\right) = 0.$$

Therefore, a solution U to (5.1.1) satisfies $\partial_t\sum_{i=1}^N u_i = 0$. Admit that $u_i(t,x) \geq 0$ for all $t > 0$ (this will be proved in the discrete setting and is proved in [14, Proposition 2.2] in the continuous setting), then the admissibility condition encoded in $\mathcal{A}$ is preserved along time.

Lemma 5.1.2. *Let* U *be a solution to* (5.1.1) *and* (5.1.3) *corresponding to an initial data* $U^0 \in L^\infty(\Omega;\mathcal{A})$*, then* $U(t,x) \in \mathcal{A}$ *for all* $(t,x) \in \mathcal{A}$*, i.e.,* $U \in L^\infty(Q_T;\mathcal{A})$*.*

The system can be derived by passing to the macroscopic limit from a random jump process in the spirit of [85, 26]. As expected because of this derivation from statistical physic considerations, the system fulfills Onsager's reciprocal relation [107, 108] and has a formal gradient flow structure. The driving functional is the mixing entropy

$$E : \begin{cases} L^\infty(\Omega; \mathcal{A}) \to \mathbb{R}, \\ U \mapsto \int_\Omega \sum_{i=1}^N u_i \log(u_i) \mathrm{d}x. \end{cases} \tag{5.1.5}$$

The next property we want to highlight at the continuous level is the decay of this entropy. Using the chain rule $\nabla c = c \nabla \log(c)$, the system (5.1.1) is formally equivalent to

$$\partial_t u_i - \operatorname{div}\left(\sum_{j=1}^N a_{i,j} u_i u_j \left(\nabla \log(u_i) - \nabla \log(u_j) \right) \right) = 0, \quad i \in [\![1, N]\!]. \tag{5.1.6}$$

Proposition 5.1.1. *E is a Lyapunov functional for the system (5.1.3)–(5.1.6). More precisely, the following entropy - entropy dissipation estimate holds:*

$$\frac{d}{dt} E(U) + \int_\Omega \left(\sum_{1 \le i < j \le N} a_{i,j} u_i u_j \left| \nabla \log(u_i) - \nabla \log u_j \right|^2 \right) \mathrm{d}x = 0. \tag{5.1.7}$$

Proof. First, we notice that thanks to the conservation of mass:

$$\frac{d}{dt} E(U) = \frac{d}{dt} \int_\Omega \sum_{i=1}^N u_i (\log(u_i) - 1) = \int_\Omega \sum_{i=1}^N \log(u_i) \partial_t u_i.$$

Then multiply Equation (5.1.6) by $\log(u_i)$ and integrate by part in order to get:

$$\int_\Omega \log(u_i) \partial_t u_i + \int_\Omega \left(\sum_{j=1}^N a_{i,j} u_i u_j \nabla \log(u_i) \cdot \left(\nabla \log(u_i) - \nabla \log(u_j) \right) \right) = 0.$$

Summing over $i \in [\![1, N]\!]$ yields the announced result thanks to the symmetry of A. □

The entropy - entropy dissipation relation (5.1.7) is key in the analysis of many cross-diffusion systems, as exposed in [91, 92]. It will also play a central role in this paper. Assume that

$$\min_{i \neq j} a_{i,j} > 0 \tag{5.1.8}$$

as it will be done in Section 5.4. As a consequence of the inequality

$$\sum_{i=1}^{N}\int_\Omega |\nabla u_i|^2 \le 4\sum_{i=1}^{N}\int_\Omega |\nabla\sqrt{u_i}|^2 \le \frac{1}{\min_{i\neq j} a_{i,j}}\int_\Omega \sum_{1\le i<j\le N} a_{i,j}u_iu_j\left|\nabla\log(u_i)-\nabla\log(u_j)\right|^2,$$

we deduce from (5.1.7) a $L^2(0,T;H^1(\Omega))^N$ estimate on U. This motivates the following notion of weak solution.

Definition 8. *A weak solution U to (5.1.1) and (5.1.3) corresponding to the initial profile $U^0\in L^\infty(\Omega;\mathcal{A})$ is a function of $L^\infty(Q_T;\mathcal{A})\cap L^2([0,T];H^1(\Omega))^N$ satisfying, $\forall i\in[\![1,N]\!]$, $\forall\varphi\in C_c^\infty([0,T)\times\overline{\Omega})$:*

$$-\iint_{Q_T} u_i\partial_t\varphi\mathrm{d}x\mathrm{d}t-\int_\Omega u_i^0\varphi(0,\cdot)\mathrm{d}x+\iint_{Q_T}\sum_{j=1}^{N} a_{i,j}\left(u_j\nabla u_i-u_i\nabla u_j\right)\nabla\varphi=0. \quad (5.1.9)$$

The regularity requirement on a weak solution U is natural in the setting where Assumption (5.1.8) holds. In this case, the solution even enjoys a stronger regularity, as established in the recent contribution [14]. In the case where (5.1.8) is not fulfilled (but under a structural assumption on the matrix A), a more involved notion of weak solution has to be introduced, cf. [84].

There is an important property that relates the model (5.1.1) to classical Fickian diffusion. As a consequence of Lemma 5.1.2, one can rewrite

$$\mathrm{div}\left(\sum_{j=1}^{N}\left(u_j\nabla u_i-u_i\nabla u_j\right)\right)=\Delta u_i,\qquad i\in[\![1,N]\!]. \quad (5.1.10)$$

As a consequence, if all the $a_{i,j}$ are equal to some $a\in\mathbb{R}$, then the system (5.1.1) reduces to N uncoupled heat equations $\partial_t u_i=a\Delta u_i$. Based on the identity (5.1.10), we can rewrite the system (5.1.1) under the form

$$\partial_t u_i-a^\star\,\Delta\,u_i-\mathrm{div}\left(\sum_{j=1}^{N}(a_{i,j}-a^\star)\left(u_j\nabla u_i-u_i\nabla u_j\right)\right)=0,\quad i\in[\![1,N]\!], \quad (5.1.11)$$

where $a^\star\in\mathbb{R}$ is arbitrary for the moment. The formulation (5.1.11) is at the basis of our discretization.

5.1.3 Objectives

The goal of this paper is to build and analyze a numerical scheme preserving the properties discussed in the previous section, namely:

— the non-negativity of the concentrations;

— the conservation of mass (Lemma 5.1.1);

— the preservation of the volume filling constraint (Lemma 5.1.2);

— the entropy-entropy dissipation relation (Proposition 5.1.1).

The construction of our scheme is the purpose of Section 5.2. In Section 5.3, we will show the existence of solutions to this scheme and the preservation of discrete counterparts to the previously listed physical properties. Section 5.4 is devoted to the convergence of the numerical scheme toward weak solutions provided Assumption (5.1.8) is satisfied. Finally, in Section 5.5, we show the outcomes of some numerical experiments.

Before entering the core of the paper, let us mention that the development of numerical analysis for cross-diffusion systems is quite recent. To our knowledge, the first convergence study of a finite volume approximation for a non-degenerate cross-diffusion problem was carried out in [8]. This contribution is based on classical quadratic energy estimate, similarly to what is proposed in [128]. The implementation of the discrete entropy method [47] for cross-diffusion systems is more recent. Let us cite [2, 4] where upstream mobility finite volume and control volume finite element schemes for a multiphase extension of the porous medium equation are studied. Upwinding is also used in [31] to approximate the solution of a system which is very close to the problem (5.1.1) under study, or in [45] for a problem in which nonlocal interactions are also considered. As a consequence of the upwind choice for the mobility, the schemes presented in [2, 4, 31] and [45] are first-order accurate in space. A natural solution to pass to order two is to rather consider mobilities given by arithmetic means [53]. The motivation of the finite element scheme proposed in [93] is also the same. However, the scheme proposed in [93] is expressed in entropy (or dual) variables (in our context $\log(u_i)$) leading to computational difficulties when the concentrations are close to 0. Other entropy stable numerical schemes have been proposed for cross-diffusion systems, as for instance discontinuous Galerkin schemes in [122], or finite volumes on staggered cartesian grids for Maxwell-Stefan cross-diffusion in [90]. Finally, let us point out that the design of entropy (or energy) stable numerical schemes for dissipative systems with formal gradient flow structure in a Riemannian geometry have been the purpose of intense research in the recent years, as shows the extensive (but not exhaustive) recent literature [36, 97, 39, 7, 106, 29, 38, 87, 10, 46] on this topic. Let us also refer to [49, 57, 119, 20, 118] for the simpler situation of gradient flows in Hilbert spaces.

Remark 5.1.1. Our study can be extended to the case where reaction terms are incorporated in the system. More precisely, one can consider a system of the form

$$\partial_t U - \operatorname{div}(D(U)\nabla U) = R(U), \tag{5.1.12}$$

where $D(U)$ is as in (5.1.2), and where the reaction term function

$$R : \begin{cases} \mathbb{R}^N \to \mathbb{R}^N \\ U \mapsto R(U) = (r_i(U))_{1\leq i\leq N} \end{cases}$$

is continuous and satisfies the following structural properties which are classically satisfied for reactive systems:

(i) Isochore process: $\sum_{i=1}^N r_i(U) = 0$ for all $U \in \mathbb{R}^N$;

(ii) Positivity preservation: $r_i(U) \geq 0$ for $U \in \mathbb{R}^N$ with $u_i \leq 0$;

(iii) Entropy dissipation: there exists $\overline{U} = (\overline{u}_i)_{1\leq i\leq N}$ in $\mathcal{A}$ with $\overline{u}_i > 0$ for all $i \in \{1, ;N\}$ such that

$$R(U) \cdot \log(U/\overline{U}) = \sum_{i=1}^N r_i(U)\left(\log(u_i) - \log(\overline{u}_i)\right) \leq 0, \qquad \forall U \in \mathcal{A}. \tag{5.1.13}$$

Because of reaction terms, the volume of each specie is no longer conserved, so that Lemma 5.1.1 does no longer hold true. However, because of Assumption ((i)) above, the total volume is conserved, hence the condition $\sum_{i=1}^N u_i(t,x) = 1$ remains true for all time. Since Assumption ((ii)) guarantees the positivity of the solution, one gets that $U(t,x)$ belongs to $L^\infty(Q_T, \mathcal{A})$ with the reaction term as well. Finally, Assumption ((iii)) on the reaction terms ensures that the relative entropy

$$E(U|\overline{U}) = \sum_{i=1}^N \int_\Omega u_i \log\left(\frac{u_i}{\overline{u}_i}\right) \geq 0 \tag{5.1.14}$$

is a Lyapunov functional for the system. This stability property allows to extend our purpose in presence of reaction. Note that in absence of reaction $R \equiv 0$, this relative entropy (one can for instance set $\overline{u}_i = ⨍_\Omega u_i^0$) coincides with the mixing entropy (5.1.5) up to an additive constant thanks to the conservation of the volume of each specie, cf. Lemma 5.1.1.

Finally, let us note that if $\overline{U} \in \mathcal{A}$ is such that $\overline{u}_i = 0$ for some $i \in \{1, \dots, N\}$, the relative entropy $E(U|\overline{U})$ is no longer well-defined. Our analysis can still be extended by showing that the mixing entropy $E(U)$ grows at most linearly with time, which is sufficient for establishing the convergence of the straightforward extension to the case $R \neq 0$ of the finite volume scheme to be presented in the

next section.

5.2 Finite Volume approximation

This section is organized as follows. First, in Section 5.2.1, we state the requirements on the mesh and fix some notations. Then in Section 5.2.2, we describe the numerical scheme to be studied in this paper. It is based on Formulation (5.1.11) of the problem. Then in Section 5.2.3, we state our two main results. The first one, namely Theorem 5.2.1, focuses on the case of a fixed mesh. We are interested in the existence of a solution to the nonlinear system corresponding to the scheme, and the dissipation of the entropy at the discrete level. More precisely, one establishes that the studied scheme satisfies a discrete entropy - entropy dissipation inequality that can be thought of as a counterpart to Proposition 5.1.1. Our second main result, namely Theorem 5.2.2, is devoted to the convergence of the scheme towards a weak solution as the time step and the mesh size tend to 0.

5.2.1 Discretization of $(0,T)\times\Omega$

The scheme we propose relies on two-point flux approximation (TPFA) finite volumes. As explained in [61, 67, 80], this approach appears to be very efficient as soon as the continuous problem to be solved numerically is isotropic and one has the freedom to choose a suitable mesh fulfilling the so-called orthogonality condition [89, 68]. We recall here the definition of such a mesh, which is illustrated in Figure 5.1.

Definition 9. *An admissible mesh of Ω is a triplet $\left(\mathcal{T},\mathcal{E},(x_K)_{K\in\mathcal{T}}\right)$ such that the following conditions are fulfilled.*

(i) Each control volume (or cell) $K\in\mathcal{T}$ is non-empty, open, polyhedral and convex. We assume that

$$K\cap L=\emptyset \quad \text{if } K,L\in\mathcal{T} \text{ with } K\neq L, \qquad \text{while} \quad \bigcup_{K\in\mathcal{T}}\overline{K}=\overline{\Omega}.$$

(ii) Each face $\sigma\in\mathcal{E}$ is closed and is contained in a hyperplane of $\mathbb{R}^d$, with positive $(d-1)$-dimensional Hausdorff (or Lebesgue) measure denoted by $m_\sigma=\mathcal{H}^{d-1}(\sigma)>0$. We assume that $\mathcal{H}^{d-1}(\sigma\cap\sigma')=0$ for $\sigma,\sigma'\in\mathcal{E}$ unless $\sigma'=\sigma$. For all $K\in\mathcal{T}$, we assume that there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K=\bigcup_{\sigma\in\mathcal{E}_K}\sigma$. Moreover, we suppose that $\bigcup_{K\in\mathcal{T}}\mathcal{E}_K=\mathcal{E}$. Given two distinct control volumes $K,L\in\mathcal{T}$, the intersection $\overline{K}\cap\overline{L}$ either reduces to a single face $\sigma\in\mathcal{E}$ denoted by $K|L$, or its $(d-1)$-dimensional Hausdorff measure is 0.

(iii) *The cell-centers* $(x_K)_{K\in\mathcal{T}}$ *satisfy* $x_K \in K$, *and are such that, if* $K, L \in \mathcal{T}$ *share a face* $K|L$, *then the vector* $x_L - x_K$ *is orthogonal to* $K|L$.

(iv) *For the boundary faces* $\sigma \subset \partial\Omega$, *we assume that either* $\sigma \subset \Gamma_D$ *or* $\sigma \subset \overline{\Gamma}_N$. *For* $\sigma \subset \partial\Omega$ *with* $\sigma \in \mathcal{E}_K$ *for some* $K \in \mathcal{T}$, *we assume additionally that there exists* $x_\sigma \in \sigma$ *such that* $x_\sigma - x_K$ *is orthogonal to* σ.

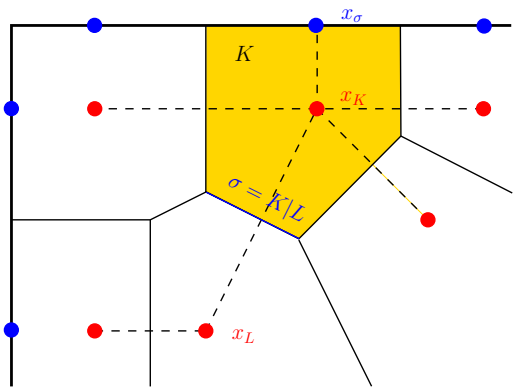


Figure 5.1 – Illustration of an admissible mesh as in Definition 9.

We denote by m_K the d-dimensional Lebesgue measure of the control volume K. The set of the faces is partitioned into two subsets: the set $\mathcal{E}_{\rm int}$ of the interior faces defined by $\mathcal{E}_{\rm int} = \{\sigma \in \mathcal{E} \mid \sigma = K|L \text{ for some } K, L \in \mathcal{T}\}$, and the set $\mathcal{E}_{\rm ext}$ of the exterior faces defined by $\mathcal{E}_{\rm ext} = \{\sigma \in \mathcal{E} \mid \sigma \subset \partial\Omega\}$. For a given control volume $K \in \mathcal{T}$, we also define $\mathcal{E}_{K,\rm int} = \mathcal{E}_K \cap \mathcal{E}_{\rm int}$ the set of its faces that belong to $\mathcal{E}_{\rm int}$. For such a face $\sigma \in \mathcal{E}_{K,\rm int}$, we may write $\sigma = K|L$, meaning that $\sigma = \overline{K} \cap \overline{L}$, where $L \in \mathcal{T}$.

Given $\sigma \in \mathcal{E}$, we let

$$d_\sigma = \begin{cases} |x_K - x_L| & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ |x_K - x_\sigma| & \text{if } \sigma \in \mathcal{E}_{\rm ext}, \end{cases} \qquad \text{and} \quad \tau_\sigma = \frac{m_\sigma}{d_\sigma}.$$

We finally introduce the size $h_\mathcal{T}$ and the regularity $\zeta_\mathcal{T}$ (which is assumed to be positive) of a discretization $(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}})$ of Ω by setting

$$h_\mathcal{T} = \max_{K\in\mathcal{T}} \operatorname{diam}(K), \qquad \zeta_\mathcal{T} = \min_{K\in\mathcal{T}} \min_{\sigma\in\mathcal{E}_K} \frac{d(x_K, \sigma)}{d_\sigma}.$$

Concerning the time discretization of $(0, T)$, we consider an increasing finite family of times $0 = t_0 < t_1 < \ldots, < t_{N_T} = T$. We denote by $\Delta t_n = t_n - t_{n-1}$ for $n \in \{1, \ldots, N_T\}$, by $\boldsymbol{\Delta t} = (\Delta t_n)_{1\le n\le N_T}$, and by $h_T = \max_{1\le n\le N_T} \Delta t_n$. In

what follows, we will use boldface notations for mesh-indexed families, typically for elements of $\mathbb{R}^{\mathcal{T}}$, $(\mathbb{R}^{\mathcal{T}})^N$, $(\mathbb{R}^{\mathcal{T}})^{N_T}$, or even $(\mathbb{R}^{\mathcal{T}})^{N\times N_T}$.

5.2.2 Numerical scheme

The initial data $U^0 \in L^\infty(\Omega;\mathcal{A})$ is discretized into

$$\boldsymbol{U}^0 = \left(\boldsymbol{u}_i^0\right)_{i\in[\![1,N]\!]} \in \left(\mathbb{R}^{\mathcal{T}}\right)^N = \left(u_{i,K}^0\right)_{K\in\mathcal{T},i\in[\![1,N]\!]}$$

by setting

$$u_{i,K}^0 = \frac{1}{m_K}\int_K u_i^0(x)\mathrm{d}x, \qquad \forall K\in\mathcal{T}, i\in[\![1,N]\!]. \tag{5.2.1}$$

Assume that $\boldsymbol{U}^{n-1} = \left(u_{i,K}^{n-1}\right)_{K\in\mathcal{T},i\in[\![1,N]\!]}$ is given for some $n\geq 1$, then we have to define how to compute $\boldsymbol{U}^n = \left(u_{i,K}^n\right)_{K\in\mathcal{T},i\in[\![1,N]\!]}$.

First, we introduce some notations. Given a discrete scalar field $\boldsymbol{c} = (c_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, we define for all cell $K\in\mathcal{T}$ and interface $\sigma\in\mathcal{E}_K$ the mirror value $c_{K\sigma}$ of c_K across σ by setting:

$$c_{K\sigma} = \begin{cases} c_L & \text{if } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ c_K & \text{if } \sigma\in\mathcal{E}_{\text{ext}}. \end{cases} \tag{5.2.2}$$

We also define the oriented and absolute jumps of $\boldsymbol{c}$ across any edge by

$$D_{K\sigma}\boldsymbol{c} = c_{K\sigma} - c_K, \quad D_\sigma \boldsymbol{c} = |D_{K\sigma}\boldsymbol{c}|, \qquad \forall K\in\mathcal{T},\ \forall\sigma\in\mathcal{E}_K.$$

The scheme is based on the formulation (5.1.11). It requires the introduction of a parameter $a^\star$ on which we only have the following requirements:

$$a^\star > 0 \qquad \text{and} \qquad a^\star \geq \min_{(i,j)} a_{i,j}. \tag{5.2.3}$$

The conservation laws are discretized in a conservative way with a time discretization relying on the backward Euler scheme:

$$m_K\frac{u_{i,K}^n - u_{i,K}^{n-1}}{\Delta t_n} + \sum_{\sigma\in\mathcal{E}_K} F_{i,K\sigma}^n = 0, \quad \forall K\in\mathcal{T},\ \forall i\in[\![1,N]\!]. \tag{5.2.4a}$$

The discrete fluxes are computed thanks to a formula based on (5.1.11) and on TPFA finite volumes:

$$F_{i,K\sigma}^n = -a^\star\tau_\sigma D_{K\sigma}\boldsymbol{u}_i^n - \tau_\sigma\left(\sum_{j=1}^N (a_{i,j}-a^\star)\left(u_{j,\sigma}^n D_{K\sigma}u_i^n - u_{i,\sigma}^n D_{K\sigma}u_j^n\right)\right), \tag{5.2.4b}$$

for all $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_K$ and $i \in [\![1, N]\!]$. Edge values $\left(u^n_{j,\sigma}\right)_j$ of the concentrations u_j appears in Formula (5.2.4b). It is deduced from $u^n_{j,K}$ and $u^n_{j,K\sigma}$ thanks to a logarithmic mean, i.e.,

$$u^n_{j,\sigma} = \begin{cases} 0 & \text{if } \min(u^n_{j,K}, u^n_{j,K\sigma}) \leq 0, \\ u^n_{j,K} & \text{if } 0 \leq u^n_{j,K} = u^n_{j,K\sigma}, \\ \dfrac{u^n_{j,K} - u^n_{j,K\sigma}}{\log(u^n_{j,K}) - \log(u^n_{j,K\sigma})} & \text{otherwise.} \end{cases} \tag{5.2.4c}$$

This choice for the edge concentration is crucial for the preservation at the discrete level of a discrete entropy - entropy dissipation inequality similar to the one highlighted in Proposition 5.1.1. Equations (5.2.4b) and (5.2.2) implies that for all $\sigma \in \mathcal{E}_{\text{ext}}$: $F^n_{i,K\sigma} = 0$, so that the no-flux boundary condition (5.1.3) is taken into account.

Remark 5.2.1. Let us highlight why the choice of a strictly positive $a^\star$ is important. Consider a mesh with two cells K, L, and one edge. We consider two species and let $u^0_K = (0, 1)$ and $u^0_L = (1, 0)$. We have: $u^0_{1,K|L} = 0$ and $u^0_{2,K|L} = 0$, hence, if $a^\star = 0$, the initial condition is a stationary solution even though this is not expected for a discretization of the heat equation. Setting $a^\star > 0$ eliminates these spurious solutions. The choice of $a^\star$ has a strong influence on the numerical outcomes, as it will be shown in Section 5.5, but we don't have a clear understanding yet on the methodology to choose an optimal $a^\star$. What seems clear is that $a^\star$ has to be chosen in the interval $[\min_{i\neq j} a_{i,j}, \max_{i\neq j} a_{i,j}]$. A tentative non-optimal formula is proposed in Section 5.5.

Remark 5.2.2. The time discretization in scheme (5.2.4) is only first-order accurate since it relies on the backward Euler approximation. Going to second-order time discretizations is tempting, but no theoretical guarantees concerning the entropy stability of the scheme can be granted then. This is due to the fact that the entropy (5.1.5) is not quadratic, hence neither the Crank-Nicolson scheme nor the BDF2 scheme can be shown to be unconditionally stable here. This lack of theoretical foundation for the entropy stability has for instance also been reported in [87].

5.2.3 Main results and organization

The first theorem proven is this paper concerns the existence of discrete solutions for a given mesh, and the preservation of the structural properties listed in Section 5.1.3:

— the mass of each specie is conserved along the time steps;

— the concentrations are (strictly) positive and sum to 1 in all the cells, i.e., $U_K^n \in \mathcal{A}$ for all $K \in \mathcal{T}$ and $n \geq 1$;

— the discrete counterpart of the entropy decays along time.

For this last property, we need to introduce the discrete entropy functional $E_{\mathcal{T}}$, which is defined by:

$$E_{\mathcal{T}}(\boldsymbol{U}) = \sum_{K\in\mathcal{T}} \sum_{i=1}^{N} m_K u_{i,K} \log u_{i,K}, \qquad \forall \boldsymbol{U} = (u_{i,K})_{K\in\mathcal{T}, i\in[\![1,N]\!]} \in \mathcal{A}^{\mathcal{T}}. \tag{5.2.5}$$

As stated in Theorem 5.2.1 below, the nonlinear system corresponding to our scheme (5.2.4) admits solutions that preserve the physical bounds on the concentrations and the decay of the entropy.

Theorem 5.2.1. *Let $(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}})$ be an admissible mesh and let $\boldsymbol{U}^0$ be defined by (5.2.1). Then, for all $1 \leq n \leq N_T$, the nonlinear system of equations (5.2.2) – (5.2.4), has a positive solution $\boldsymbol{U}^n \in \mathcal{A}^{\mathcal{T}}$. Moreover, such a solution satisfies $E_{\mathcal{T}}(\boldsymbol{U}^n) \leq E_{\mathcal{T}}(\boldsymbol{U}^{n-1})$ for all $n \in [\![1, N_T]\!]$, $\sum_{K\in\mathcal{T}} m_K u_{i,K}^n = \int_\Omega u_i^0$ for all $i \in [\![1, N]\!]$ and $n \in [\![0, N_T]\!]$.*

The proof of Theorem 5.2.1 will be the purpose of Section 5.3. With a discrete solution $(\boldsymbol{U}^n)_{1\leq n\leq N_T}$ to the scheme (5.2.4) at hand, we can define the piecewise constant approximate solution $U_{\mathcal{T},\boldsymbol{\Delta t}} = (u_{i,\mathcal{T},\boldsymbol{\Delta t}})_{i\in[\![1,N]\!]} : Q_T \to \mathcal{A}$ defined almost everywhere by

$$U_{\mathcal{T},\boldsymbol{\Delta t}}(t,x) = U_K^n \quad \text{if } (t,x) \in (t_{n-1}, t_n] \times K.$$

This definition will be developed in Section 5.4 and supplemented by other reconstruction operators. Let $\left(\mathcal{T}_m, \mathcal{E}_m, (x_K)_{K\in\mathcal{T}_m}\right)_{m\geq 1}$ be a sequence of admissible discretizations with $h_{\mathcal{T}_m}, h_{T,m}$ tending to 0 as m tends to $+\infty$, while the regularity $\zeta_{\mathcal{T}_m}$ remains uniformly bounded from below by a positive constant $\zeta^\star$. Thanks to Theorem 5.2.1, we dispose of a family $\boldsymbol{U}_m$ of solutions to our scheme. The convergence of $\boldsymbol{U}_m$ is the purpose of Theorem 5.2.2 whose proof is detailed in Section 5.4.

Theorem 5.2.2. *Assume that the nondegeneracy assumption (5.1.8) holds. Given any sequence of solutions $\boldsymbol{U}_m = \left(u_{i,K}^n\right)_{i\in[\![1,N]\!], K\in\mathcal{T}_m, 1\leq n\leq N_{T,m}}$, there exists at least one $U \in L^\infty(Q_T; \mathcal{A}) \cap L^2((0,T); H^1(\Omega))$ such that, up to a subsequence,*

$$\boldsymbol{U}_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} U \quad \textit{strongly in } L^p(Q_T), \textit{ for any } 1 \leq p < \infty, \tag{5.2.6}$$

Moreover, U is a weak solution in the sense of Definition 8.

Remark 5.2.3. Theorem 5.2.2 above establishes the convergence of the scheme, but no error estimate can be deduced from its proof. Indeed, its proof can be thought of as an adaptation to the discrete setting of an existence proof based on compactness arguments, as the for instance proposed in [91]. The derivation of error estimates is different since it relies on the perturbation of uniqueness proofs. As far as we know, uniqueness for the problem under consideration is an open question excepted in the one-dimensional setting [14], but a natural approach would be to derive error estimates based on the use of the relative entropy, as done for instance in [37] for hyperbolic systems or in [77, 79] for the compressible Navier-Stokes problem. However, the aforementioned strategy generally leads to under-optimal error estimates. The recovery of optimal error estimates for finite volume approximation of diffusion equations on unstructured grids has only been achieved recently [65]. The extension of this optimal result to our much more complex cross-diffusion system appears to be an interesting and challenging issue.

5.3 Numerical analysis on a fixed mesh

This section is devoted to the proof of Theorem 5.2.1. In Section 5.3.1, we establish a priori estimates on a slightly modified scheme that will be shown to reduce to the original scheme (5.2.4). Then in Section 5.3.2, we apply a topological degree argument to prove the existence of solutions to our scheme. Section 5.3.3 is devoted to the proof of the entropy dissipation property.

To prove the existence of solutions to the system of equations (5.2.4), we need the inequality $\sum_i u_{i,\sigma} \leq 1$. We then slightly modify (5.2.4) by adding the following equation:

$$\widetilde{u}^n_{i,\sigma} = \frac{u^n_{i,\sigma}}{\max(1, \sum_{j=1}^N u^n_{j,\sigma})},$$

and replacing $u^n_{i,\sigma}$ by $\widetilde{u}^n_{i,\sigma}$ in (5.2.4b). We will denote this new system (S) and see in Proposition 5.3.2 that its solutions satisfy $\sum_i u_{i,\sigma} \leq 1$, so that $\widetilde{u}^n_{i,\sigma} = u^n_{i,\sigma}$. Whence they also satisfy the original system of equations.

5.3.1 A priori estimates

The first lemma shows the nonnegativity of the solutions to (S).

Lemma 5.3.1. *Given a nonnegative $\boldsymbol{U}^{n-1}$, any solution $\boldsymbol{U}^n$ to (S) is also nonnegative.*

Proof. Let $\boldsymbol{U}^n$ be a solution of (S) and let $i \in [\![1, N]\!]$. We consider a cell $K \in \mathcal{T}$ where $\boldsymbol{u}^n_i$ reaches it's minimum, i.e., $u^n_{i,K} \leq u^n_{i,L}$ for all $L \in \mathcal{T}$, and assume for

contradiction that $u_{i,K}^n$ is (strictly) negative. Equation (5.2.4b) then gives:

$$m_K \frac{u_{i,K}^n - u_{i,K}^{n-1}}{\Delta t_n} = -\sum_{\sigma \in \mathcal{E}_K} F_{K\sigma}^n.$$

The term on the left hand side is negative since $u_{i,K}^{n-1} \geq 0 > u_{i,K}^n$, whereas the right-hand side may be simplified noticing that $\widetilde{u}_{i,\sigma}^n = 0$:

$$\sum_{\sigma \in \mathcal{E}_K} a^\star \tau_\sigma D_{K\sigma} \boldsymbol{u}_i^n + \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma \sum_{j=1}^N (a_{i,j} - a^\star) \widetilde{u}_{j,\sigma}^n D_{K\sigma} \boldsymbol{u}_i^n = -\sum_{\sigma \in \mathcal{E}_K} F_{K\sigma}^n < 0.$$

Noticing that $D_{K\sigma} \boldsymbol{u}_i^n \geq 0$, $\widetilde{u}_{j,\sigma}^n \geq 0$, and $\sum_{j=1}^N \widetilde{u}_{j,\sigma}^n \leq 1$ we obtain that

$$0 \leq \sum_{\sigma \in \mathcal{E}_K} a^\star (1 - \sum_{j=1}^N \widetilde{u}_{j,\sigma}^n) \tau_\sigma D_{K\sigma} \boldsymbol{u}_i^n < 0,$$

which is absurd, hence the desired result. □

Let us now show that the concentrations sum to 1 in all the cells.

Lemma 5.3.2. *Given $\boldsymbol{U}^{n-1}$ in $\mathcal{A}^{\mathcal{T}}$, any solution $\boldsymbol{U}^n$ to (S) is also in $\mathcal{A}^{\mathcal{T}}$.*

Proof. Thanks to Lemma 5.3.1, it suffices to show that $\sum_{i=1}^N u_{i,K}^n = 1$ for all $K \in \mathcal{T}$. Let $\boldsymbol{U}^n$ be a solution to (S). Using (5.2.4b) in (5.2.4a) and summing over the species leads to:

$$\begin{aligned} &\frac{\sum_{i=1}^N u_{i,K}^n - \sum_{i=1}^N u_{i,K}^{n-1}}{\Delta t_n} m_K - a^\star \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \sum_i \boldsymbol{u}_i \\ &\quad - \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma \sum_i \left(\sum_{j=1}^N (a_{i,j} - a^\star) \left(\widetilde{u}_{j,\sigma}^n D_{K\sigma} \boldsymbol{u}_i - \widetilde{u}_{i,\sigma}^n D_{K\sigma} \boldsymbol{u}_j \right) \right) = 0, \quad \forall K \in \mathcal{T}. \end{aligned}$$

The third term of the left-hand side vanishes thanks to the symmetry of A, so that

$$\frac{\sum_{i=1}^N u_{i,K}^n - \sum_{i=1}^N u_{i,K}^{n-1}}{\Delta t_n} m_K - a^\star \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \sum_i \boldsymbol{u}_i = 0, \quad \forall K \in \mathcal{T}.$$

The discrete quantity $\sum_i \boldsymbol{u}_i$ is solution to the classical backward Euler TPFA scheme for the heat equation, which is well posed. So $\sum_i \boldsymbol{u}_i^n = \sum_i \boldsymbol{u}_i^{n-1} = \mathbf{1}$ is its unique solution, hence the desired result. □

5.3.2 Existence of solutions

Using the tools exposed in the previous subsection, we may derive the existence of a solution to (S):

Proposition 5.3.1. *Given $\boldsymbol{U}^{n-1}$ in $\mathcal{A}^{\mathcal{T}}$, there exists at least one solution to (S) in $\mathcal{A}^{\mathcal{T}}$.*

Proof. The proof relies on a topological degree argument [98, 56]. The idea is to transform continuously our complex nonlinear system into a linear system while guaranteeing that the a priori estimates controlling the solution remain valid all along the homotopy. We sketch the main ideas of the proof, making the homotopy explicit. We are interested in the existence of zeros for a functional

$$\boldsymbol{\mathcal{H}} : \begin{cases} [0,1] \times (\mathbb{R}^N)^{\mathcal{T}} \to (\mathbb{R}^N)^{\mathcal{T}} \\ (\lambda, \boldsymbol{U}) \mapsto \boldsymbol{\mathcal{H}}(\lambda, \boldsymbol{U}) \end{cases}$$

that boils down to the scheme (S) when $\lambda = 1$. In our case, we set:

$$\begin{aligned} \boldsymbol{\mathcal{H}}(\lambda, \boldsymbol{U})_{i,K} &= \frac{u_{i,K} - u_{i,K}^{n-1}}{\Delta t_n} m_K - a^\star \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D_{K\sigma} \boldsymbol{u}_i \\ &- \lambda \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma \left(\sum_{j=1}^{N} (a_{i,j} - a^\star) \left(\widetilde{u}_{j,\sigma} D_{K\sigma} \boldsymbol{u}_i - \widetilde{u}_{i,\sigma} D_{K\sigma} \boldsymbol{u}_j \right) \right), \quad \forall K \in \mathcal{T}, \forall i \in [\![1, N]\!]. \end{aligned} \tag{5.3.1}$$

One notices that $\boldsymbol{\mathcal{H}}(0, \boldsymbol{U}) = \boldsymbol{0}$ is the classical heat equation, the solution of which belongs to $\mathcal{A}^{\mathcal{T}}$. Therefore, fixing $\eta > 0$, the relatively compact open set

$$\mathcal{A}_\eta^{\mathcal{T}} = \left\{ \boldsymbol{U} \in \mathbb{R}^{\mathcal{T}} \;\middle|\; \inf_{\boldsymbol{V} \in \mathcal{A}^{\mathcal{T}}} \|\boldsymbol{U} - \boldsymbol{V}\| < \eta \right\}$$

has a topological degree equal to 1. Note that the choice of the norm in the definition of $\mathcal{A}_\eta^{\mathcal{T}}$ is not important since the dimension is finite. Moreover, thanks to Lemma 5.3.2, the solutions $\boldsymbol{u}^{(\lambda)}$ of $\boldsymbol{\mathcal{H}}(\lambda, \boldsymbol{U}) = \boldsymbol{0}$ remains in $\mathcal{A}^{\mathcal{T}}$, thus in the interior of $\mathcal{A}_\eta^{\mathcal{T}}$. Thus the topological degree of $\mathcal{A}_\eta^{\mathcal{T}}$ for $\lambda = 1$ is still equal to 1, hence the existence of (at least) one solutions to (S). Since $\eta > 0$ is arbitrary, then there is a solution in $\mathcal{A}^{\mathcal{T}} = \bigcap_{\eta > 0} \mathcal{A}_\eta^{\mathcal{T}}$. □

To prove the Theorem 5.2.1, we need to transfer this existence result on the original system.

Proposition 5.3.2. *A solution $\boldsymbol{U}^n$ of (S) is a solution of (5.2.4). Reciprocally, a solution of (5.2.4) in $\mathcal{A}^{\mathcal{T}}$ is a solution of (S).*

Proof. Let $\boldsymbol{U}^n$ be a solution of (S). A simple convexity argument shows that the logarithmic mean of two nonnegative number is smaller than the arithmetic mean, so that $u^n_{i,\sigma} \leq \frac{u^n_{i,K}+u^n_{i,K\sigma}}{2}$. Summing w.r.t. $i \in [\![1, N]\!]$ and using that the solution $\boldsymbol{U}$ of (S) belongs to $\mathcal{A}^{\mathcal{T}}$, one gets that $\sum_i u^n_{i,\sigma} \leq 1$ for all $\sigma \in \mathcal{E}$. Therefore $\widetilde{u}^n_{i,\sigma} = u^n_{i,\sigma}$ and $\boldsymbol{U}^n$ is a also solution to (5.2.4). The proof of the reverse implication follows the same lines. □

5.3.3 Entropy dissipation

We intend here to prove a discrete counterpart to Proposition 5.1.1. The proof will be very similar and requires a discrete counterpart of the conservation of mass (Lemma 5.1.1).

Lemma 5.3.3. *Given any $\boldsymbol{U}^{n-1} \in \mathcal{A}^{\mathcal{T}}$, any solution $\boldsymbol{U}^n$ to (5.2.4) satisfies:*

$$\sum_{K\in\mathcal{T}} m_K u^n_{i,K} = \sum_{K\in\mathcal{T}} m_K u^{n-1}_{i,K} = \int_\Omega u^0_i \mathrm{d}x, \quad \forall i \in [\![1, N]\!].$$

The proof of this lemma is a straightforward calculation based on equation (5.2.4a), the conservativity of the fluxes, and the definition (5.2.1) of the discrete initial condition. With this lemma and Proposition 5.3.2, we can refine the result Lemma 5.3.1 to get the strict positivity of any solution to (5.2.4) belonging to $\mathcal{A}^{\mathcal{T}}$.

Lemma 5.3.4. *Let $\boldsymbol{U}^{n-1} \in \mathcal{A}^{\mathcal{T}}$ be such that $\sum_K m_K u^{n-1}_{i,K} > 0$ for all $i \in [\![1, N]\!]$, then any solution to (5.2.4) in $\mathcal{A}^{\mathcal{T}}$ is positive: $u^n_{i,K} > 0$ for all $i \in [\![1, N]\!]$ and all $K \in \mathcal{T}$.*

Proof. Let $\boldsymbol{U}^n \in \mathcal{A}^{\mathcal{T}}$ be a solution to the scheme (5.2.4), and let $i \in [\![1, N]\!]$. We know from Lemma 5.3.1 that $\boldsymbol{u}^n_i \geq \boldsymbol{0}$. Assume for contradiction that there exists one cell K such that $u^0_{i,K}$ vanishes. Using Lemma 5.3.3 and the connectivity of Ω, there exists $\sigma = K|L \in \mathcal{E}^{\mathrm{int}}$ such that $u^n_{i,K} = 0$ and $u^n_{i,L} > 0$. Then $u^n_{i,\sigma} = 0$ and as in the proof of Lemma 5.3.1:

$$a^\star(1 - \sum_{j=1}^N u^n_{j,\sigma})\tau_\sigma D_{K\sigma}\boldsymbol{u}^n_i \leq 0.$$

Using $u^n_{j,\sigma} \leq \frac{u^n_{j,K}+u^n_{j,L}}{2}$ and $u_{i,\sigma} = 0$ we deduce that

$$\sum_{j=1}^{N} u^n_{j,\sigma} \leq \sum_{j\neq i}^{N} \frac{u^n_{j,K} + u^n_{j,L}}{2} \leq 1 - \frac{u^n_{i,L}}{2} < 1.$$

Therefore $a^\star(1 - \sum_{j=1}^N \widetilde{u}^n_{j,\sigma})\tau_\sigma > 0$, and since $D_{K\sigma}\boldsymbol{u}^n_i > 0$, we deduce that:

$$0 < a^\star(1 - \sum_{j=1}^{N} \widetilde{u}^n_{j,\sigma})\tau_\sigma D_{K\sigma}\boldsymbol{u}^n_i \leq 0.$$

As this statement is absurd, our assumption was false, hence the desired result. □

As in the continuous case, we will use the conservation of mass (Lemma 5.3.3) and a discrete equivalent of the chain rule $\nabla c = c\nabla \log c$. This equivalent writes

$$D_{K\sigma}\boldsymbol{u}^n_i = u^n_{i,\sigma} D_{K\sigma} \log(\boldsymbol{u}^n_i), \quad \forall i \in [\![1, N]\!],\ \forall K \in \mathcal{T}. \tag{5.3.2}$$

The above discrete chain rule follows from the definition (5.2.4c) of $u^n_{i,\sigma}$ and the positivity of solutions to (5.2.4) which gives a sense to $\log(\boldsymbol{u}^n_i)$.

Using (5.3.2) in (5.2.4b), $\boldsymbol{U}^n$ satisfies

$$\frac{u^n_{i,K} - u^{n-1}_{i,K}}{\Delta t_n} m_K - \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma \left(\sum_{j=1}^{N} (a_{i,j} - a^\star) u^n_{i,\sigma} u^n_{j,\sigma} \left(D_{K\sigma} \log(\boldsymbol{u}_i) - D_{K\sigma} \log(\boldsymbol{u}_j) \right) \right) - a^\star \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D_{K\sigma}\boldsymbol{u}^n_i = 0, \quad \forall K \in \mathcal{T}, \forall i \in [\![1, N]\!]. \tag{5.3.3}$$

This reformulation is suitable for proving a discrete entropy - entropy dissipation inequality, which should be seen as a discrete counterpart of Proposition 5.1.1.

Proposition 5.3.3. *Given $\boldsymbol{U}^{n-1}$ in $\mathcal{A}^\mathcal{T}$, any solution $\boldsymbol{U}^n \in \mathcal{A}^\mathcal{T}$ to (5.2.4) satisfies*

$$E_\mathcal{T}(\boldsymbol{U}^n) - E_\mathcal{T}(\boldsymbol{U}^{n-1}) + \Delta t_n \min_{1\leq i,j\leq N} a_{i,j} \sum_{\sigma\in\mathcal{E}} \sum_{i=1}^{N} \tau_\sigma u^n_{i,\sigma} (D_{K\sigma} \log(\boldsymbol{u}^n_i))^2 \leq 0. \tag{5.3.4}$$

In particular, $E_\mathcal{T}(\boldsymbol{U}^n) \leq E_\mathcal{T}(\boldsymbol{U}^{n-1})$.

Proof. Multiplying equation (5.3.3) by $\Delta t_n \log(u^n_{i,K})$ and summing over the cells

and species leads to:

$$\sum_{K\in\mathcal{T}}\sum_{i=1}^{N}(u_{i,K}^n\log(u_{i,K}^n)-u_{i,K}^{n-1}\log(u_{i,K}^n))m_K+\Delta t_n a^\star\sum_{\sigma\in\mathcal{E}}\sum_{i=1}^{N}\tau_\sigma u_{i,\sigma}^n(D_{K\sigma}\log(\boldsymbol{u}_i^n))^2$$
$$-\Delta t_n\sum_{K\in\mathcal{T}}\sum_{i=1}^{N}\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma\left(\sum_{j=1}^{N}(a_{i,j}-a^\star)u_{j,\sigma}^n u_{i,\sigma}^n\log(u_{i,K}^n)D_{K\sigma}\big(\log(\boldsymbol{u}_i^n)-\log(\boldsymbol{u}_j^n)\big)\right)=0. \tag{5.3.5}$$

Using the symmetry of the matrix A and discrete integration by part, both in space and with respect to the species, we have:

$$\sum_{K\in\mathcal{T}}\sum_{i=1}^{N}\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma\left(\sum_{j=1}^{N}(a_{i,j}-a^\star)u_{j,\sigma}^n u_{i,\sigma}^n\log(u_{i,K}^n)D_{K\sigma}\big(\log(\boldsymbol{u}_i^n)-\log(\boldsymbol{u}_j^n)\big)\right)=$$
$$-\sum_{\sigma\in\mathcal{E}}\tau_\sigma\left(\sum_{1\le i<j\le N)}(a_{i,j}-a^\star)u_{j,\sigma}^n u_{i,\sigma}^n\Big(D_{K\sigma}\big(\log(\boldsymbol{u}_i^n)-\log(\boldsymbol{u}_j^n)\big)\Big)^2\right). \tag{5.3.6}$$

On the other hand, the convexity of $c\log(c)$ yields:

$$u_{i,K}^n-u_{i,K}^{n-1}+u_{i,K}^n\log(u_{i,K}^n)-u_{i,K}^{n-1}\log(u_{i,K}^n)\ge u_{i,K}^n\log(u_{i,K}^n)-u_{i,K}^{n-1}\log(u_{i,K}^{n-1}).$$

Combining this inequality with Equation (5.3.6) and Lemma 5.3.3 in (5.3.5) provides:

$$E_{\mathcal{T}}(\boldsymbol{U}^n)-E_{\mathcal{T}}(\boldsymbol{U}^{n-1})+\Delta t_n a^\star\sum_{\sigma\in\mathcal{E}}\sum_{i=1}^{N}\tau_\sigma u_{i,\sigma}^n(D_{K\sigma}\log(\boldsymbol{u}_i^n))^2$$
$$+\Delta t_n\sum_{\sigma\in\mathcal{E}}\tau_\sigma\left(\sum_{1\le i<j\le N}(a_{i,j}-a^\star)u_{j,\sigma}^n u_{i,\sigma}^n\Big(D_{K\sigma}\big(\log(\boldsymbol{u}_i^n)-\log(\boldsymbol{u}_j^n)\big)\Big)^2\right)\le 0.$$

Using the hypothesis $0\le\min a_{i,j}\le a^\star$ together with

$$\sum_{i=1}^{N}u_{i,\sigma}^n(D_{K\sigma}\log(\boldsymbol{u}_i^n))^2-\left(\sum_{1\le i<j\le N}u_{j,\sigma}^n u_{i,\sigma}^n\Big(D_{K\sigma}\big(\log(\boldsymbol{u}_i^n)-\log(\boldsymbol{u}_j^n)\big)\Big)^2\right)=$$
$$\sum_{i=1}^{N}u_{i,\sigma}^n\left(1-\sum_{j=1}^{N}u_{j,\sigma}^n\right)(D_{K\sigma}\log(\boldsymbol{u}_i^n))^2\ge 0, \tag{5.3.7}$$

we deduce that (5.3.4) holds. □

The proof of Theorem 5.2.1 is now complete.

5.4 Convergence analysis

The goal of this Section is to prove Theorem 5.2.2, which states the convergence of the approximate solution towards a weak solution to the continuous problem in the sense of Definition 8 under the nondegeneracy condition (5.1.8). We could extend this result on several other special cases including the one treated in [31]. We hint that the optimal assumption would be that the zeros of the diffusion matrix form a cluster-graph. However, we stick to the study of the non-degenerate case for the sake of simplicity.

We consider here a sequence $\left(\mathcal{T}_m, \mathcal{E}_m, (x_K)_{K\in\mathcal{T}_m}\right)_{m\geq 1}$ of admissible discretizations with $h_{\mathcal{T}_m}, h_{T,m}$ tending to 0 as m tends to $+\infty$, while the regularity $\zeta_{\mathcal{T}_m}$ remains uniformly bounded from below by a positive constant $\zeta^\star$. Theorem 5.2.1 provides the existence of a family of discrete solutions $\boldsymbol{U}_m = \left(u_{i,K}^n\right)_{i\in[\![1,N]\!],K\in\mathcal{T}_m,1\leq n\leq N_m}$. To prove Theorem 5.2.2, we first establish in Section 5.4.2 some compactness properties on the family of piecewise constant approximate solutions $U_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}$. Then we identify the limit as a weak solution in Section 5.4.3. In order to enlighten the notations, we remove the subscript m as soon as it is not necessary for understanding.

5.4.1 Reconstruction operators

To carry out the convergence analysis, we introduce some reconstruction operators following the methodology proposed in [64]. The operators $\pi_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}} \to L^\infty(\Omega)$ and $\pi_{\mathcal{T},\boldsymbol{\Delta t}} : \left(\mathbb{R}^{\mathcal{T}}\right)^{N_T} \to L^\infty(Q_T)$ are defined respectively by

$$\pi_{\mathcal{T}}\boldsymbol{f}(x) = f_K \quad \text{if } x \in K, \qquad \forall \boldsymbol{f} = (f_K)_{K\in\mathcal{T}},$$

and

$$\pi_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{f}(t,x) = f_K^n \quad \text{if } (t,x) \in (t_{n-1},t_n]\times K, \qquad \forall \boldsymbol{f} = (f_K^n)_{K\in\mathcal{T},1\leq n\leq N_T}.$$

These operators allow to pass from the discrete solution $(\boldsymbol{U}^n)_{1\leq n\leq N_T}$ to the approximate solution since

$$u_{i,\mathcal{T},\boldsymbol{\Delta t}} = \pi_{\mathcal{T},\boldsymbol{\Delta t}}\left(\boldsymbol{u}_i^n\right)_n, \quad \forall i \in [\![1,N]\!].$$

In order to carry out the analysis, we further need to introduce approximate

gradient reconstruction. For $\sigma = K|L \in \mathcal{E}_{\rm int}$, we denote by Δ_σ the diamond cell corresponding to σ, which is the interior of the convex hull of $\{\sigma, x_K, x_L\}$. For $\sigma \in \mathcal{E}_{\rm ext}$, the diamond cell Δ_σ is defined as the interior of the convex hull of $\{\sigma, x_K\}$. The approximate gradient $\nabla_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}} \to L^2(\Omega)^d$ we use in the analysis is merely weakly consistent (unless $d = 1$) and takes its source in [48, 66]. It is piecewise constant on the diamond cells Δ_σ, and it is defined as follows:

$$\nabla_{\mathcal{T}} \boldsymbol{f}(x) = d \frac{D_{K\sigma} \boldsymbol{f}}{d_\sigma} n_{K\sigma} \quad \text{if } x \in \Delta_\sigma, \qquad \forall \boldsymbol{f} \in \mathbb{R}^{\mathcal{T}},$$

where $n_{K\sigma}$ is the outer-pointing normal of K at σ. We also define $\nabla_{\mathcal{T},\boldsymbol{\Delta t}} : \mathbb{R}^{\mathcal{T}\times N_T} \to L^2(Q_T)^d$ by setting

$$\nabla_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{f}(t,\cdot) = \nabla_{\mathcal{T}} \boldsymbol{f}^n \quad \text{if } t \in (t_{n-1}, t_n], \qquad \forall \boldsymbol{f} = (\boldsymbol{f}^n)_{1\le n\le N_T} \in \mathbb{R}^{\mathcal{T}\times N_T}.$$

It follows from the definition of the approximate gradient that

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K\sigma} \boldsymbol{f} D_{K\sigma} \boldsymbol{g} = \frac{1}{d} \int_\Omega \nabla_{\mathcal{T}} \boldsymbol{f} \cdot \nabla_{\mathcal{T}} \boldsymbol{g} \mathrm{d}x, \qquad \forall \boldsymbol{f}, \boldsymbol{g} \in \mathbb{R}^{\mathcal{T}}. \tag{5.4.1}$$

This implies in particular that

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma |D_\sigma \boldsymbol{f}|^2 = \frac{1}{d} \int_\Omega |\nabla_{\mathcal{T}} \boldsymbol{f}|^2 \mathrm{d}x, \qquad \forall \boldsymbol{f} \in \mathbb{R}^{\mathcal{T}}. \tag{5.4.2}$$

5.4.2 Compactness properties

In this subsection, we take advantage of Proposition 5.3.3 and of the non-degeneracy assumption (5.1.8) to get enough compactness for the convergence.

Lemma 5.4.1. *There exists C depending only on Ω and $\min_{i\neq j} a_{i,j}$ such that*

$$\sum_{i=1}^N \iint_{Q_T} |\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \sqrt{\boldsymbol{u}_{i,m}}|^2 + \left(\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \sqrt{\boldsymbol{u}_{i,m}}\right)^2 \mathrm{d}x\mathrm{d}t \le C, \quad \forall m \ge 1.$$

Proof. We get rid of the subscript m for the ease of reading. The L^∞ bound on $\boldsymbol{U}$ yields immediately the L^2 estimate on $\pi_{\mathcal{T},\boldsymbol{\Delta t}} \sqrt{\boldsymbol{u}_i}$. The proof thus consists in proving the bound on the discrete gradient. Le us focus on the proof of

$\iint_{Q_T} |\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\sqrt{\boldsymbol{u}_i}|^2 \mathrm{d}x\mathrm{d}t \leq C$ for some fixed $i \in [\![1,N]\!]$. Thanks to (5.4.2), we have

$$\begin{aligned}\iint_{Q_T} |\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\sqrt{\boldsymbol{u}_i}|^2 &= d\sum_{n=1}^{N_T}\Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma |D_\sigma \sqrt{\boldsymbol{u}_i^n}|^2,\\ &= d\sum_{n=1}^{N_T}\Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \check{u}_{i\sigma}^n |D_\sigma \log(\boldsymbol{u}_i^n)|^2,\end{aligned}$$

where $\check{u}_{i\sigma}^n = 4\frac{\left(D_\sigma\sqrt{\boldsymbol{u}_i^n}\right)^2}{\left(D_\sigma \log(\boldsymbol{u}_i^n)\right)^2}$. It results from Cauchy-Schwarz inequality that

$$4\left(\sqrt{a}-\sqrt{b}\right)^2 \leq (a-b)(\log(a)-\log(b)), \qquad \forall (a,b)\in(0,+\infty),$$

so that $\check{u}_{i\sigma}^n \leq u_{i\sigma}^n$. Therefore, Proposition 5.3.3 provides:

$$\min_{i\neq j} a_{i,j} \sum_{i=1}^{N} \iint_{Q_T} |\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\sqrt{\boldsymbol{u}_i}|^2 \leq \frac{d}{4}\left(E_{\mathcal{T}}(\boldsymbol{U}^0) - E_{\mathcal{T}}(\boldsymbol{U}^{N_T})\right).$$

As $E_{\mathcal{T}}$ is bounded between $-m_\Omega$ and 0 and as, by hypothesis, $\min a_{i,j} > 0$, we obtain the desired bound. □

The inequality $2D_\sigma\sqrt{\boldsymbol{u}_i^n} \geq D_\sigma \boldsymbol{u}_i^n$ and Lemma 5.4.1 yield the following discrete $L^2(0,T;H^1(\Omega))$ estimate on $\boldsymbol{u}_i$.

Corrolary 5.4.2. *There exists C depending only on Ω and $\min_{i\neq j} a_{i,j}$ such that*

$$\sum_{i=1}^{N} \iint_{Q_T} |\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{u}_{i,m}|^2 + \left(\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{u}_{i,m}\right)^2 \mathrm{d}x\mathrm{d}t \leq C, \qquad \forall m \geq 1.$$

The following proposition is about the relative compactness of the approximate solution and of the weakly consistent approximate gradient.

Proposition 5.4.1. *Let $(\boldsymbol{U}_m)$ be the family of discrete solutions. There exists at least one $U \in L^\infty(Q_T;\mathcal{A}) \cap L^2((0,T);H^1(\Omega))$ such that, up to a subsequence, for all $i \in [\![1,N]\!]$:*

$$\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{u}_{i,m} \underset{m\to\infty}{\longrightarrow} u_i \quad \textit{strongly in } L^2(Q_T), \tag{5.4.3}$$

$$\nabla_{\mathcal{T}_m,\boldsymbol{\Delta t}_m}\boldsymbol{u}_{i,m} \underset{m\to\infty}{\longrightarrow} \nabla u_i \quad \textit{weakly in } L^2(Q_T)^d. \tag{5.4.4}$$

Proof. We drop the subscript m for clarity. The proof of this result relies on a discrete Aubin-Lions lemma [78, Lemma 3.4] on the particular setting of [31,

Lemma 9]. Define the discrete $L^2(0,T;(H^1(\Omega))')$ norm by duality as follows:

$$\|\boldsymbol{v}\|_{-1} = \sup\left\{\int_\Omega \pi_{\mathcal{T}}\boldsymbol{v}\pi_{\mathcal{T}}\boldsymbol{\varphi},\ \|\pi_{\mathcal{T}}\boldsymbol{\varphi}\|_{L^2}^2 + \|\nabla_{\mathcal{T}}\boldsymbol{\varphi}\|_{L^2}^2 = 1\right\},\qquad \forall \boldsymbol{v}\in\mathbb{R}^{\mathcal{T}}.$$

Therefore if $\|\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_i\|_{L^2(Q_T)} \le C$ and $\sum_n \|\boldsymbol{u}_i^n - \boldsymbol{u}_i^{n-1}\|_{-1} \le C$, then, up to a subsequence, $\pi_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_i$ tends towards some u_i in $L^2(Q_T)$, while $\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_i$ converges weakly towards ∇u_i. In particular, $U\in L^2(0,T;H^1(\Omega))^N$.

Corollary 5.4.2 provides the L^2 bound on $\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_i$. For the other inequality, we let $\boldsymbol{\varphi}\in\mathbb{R}^{\mathcal{T}}$, $n\in[\![1,N_T]\!]$ and $i\in[\![1,N]\!]$. It follows from (5.2.4a) that

$$\int_\Omega \pi_{\mathcal{T}}\left(\boldsymbol{u}_i^n - \boldsymbol{u}_i^{n-1}\right)\pi_{\mathcal{T}}\boldsymbol{\varphi} = -\Delta t_n \sum_{K\in\mathcal{T}}\varphi_K \sum_{\sigma\in\mathcal{E}_K} F_{i,K\sigma}^n.$$

Using (5.2.4b), this yields

$$\begin{aligned}\frac{1}{\Delta t_n}\int_\Omega \pi_{\mathcal{T}}\left(\boldsymbol{u}_i^n - \boldsymbol{u}_i^{n-1}\right)\pi_{\mathcal{T}}\boldsymbol{\varphi} &= \sum_{\sigma\in\mathcal{E}} a^\star \tau_\sigma D_{K\sigma}\boldsymbol{u}_i^n D_{K\sigma}\boldsymbol{\varphi}\\ &\quad + \sum_{\sigma\in\mathcal{E}}\tau_\sigma\left(\sum_{j=1}^N (a_{i,j}-a^\star)\left(u_{j,\sigma}^n D_{K\sigma}u_i^n - u_{i,\sigma}^n D_{K\sigma}u_j^n\right)\right) D_{K\sigma}\boldsymbol{\varphi}.\end{aligned}$$

Using the Cauchy-Schwarz inequality, the L^∞ bound on $\left(u_{i,\sigma}^n\right)_{\sigma\in\mathcal{E},i\in[\![1,N]\!]}$ and (5.4.1) then leads to

$$\begin{aligned}\frac{1}{\Delta t_n}\int_\Omega \pi_{\mathcal{T}}\left(\boldsymbol{u}_i^n - \boldsymbol{u}_i^{n-1}\right)\pi_{\mathcal{T}}\boldsymbol{\varphi} &\le a^\star \|\nabla_{\mathcal{T}}\boldsymbol{u}_i^n\|_{L^2(\Omega)}\|\nabla_{\mathcal{T}}\boldsymbol{\varphi}\|_{L^2(\Omega)}\\ &\quad + \|\nabla_{\mathcal{T}}\boldsymbol{\varphi}\|_{L^2(\Omega)}\sum_{j=1}^N |a_{i,j}-a^\star|\left(\|\nabla_{\mathcal{T}}\boldsymbol{u}_i^n\|_{L^2(\Omega)} + \|\nabla_{\mathcal{T}}\boldsymbol{u}_j^n\|_{L^2(\Omega)}\right).\end{aligned}$$

By definition of the discrete $(H^1(\Omega))'$ norm, we have

$$\left\|\frac{\boldsymbol{u}_i^n - \boldsymbol{u}_i^{n-1}}{\Delta t_n}\right\|_{-1} \le a^\star\|\nabla_{\mathcal{T}}\boldsymbol{u}_i^n\|_{L^2(\Omega)} + \sum_{j=1}^N |a_{i,j}-a^\star|\left(\|\nabla_{\mathcal{T}}\boldsymbol{u}_i^n\|_{L^2(\Omega)} + \|\nabla_{\mathcal{T}}\boldsymbol{u}_j^n\|_{L^2(\Omega)}\right).$$

Using Corollary 5.4.2 again provides that $\sum_n \|\boldsymbol{u}_i^n - \boldsymbol{u}_i^{n-1}\|_{-1} \le C$. The relative compactness properties on $\pi_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_i$ and $\nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_i$ follow.

We still have to prove that U is in $L^\infty(Q_T;\mathcal{A})$. Let $i\in[\![1,N]\!]$ and let $\varphi_i\in L^2(Q_T)$ be zero where the limit u_i is nonnegative and 1 where the limit is negative,

then

$$\int_{Q_T} \varphi_i \pi_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{u}_i \geq 0 \qquad \text{and} \qquad \int_{Q_T} \varphi_i \pi_{\mathcal{T},\boldsymbol{\Delta t}} \boldsymbol{u}_i \underset{m\to+\infty}{\longrightarrow} \int_{Q_T} u_i \varphi_i \leq 0.$$

Therefore, $\int_{Q_T} u_i \varphi_i = 0$, so that u_i is nonnegative. Finally, the linearity of the limit yields $\sum_{i=1}^N u_i = 1$. □

Remark 5.4.1. The uniform $L^\infty(Q_T)$ bound on $\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \boldsymbol{U}_m$ together with the strong convergence in $L^2(Q_T)$ yield (5.2.6) thanks to Hölder's inequality:

$$\pi_{\mathcal{T}_m,\boldsymbol{\Delta t}_m} \boldsymbol{U}_m \underset{m\to\infty}{\longrightarrow} U \quad \text{strongly in } L^p(Q_T)^N, \text{ for any } 1 \leq p < \infty.$$

We also need convergence properties for the face values $u_{i,\sigma}$. We can reconstruct an approximate solution $u_{i,\mathcal{E},\boldsymbol{\Delta t}}$ which is piecewise constant on the diamond cells by setting, for all $i \in [\![1, N]\!]$:

$$u_{i,\mathcal{E},\boldsymbol{\Delta t}}(t,x) = u_{i,\sigma}^n \quad \text{if} \quad (t,x) \in (t_{n-1}, t_n] \times \Delta_\sigma, \quad \sigma \in \mathcal{E}.$$

Lemma 5.4.3. *We have, for any $i \in [\![1, N]\!]$:*

$$u_{i,\mathcal{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} u_i \quad \text{in} \quad L^p(Q_T), \text{ for any } 1 \leq p < \infty,$$

where U is as in Proposition 5.4.1.

Proof. Here again, we get rid of m for clarity, and show the convergence for a specific value of p. The convergence for any finite p follows from the $L^\infty(Q_T)$ bound on $u_{i,\mathcal{E}_m,\boldsymbol{\Delta t}_m}$ and Hölder's inequality. Since $u_{i,\mathcal{T},\boldsymbol{\Delta t}}$ converges towards u_i in $L^1(Q_T)$, and since $u_{i,\mathcal{E},\boldsymbol{\Delta t}}$ is uniformly bounded, it suffices to show that $\|u_{i,\mathcal{E},\boldsymbol{\Delta t}} - u_{i,\mathcal{T},\boldsymbol{\Delta t}}\|_{L^1(Q_T)}$ tends to 0. Denote by $\Delta_{K\sigma}$ the half-diamond cell which is defined as the interior of the convex hull of $\{x_K, \sigma\}$ for $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, then the following geometrical relation holds:

$$m(\Delta_{K\sigma}) = \frac{1}{d} m_\sigma \mathrm{dist}(x_K, \sigma) \leq \frac{h_\mathcal{T}}{d} m_\sigma.$$

As a consequence,

$$\begin{aligned}\|u_{i,\mathcal{E},\boldsymbol{\Delta t}} - u_{i,\mathcal{T},\boldsymbol{\Delta t}}\|_{L^1(Q_T)} &= \sum_{n=1}^{N_T} \Delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_{\Delta_{K\sigma}} |u_{i,K}^n - u_{i,\sigma}^n| \\ &\leq \frac{h_\mathcal{T}}{d} \sum_{n=1}^{N_T} \Delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_\sigma |u_{i,K}^n - u_{i,\sigma}^n|.\end{aligned}$$

As we have $u_{i,K}^n = u_{i,\sigma}^n$, the contributions corresponding to the boundary edges vanish. For $\sigma \in \mathcal{E}_{\rm int}$, $u_{i,\sigma}$ is an average of $u_{i,K}$ and $u_{i,K\sigma}$, hence $|u_{i,K}^n - u_{i,\sigma}^n| \leq |u_{i,K}^n - u_{i,K\sigma}^n|$. Therefore, we obtain that

$$\begin{aligned}\|u_{i,\mathcal{E},\boldsymbol{\Delta t}} - u_{i,\mathcal{T},\boldsymbol{\Delta t}}\|_{L^1(Q_T)} \leq & \frac{h_\mathcal{T}}{d} \sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} 2m_\sigma \left|D_\sigma \boldsymbol{u}_i^n\right| \\ \leq & 2\frac{h_\mathcal{T}}{d} \left(\sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} m_\sigma d_\sigma\right)^{\frac12} \left(\sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \left|D_\sigma \boldsymbol{u}_i^n\right|^2\right)^{\frac12}.\end{aligned}$$

We deduce from Corollary 5.4.2 that $\|u_{i,\mathcal{E},\boldsymbol{\Delta t}} - u_{i,\mathcal{T},\boldsymbol{\Delta t}}\|_{L^1(Q_T)} \leq Ch_\mathcal{T}$, hence $u_{i,\mathcal{E},\boldsymbol{\Delta t}}$ and $u_{i,\mathcal{T},\boldsymbol{\Delta t}}$ share the same limit in $L^1(Q_T)$. □

5.4.3 Convergence towards a weak solution

The last step to conclude the proof of Theorem 5.2.2 is to identify the limit value U exhibited in Proposition 5.4.1 as a weak solution to (5.1.1), (5.1.3) corresponding to the initial profile $U \in L^\infty(\Omega;\mathcal{A})$. This is the purpose of our last statement.

Proposition 5.4.2. *Let U be as in Proposition 5.4.1, then U is a weak solution in the sense of Definition 8.*

Proof. We drop again the subscript m for the sake of readability, and let $i \in [\![1, N]\!]$, $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$, then define $\boldsymbol{\varphi} = (\varphi_K^n)$ by $\varphi_K^n = \varphi(x_K, t_n)$ for all $n \in \{0,\dots,N_T\}$ and $K \in \mathcal{T}$. Multiplying (5.2.4a) by $\Delta t_n \varphi_K^{n-1}$, then summing over $K \in \mathcal{T}$ and $n \in \{1,\dots,N_T\}$ leads to

$$T_1 + T_2 + T_3 = 0, \tag{5.4.5}$$

where we have set

$$\begin{aligned}T_1 &= \sum_{n=1}^{N_T} \sum_{K\in\mathcal{T}} m_K (u_{i,K}^n - u_{i,K}^{n-1}) \varphi_K^{n-1}, \\ T_2 &= \sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma a^\star D_{K\sigma}\boldsymbol{u}_i^n D_{K\sigma}\boldsymbol{\varphi}^{n-1}, \\ T_3 &= \sum_{n=1}^{N_T} \Delta t_n \sum_{\sigma\in\mathcal{E}} \tau_\sigma \sum_{j=1}^{N} (a_{i,j} - a^\star) \left(u_{j\sigma}^n D_{K\sigma}\boldsymbol{u}_i^n - u_{i\sigma}^n D_{K\sigma}\boldsymbol{u}_j^n\right) D_{K\sigma}\boldsymbol{\varphi}^{n-1}.\end{aligned}$$

The term T_1 can be rewritten as

$$T_1 = \sum_{n=1}^{N_T} \Delta t_n \sum_{K\in\mathcal{T}} m_K u_{i,K}^n \frac{\varphi_K^{n-1} - \varphi_K^n}{\Delta t_n} - \sum_{K\in\mathcal{T}} m_K u_{i,K}^0 \varphi_K^0,$$

so that it follows from the convergence of $\pi_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{U}$ towards U and of $\pi_{\mathcal{T}}\boldsymbol{U}^0$ towards U^0 together with the regularity of φ that

$$T_1 \underset{m\to\infty}{\longrightarrow} -\iint_{Q_T} u_i \partial_t \varphi \mathrm{d}x\mathrm{d}t - \int_\Omega u_i^0 \varphi(0,\cdot)\mathrm{d}x. \tag{5.4.6}$$

To treat the term T_2, we introduce a strongly consistent reconstruction of the gradient. Following [63] (see [52] for a practical example), one can reconstruct a second approximate gradient operator $\widehat{\nabla}_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}} \to L^\infty(\Omega)^d$ such that

$$\int_{\Delta_\sigma} \nabla_{\mathcal{T}}\boldsymbol{u} \cdot \widehat{\nabla}_{\mathcal{T}}\boldsymbol{v}\mathrm{d}x = \tau_\sigma D_{K\sigma}\boldsymbol{u} D_{K\sigma}\boldsymbol{v}, \qquad \forall \boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^{\mathcal{T}}, \forall \sigma \in \mathcal{E},$$

and which is strongly consistent, i.e.,

$$\widehat{\nabla}_{\mathcal{T}}\boldsymbol{\varphi}^n \underset{h_{\mathcal{T}}\to 0}{\longrightarrow} \nabla\varphi(\cdot,t_n) \text{ uniformly in } \overline{\Omega}, \quad \forall n \in \{1,\dots,N_T\},$$

thanks to the smoothness of φ. Using this tool, the terms T_2 and T_3, are easy to treat. The first one can be rewritten as:

$$T_2 = a^\star \iint_{Q_T} \nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_i \cdot \widehat{\nabla}_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{\varphi}\mathrm{d}x\mathrm{d}t,$$

so that

$$T_2 \underset{m\to\infty}{\longrightarrow} a^\star \iint_{Q_T} \nabla u_i \cdot \nabla\varphi \mathrm{d}x\mathrm{d}t. \tag{5.4.7}$$

On the other hand, the term T_3 rewrites

$$T_3 = \iint_{Q_T} \sum_{j=1}^{N} (a_{i,j} - a^\star) \left(u_{j,\mathcal{E},\boldsymbol{\Delta t}} \nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_i - u_{i,\mathcal{E},\boldsymbol{\Delta t}} \nabla_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{u}_j\right) \widehat{\nabla}_{\mathcal{T},\boldsymbol{\Delta t}}\boldsymbol{\varphi},$$

so that

$$T_3 \underset{m\to\infty}{\longrightarrow} \iint_{Q_T} \sum_{j=1}^{N} (a_{i,j} - a^\star) \left(u_j \nabla u_i - u_i \nabla u_j\right) \nabla\varphi. \tag{5.4.8}$$

Combining (5.4.5), (5.4.6), (5.4.7), and (5.4.8), we obtain that

$$-\iint_{Q_T} u_i \partial_t \varphi \mathrm{d}x\mathrm{d}t - \int_\Omega u_i^0 \varphi(0,\cdot)\mathrm{d}x + a^\star \iint_{Q_T} \nabla u_i \cdot \nabla \varphi \mathrm{d}x\mathrm{d}t$$
$$+ \iint_{Q_T} \sum_{j=1}^N (a_{i,j} - a^\star)\left(u_j \nabla u_i - u_i \nabla u_j\right) \nabla \varphi \mathrm{d}x\mathrm{d}t = 0, \quad \forall \varphi \in C_c^\infty([0,T) \times \overline{\Omega}).$$

Using $U \in \mathcal{A}$ and the relation (5.1.10), we recover the weak formulation (5.1.9). □

5.5 Numerical results

The numerical scheme has been implemented using MATLAB. The nonlinear system corresponding to the scheme is solved thanks to a variation of the Newton method with stopping criterion $\|\boldsymbol{U}^{n,k+1} - \boldsymbol{U}^{n,k}\|_\infty \leq 10^{-12}$ for successful convergence or 20 iteration for failed convergence. The solution of the Newton iteration, $\boldsymbol{U}^{n,k+1/3}$, is then "projected" on $\mathcal{A}$ by setting $\boldsymbol{U}^{n,k+2/3} = \max(\boldsymbol{U}^{n,k+1/3}, 10^{-10}\tau)$, and then for all $K \in \mathcal{T}$: $U_K^{n,k+1} = U_K^{n,k+2/3}/(\sum_{i=1}^N u_{i,K}^{n,k+2/3})$. The seed of the algorithm is the solution to N uncoupled heat equations with diffusion coefficients all equal to $a^\star$.

For the first time step, we also make use of a continuation method based on the intermediate diffusion coefficients $a_{i,j}^\lambda = \lambda a_{ij} + (1-\lambda)a^\star$ with $\lambda \in [0,1]$. The parameter λ is originally set to 1. If Newton's method does not converge, we let $\lambda = (\lambda + \lambda_{\text{prev}})/2$ where λ_{prev} is originally set to 0. If Newton's method converges, we let $\lambda_{\text{prev}} = \lambda$ and $\lambda = 1$.

5.5.1 Convergence under grid refinement

Our first test case is devoted to the convergence analysis of the scheme in a one-dimensional setting $\Omega = (0,1)$. Two different initial conditions are considered: U_s^0 is smooth and vanished point-wise at the boundary of Ω, whereas U_r^0 is discontinuous and vanishes on intervals of Ω:

$$u_{1,s}^0(x) = \frac{1}{4} + \frac{1}{4}\cos(\pi x), \quad u_{2,s}^0(x) = \frac{1}{4} + \frac{1}{4}\cos(\pi x), \quad u_{3,s}^0(x) = \frac{1}{2} - \frac{1}{2}\cos(\pi x),$$
$$u_{1,r}^0 = 1_{[\frac{3}{8},\frac{5}{8}]}, \qquad u_{2,r}^0 = 1_{(\frac{1}{8},\frac{3}{8})} + 1_{(\frac{5}{8},\frac{7}{8})}, \qquad u_{3,r}^0 = 1_{[0,\frac{1}{8}]} + 1_{[\frac{7}{8},1]}.$$

We also consider three cross-diffusion coefficients matrices, a first one A^{Lap} corresponding to 3 uncoupled heat equation, a second one A^{reg} called regular with positive off-diagonal coefficients, and a third one A^{sing} called singular with a few

null off-diagonal coefficients:

$$A^{\mathrm{Lap}} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \qquad A^{\mathrm{reg}} = \begin{pmatrix} 0 & 0.2 & 1 \\ 0.2 & 0 & 0.1 \\ 1 & 0.1 & 0 \end{pmatrix}, \qquad A^{\mathrm{sing}} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0.1 \\ 1 & 0.1 & 0 \end{pmatrix}.$$

For the convergence tests, we have let $a^\star = 0.1$ and the meshes are uniform discretisations of $[0,1]$ from 2^5 cells to 2^{15} cells. The approximate solutions are compared to a reference solution which is analytical when $A = A^{\mathrm{Lap}}$ and computed on the finest grid (2^{15} cells) when $A = A^{\mathrm{reg}}$ or $A = A^{\mathrm{sing}}$. The final time is 0.25, and the time discretisation is fixed with a time step of $\Delta t = 2^{-18}$. In the case $A = A^{\mathrm{Lap}}$, we also report in Figure 5.2 results with 128 times finer time step, i.e. $\Delta t = 2^{-25}$. One notices that our scheme is second-order accurate in space in the setting presented in this paper ($A = A^{\mathrm{reg}}$), but only first-order accurate when confronted to non-diffusive discontinuities. We call non-diffusive discontinuities a spatial discontinuity of u_1^0 and u_2^0 (recall that $a_{1,2} = 0$ in A^{sing}) for which u_3^0 is equal to 0 on both sides of the discontinuity, so that the contributions corresponding to $a_{1,3}$ and $a_{2,3}$ vanish at $t = 0$. The origin of this lower order might lie in the difficulty to compute accurately the near-zero concentrations in the neighborhood of such discontinuities. We also notice in Figure 5.2 the prevalence of the error in time when comparing with respect to an analytical solution, which will motivate the development of higher-order in time methods already discussed in Remark 5.2.2.

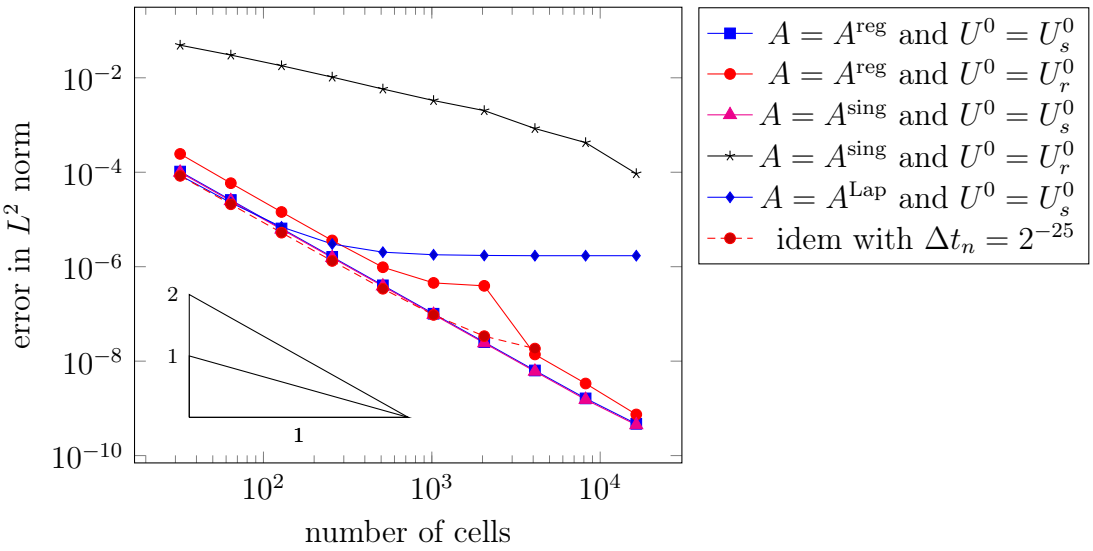


Figure 5.2 – Error with respect to reference solution.

5.5.2 On the influence of the parameter $a^\star$

The choice of $a^\star$ is a natural question concerning our scheme. The equation (5.2.3) gives a lower bound: $a^\star > 0$. The existence of an upper bound is not as clear. Equation (5.3.7) shows that for large $a^\star$, we over-estimate the diffusion. The optimal value of $a^\star$ depends on many variables such as the initial condition, the final time, and the mesh. Optimal choices of $a^\star$ are reported in Table 5.1. Notice that the optimal value is test cases dependent, since it is affected by the initial condition and by the final time.

			$A = A^{\text{reg}}$		$A = A^{\text{sing}}$	
			$U^0 = U_s^0$	$U^0 = U_r^0$	$U^0 = U_s^0$	$U^0 = U_r^0$
nb. of cells	32	$T = 0.125$	0.86	0.21	0.79	0.0023
		$T = 0.25$	0.67	0.13	0.49	0.00082
	128	$T = 0.125$	0.86	0.17	0.79	0.00050
		$T = 0.25$	0.67	0.11	0.49	0.00049

Table 5.1 – Values of $a^\star_{\text{opt}}$ for different parameters. $a^\star_{\text{opt}}$ is computed with respect to the reference solution of Section 5.5.1 for the L^2 norm.

One notices on Fig. 5.3 that the dependency of the quality of the results is strong for the initial data U_r^0. This is due to the presence of vanishing concentrations in some cells, so that the choice $a^\star = 0$ would allow for spurious solutions as highlighted in Remark 5.2.1. In this situation, the choice of $a^\star$ strongly affects the quality of the results, especially for the first time steps where some concentrations are still close to 0. The numerical experiment and homogeneity considerations suggest the following suboptimal rule for choosing $a^\star$:

$$a^\star = \min\left\{ \max_{i\neq j} a_{i,j} \ ; \ \max\left\{ \min_{i\neq j} a_{i,j}, \ \epsilon \frac{h_{\mathcal{T}}^2}{\tau} \right\} \right\}, \tag{5.5.1}$$

where $h_{\mathcal{T}}$ is the mesh size, τ the current time step and ϵ a small parameter to be tuned by the user. Another interesting feature of Figure 5.3 is the behavior of the curve corresponding to $A = A^{\text{Lap}}$. The classical TPFA scheme for the heat equation corresponding to $a^\star = 1$ is outperformed in terms of accuracy by the scheme corresponding to higher value of $a^\star \simeq 9$. The introduction of enhanced diffusion by picking high values of $a^\star$ is not covered by formula (5.5.1).

5.5.3 A 2D test case with reaction

Our second test is two-dimensional. We choose A^{sing} as the diffusion matrix and $a^\star = 0.1$. The domain $\Omega = (0, 22) \times (0, 16)$ is discretized into a cartesian grid

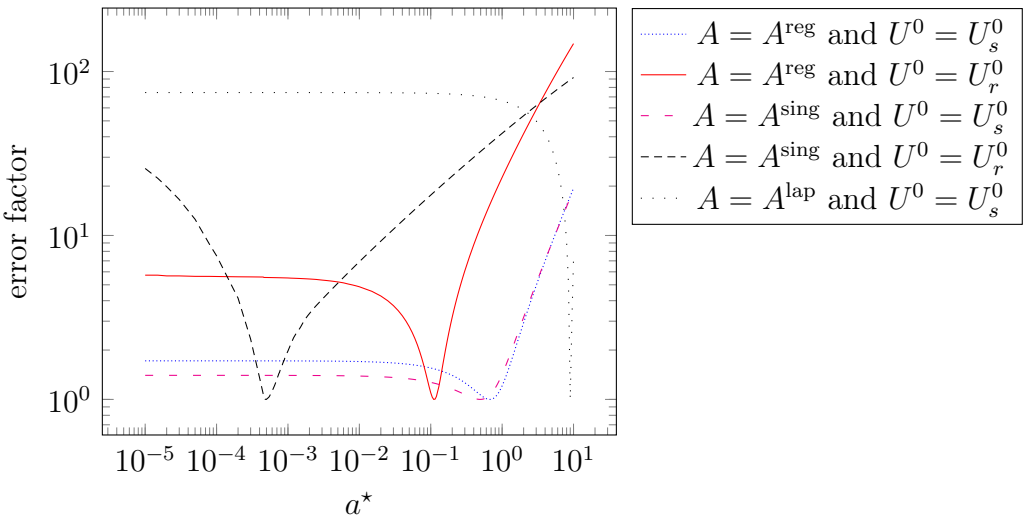


Figure 5.3 – Evolution of the ratio $\frac{\|U_{a^\star} - U_{\mathrm{ref}}\|_2}{\|U_{a^\star_{\mathrm{opt}}} - U_{\mathrm{ref}}\|_2}$, where $U_{a^\star}$ is computed with 2^7 cells and U_{ref} is as in Section 5.5.1.

made of 110×80 cells which is presented along with initial condition in Figure 5.4. We use a uniform time stepping with $\tau = 2^{-3}$. To illustrate Remark 5.1.1, we introduce the following reaction:

$$e_1 + e_3 \underset{1}{\overset{1000}{\rightleftharpoons}} 2e_2,$$

which translates as follows in the source term $R = (r_1, r_2, r_3)^T$:

$$r_1(U) = (u_2^+)^2 - 1000 u_1^+ u_3^+ \qquad r_2(U) = -2r_1(U) \qquad r_3(U) = r_1(U).$$

The reaction term R obviously satisfies Assumptions ((i)) and ((ii)) of Remark 5.1.1. Let us now discuss Assumption ((iii)). We have for all $U, \overline{U} \in \mathcal{A}$:

$$R(U).\log(U/\overline{U}) = r_1(U)\Big(\big[\log(u_1 u_3) - \log(u_2^2)\big] - \big[\log(\overline{u_1 u_3}) - \log(\overline{u_2}^2)\big]\Big).$$

The above expression is then nonpositive for any $\overline{U}$ such that $\log(\overline{u_1 u_3}) - \log(\overline{u_2}^2) = -\log(1000)$, or equivalently $R(\overline{U}) = 0$. A particular choice for such a $\overline{U}$ is the steady state U^∞, which is constant w.r.t. to space and determined as follows. Denote by α the average advancement of the reaction, then

$$u_1^\infty(x) = \frac{9}{44} - \alpha, \quad u_2^\infty(x) = \frac{2}{11} + 2\alpha, \quad u_3^\infty(x) = \frac{27}{44} - \alpha.$$

For $\alpha = 0$, the fraction of each specie is the ratio of corresponding occupied area in Figure 5.4, i.e. $\frac{1}{m_\Omega}\int_\Omega U^0 = (\frac{9}{44}, \frac{2}{11}, \frac{27}{44})^T$. The value of α is determined by imposing that $R(U^\infty) = 0$, which amounts to find a root of a polynomial of degree two. Among the two roots, only the choice $\alpha = \frac{-5\sqrt{206530}+4504}{10956}$ yields a non-negative U^∞. The time evolution of the relative energy $E_{\mathcal{T}}(\boldsymbol{U}|\boldsymbol{U}^\infty)$ is plotted on Figure 5.5, showing exponential decay to the steady-state even though the diffusion matrix is singular. Snapshots showing the evolution of the concentration profiles are presented in Figure 5.6.

To compute the solutions to our numerical scheme, we have to adapt the continuation procedure sketched at the beginning of Section 5.5 to include source terms. Roughly speaking, we solve the discrete counterpart to

$$\partial_t u_i - a^\star \Delta u_i - \lambda \operatorname{div}\left(\sum_{j=1}^{N}(a_{i,j} - a^\star)\big(u_j \nabla u_i - u_i \nabla u_j\big)\right) = \mu r_i(U),$$

where the source terms are discretized in a fully implicit way. Due to the stiffness of the reaction terms, we have to treat the reaction first then the cross-diffusion effects. More precisely, given $\boldsymbol{U}^{n-1} \in \mathcal{A}^{\mathcal{T}}$, we initialise the iterative method for the computation of $\boldsymbol{U}^n$ with $\boldsymbol{U}^{n,0}$ defined as the unique solution to the N uncoupled heat equations corresponding to $\lambda = \mu = 0$. Then one tries to solve the system corresponding to $\lambda = \mu = 1$. If the modified Newton's method with truncation and rescaling sketched at the beginning of Section 5.5 fails to converge, then one sets $\lambda = 0$ and $\mu = \frac{1}{2}$. Then one use a similar continuation method to the one described at the beginning of Section 5.5 to increase μ until it reaches the value 1. Then the continuation method is used again to increase the value of λ until $\lambda = 1$ is reached.

5.6 Conclusion

We proposed a finite volume scheme based on two-point flux approximation for a degenerate cross-diffusion system. The scheme was designed to preserve the key properties of the continuous system, namely the positivity of the solutions, the constraint on the composition and the decay of the entropy. The scheme requires the introduction of a positive parameter $a^\star$ to avoid unphysical solutions. This parameter plays an important role in the convergence proof, which is carried out under a non-degeneracy assumption. Its importance is also confirmed in the numerical experiments, in particular in the presence of initial profiles with concentrations vanishing in some parts of the computational domain.

Acknowledgements
The authors acknowledge support from the Labex CEMPI (ANR-11-LABX-0007-

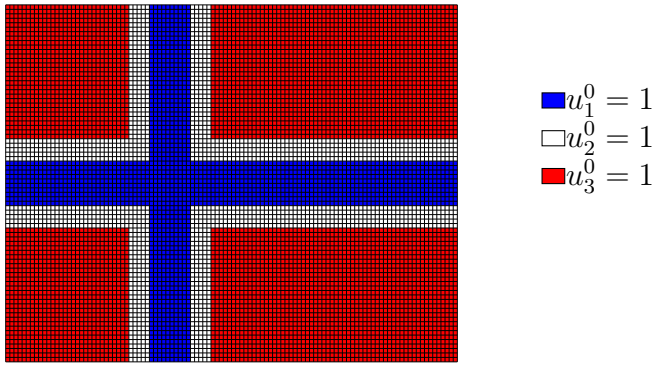


Figure 5.4 – Initial configuration U^0 for the concentrations

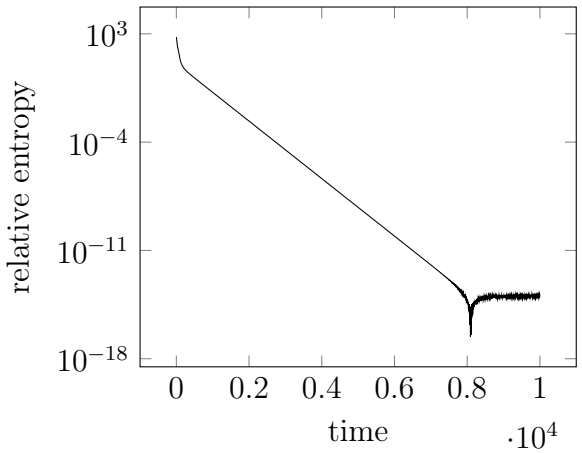
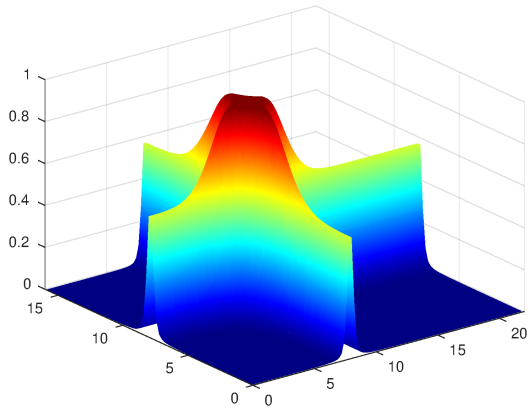
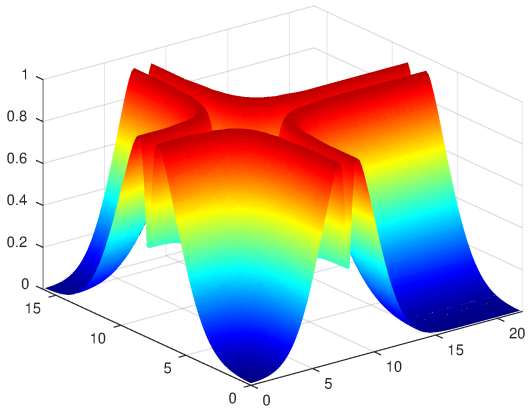


Figure 5.5 – $|E_{\mathcal{T}}(\boldsymbol{U}|\boldsymbol{U}^\infty)|$ as a function of time.

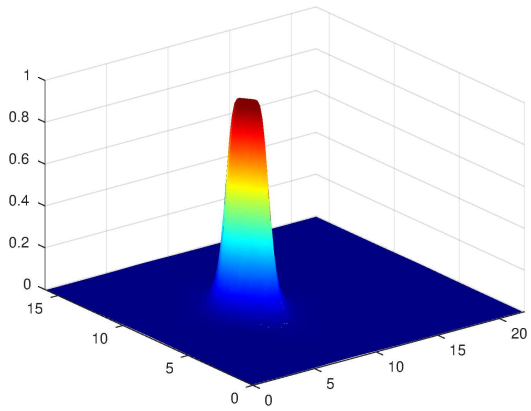
01). Clément Cancès also acknowledges support from the COMODO project (ANR-19-CE46-0002), and he warmly thanks Virginie Ehrlacher and Laurent Monasse for stimulating discussions that were at the origin of this work.



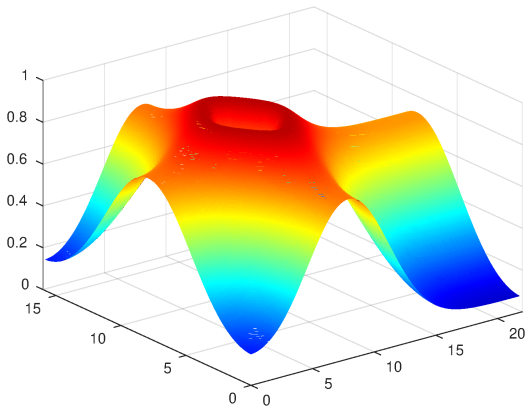
Profile of u_1 at time $t = 20$



Profile of u_2 at time $t = 20$



Profile of u_1 at time $t = 50$



Profile of u_2 at time $t = 50$

Figure 5.6 – Concentration configurations for various times. The concentration of the third specie can be deduced thanks to $u_1 + u_2 + u_3 = 1$

Appendix A

About the inverse activity coefficients

In [70] and [71], the system derived in the general introduction introduction (1.1.10) is studied and simulated using activity coefficients $a_i = \exp h_i$ and $\beta_i = \frac{c_i}{a_i}$. The ability to compute β knowing a but not c allows for another formulation of the PDE system:

$$\partial_t \beta_i(a) a_i - \operatorname{div} D_i \beta_i(a)(\nabla a_i + a_i \nabla \Phi) = 0.$$

Moreover, we have seen in Chapter 2 that the activity based scheme did not preserve the concentration averaging property (see Lemma 2.3.1). The ability to compute β as a function of a could solve this problem by setting: $\beta_{K|L} = \beta(a_{K|L})$ instead of $\frac{\beta_K + \beta_L}{2}$. A first step in this direction has been made in [70, Appendix D] in the Bikermann setting (i.e. same molar masses and molar volume for each species), and a general result has been conjectured. In this appendix, we propose a mostly constructive solution to the following non linear system. Given $(k_i)_{1\leq i\leq N}, (v_i)_{0\leq i\leq N}$ non negative parameters, $(f_i)_{1\leq i\leq N}$ positive parameters describing different characteristics of the species, $a = (a_i)_{1\leq i\leq N}, p$ describing the state of the system, is there a unique $\beta = (\beta_i)_{1\leq i\leq N}$ such that:

$$\begin{aligned}
&h_i = \log(\frac{c_i}{\overline{c}}) - k_i \log(\frac{c_0}{\overline{c}}) + v_i p && \forall i \in [\![1, N]\!], \\
&a_i = \exp h_i, \qquad a_i \beta_i = c_i, && \forall i \in [\![1, N]\!], \qquad \text{(S)} \\
&c_0 = \frac{1}{v_0} - \sum_{i=1}^{N} f_i c_i, \qquad \overline{c} = \sum_{i=0}^{N} c_i, \qquad 0 \leq c_0 < \overline{c}.
\end{aligned}$$

By definition of a we have:

$$a_i = \frac{c_i}{\overline{c}} \left(\frac{c_0}{\overline{c}}\right)^{-k_i} \exp v_i p.$$

Thus, we introduce $x = \frac{c_0}{\overline{c}}$, $\alpha_i = \exp(v_i p)$. We first show that there exist a unique possible value for x, and then we compute the full solution of (S) using this molar fraction. Notice that thanks to the bounds on c_0, we have $x \in [0, 1)$. The system (S) yields:

$$\beta_i = \frac{\overline{c}}{\alpha_i} x^{k_i}, \qquad \forall i \in [\![1, N]\!], \tag{A.0.1}$$

$$c_0 = \frac{1}{v_0} - \sum_{i=1}^{N} f_i a_i \beta_i, \tag{A.0.2}$$

$$\overline{c} = c_0 + \sum_{i=1}^{N} a_i \beta_i, \tag{A.0.3}$$

$$0 \leq c_0 < \overline{c}. \tag{A.0.4}$$

Multiplying (A.0.1) by a and summing over the species leads to:

$$\sum_{i=1}^{N} a_i \beta_i = \overline{c} \sum_{i=1}^{N} \frac{a_i}{\alpha_i} x^{k_i}.$$

Noticing that the left-hand side also appears in (A.0.3), we get:

$$\overline{c} - c_0 = \overline{c} \sum_{i=1}^{N} \frac{a_i}{\alpha_i} x^{k_i}.$$

Dividing by $\overline{c}$ then provides:

$$\sum_{i=1}^{N} \frac{a_i}{\alpha_i} x^{k_i} = 1 - x. \tag{A.0.5}$$

By monotony, we dispose of a unique $x \in (0, 1)$ satisfying (A.0.5) [1]. We still have one degree of freedom in the choice of c_0 and $\overline{c}$. Notice that thanks to (A.0.2) c_0 must satisfy:

$$c_0 = \frac{1}{v_0} - \overline{c} \sum_{i=1}^{N} \frac{f_i a_i}{\alpha_i} x^{k_i} = \frac{1}{v_0} - c_0 \sum_{i=1}^{N} \frac{f_i a_i}{\alpha_i} x^{k_i - 1},$$

hence there exists at most one solution to (S). Reciprocally, one can check that the solution we have constructed satisfies (S).

As the resolution of (A.0.5) is not explicit, our solution is not fully constructive.

1. In the Bikermann setting, (A.0.5) is a linear equation.

However, under the assuption that the solvated ions are bigger than the solvent molecules we have $k_i \geq 1$. This physically sound assumption yields the convexity of the function considered, therefore we have a convergence result of Newton's method.

Bibliography

[1] D. Abdel, P. Vágner, J. Fuhrmann, and P. Farrell. Modelling charge transport in perovskite solar cells: Potential-based and limiting ion depletion. 2020. Medium: PDF Publisher: Weierstrass Institute.

[2] A. Ait Hammou Oulhaj. Numerical analysis of a finite volume scheme for a seawater intrusion model with cross-diffusion in an unconfined aquifer. Numer. Methods Partial Differential Equations, 34(3):857–880, 2018.

[3] A. Ait Hammou Oulhaj, C. Cancès, and C. Chainais-Hillairet. Numerical analysis of a nonlinearly stable and positive Control Volume Finite Element scheme for Richards equation with anisotropy. ESAIM: Mathematical Modelling and Numerical Analysis, 52(4):1532–1567, 2018.

[4] A. Ait Hammou Oulhaj and D. Maltese. Convergence of a positive nonlinear control volume finite element scheme for an anisotropic seawater intrusion model with sharp interfaces. Numer. Methods Partial Differential Equations, 36(1):133–153, 2019.

[5] J. M. Aldaz. A stability version of Hölder's inequality. Journal of Mathematical Analysis and Applications, 343(2):842–852, July 2008.

[6] J. M. Aldaz. Strengthened Cauchy-Schwarz and Hölder inequalities. JIPAM. Journal of Inequalities in Pure & Applied Mathematics [electronic only], 10(4):Paper No. 116, 6 p., 2009.

[7] L. Almeida, F. Bubba, and C. Perthame, B. Pouchol. Energy and implicit discretization of the Fokker-Planck and Keller-Segel type equations. Netw. Heterog. Media, 14(1):23–41, 2019.

[8] B. Andreianov, M. Bendahmane, and R. Ruiz-Baier. Analysis of a finite volume method for a cross-diffusion model in population dynamics. Math. Models Methods Appl. Sci., 21(2):307–344, 2011.

[9] B. Andreianov, C. Cancès, and A. Moussa. A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic PDEs. J. Funct. Anal., 273(12):3633–3670, 2017.

[10] R. Bailo, J. A. Carrillo, H. Murakawa, and M. Schmidtchen. Convergence of a fully discrete and energy-dissipating finite-volume scheme for aggregation-diffusion equations. arXiv:2002.10821, 2020.

[11] A. Bakhta and V. Ehrlacher. Cross-diffusion systems with non-zero flux and moving boundary conditions. ESAIM: Mathematical Modelling and Numerical Analysis, 52(4):1385–1415, July 2018.

[12] M. Z. Bazant, M. S. Kilic, B. D. Storey, and A. Ajdari. Towards an understanding of induced-charge electrokinetics at large applied voltages in concentrated solutions. Advances in colloid and interface science, 152(1-2):48–88, 2009.

[13] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math., 84(3):375–393, 2000.

[14] J. Berendsen, M. Burger, V. Ehrlacher, and J.-F. Pietschmann. Uniqueness of strong solutions and weak–strong stability in a system of cross-diffusion equations. Journal of Evolution Equations, 20(2):459–483, June 2020.

[15] M. Bessemoulin-Chatard. A finite volume scheme for convection–diffusion equations with nonlinear diffusion derived from the Scharfetter–Gummel scheme. Numerische Mathematik, 121(4):637–670, Aug. 2012.

[16] M. Bessemoulin-Chatard, C. Chainais-Hillairet, and F. Filbet. On discrete functional inequalities for some finite volume schemes. IMA J. Numer. Anal., 35:1125–1149, 2015.

[17] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. SIAM review, 59(1):65–98, 2017.

[18] J. Blakemore. Approximations for fermi-dirac integrals, especially the function f12 (η) used to describe electron density in a semiconductor. Solid-State Electronics, 25(11):1067–1076, 1982.

[19] A. Bondesan and M. Briant. Stability of the Maxwell–Stefan System in the Diffusion Asymptotics of the Boltzmann Multi-species Equation. Communications in Mathematical Physics, 382(1):381–440, Feb. 2021.

[20] J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu. The gradient flow structures of thermo-poro-visco-elastic processes in porous media. arXiv:1907.03134, 2019.

[21] D. Bothe. On the Maxwell-Stefan Approach to Multicomponent Diffusion. In J. Escher, P. Guidotti, M. Hieber, P. Mucha, J. W. Prüss, Y. Shibata, G. Simonett, C. Walker, and W. Zajaczkowski, editors, Parabolic Problems: The Herbert Amann Festschrift, Progress in Nonlinear Differential Equations and Their Applications, pages 81–93. Springer, Basel, 2011.

[22] D. Bothe and P.-E. Druet. On the structure of continuum thermodynamical diffusion fluxes – A novel closure scheme and its relation to the Maxwell-Stefan and the Fick-Onsager approach. Aug. 2020. arXiv: 2008.05327.

[23] K. Brenner, C. Cancès, and D. Hilhorst. Finite volume approximation for an immiscible two-phase flow in porous media with discontinuous capillary pressure. Computational Geosciences, 17(3):573–597, June 2013.

[24] K. Brenner, J. Droniou, R. Masson, and E. H. H. Quenjel. Total-velocity-based finite volume discretization of two-phase Darcy flow in highly heterogeneous media with discontinuous capillary pressure. Feb. 2021.

[25] F. Brochard, J. Jouffroy, and P. Levinson. Polymer-polymer diffusion in melts. Macromolecules, 16(10):1638–1641, 1983.

[26] M. Burger, M. Di Francesco, J.-F. Pietschmann, and B. Schlake. Nonlinear Cross-Diffusion with Size Exclusion. SIAM Journal on Mathematical Analysis, 42(6):2842–2871, Jan. 2010.

[27] M. Burger, B. Schlake, and M.-T. Wolfram. Nonlinear Poisson–Nernst–Planck equations for ion flux through confined geometries. Nonlinearity, 25(4):961–990, Mar. 2012.

[28] C. Cancès. Energy stable numerical methods for porous media flow type problems. Oil & Gas Science and Technology - Revue d'IFP Energies nouvelles, 73:78, 2018.

[29] C. Cancès, C. Chainais-Hillairet, J. Fuhrmann, and B. Gaudeul. A numerical-analysis-focused comparison of several finite volume schemes for a unipolar degenerate drift-diffusion model. IMA Journal of Numerical Analysis, July 2020.

[30] C. Cancès, C. Chainais-Hillairet, J. Fuhrmann, and B. Gaudeul. On four numerical schemes for a unipolar degenerate drift-diffusion model. In R. Klöfkorn, F. Radu, E. Keijgavlen, and J. Fuhrmann, editors, Finite Volumes for Complex Applications IX, Bergen (Norway), June 2020. Springer, 2020.

[31] C. Cancès, C. Chainais-Hillairet, A. Gerstenmayer, and A. Jüngel. Finite-volume scheme for a degenerate cross-diffusion model motivated from ion transport. Numerical Methods for Partial Differential Equations, 35(2):545–575, 2019.

[32] C. Cancès, C. Chainais-Hillairet, M. Herda, and S. Krell. Large time behavior of nonlinear finite volume schemes for convection-diffusion equations. SIAM Journal on Numerical Analysis, Sept. 2020.

[33] C. Cancès, C. Chainais-Hillairet, and S. Krell. Numerical Analysis of a Nonlinear Free-Energy Diminishing Discrete Duality Finite Volume Scheme

for Convection Diffusion Equations. Computational Methods in Applied Mathematics, 18(3):407–432, July 2018. Special issue on "Advanced numerical methods: recent developments, analysis and application".

[34] C. Cancès and B. Gaudeul. A convergent entropy diminishing finite volume scheme for a cross-diffusion system. SIAM Journal on Numerical Analysis, 58:2684–2710, 2020.

[35] C. Cancès and C. Guichard. Convergence of a nonlinear entropy diminishing Control Volume Finite Element scheme for solving anisotropic degenerate parabolic equations. Mathematics of Computation, 85(298):549–580, 2016.

[36] C. Cancès and C. Guichard. Numerical Analysis of a Robust Free Energy Diminishing Finite Volume Scheme for Parabolic Equations with Gradient Structure. Foundations of Computational Mathematics, 17(6):1525–1584, Dec. 2017.

[37] C. Cancès, H. Mathis, and N. Seguin. Error Estimate for Time-Explicit Finite Volume Approximation of Strong Solutions to Systems of Conservation Laws. SIAM J. Numer. Anal., 54(2):1263–1287, 2016.

[38] C. Cancès and F. Nabet. Finite volume approximation of a two-phase two fluxes degenerate cahn-hilliard model. HAL: hal-02561981, 2020.

[39] C. Cancès, F. Nabet, and M. Vohralík. Convergence and a posteriori error analysis for energy-stable finite element approximations of degenerate parabolic equations. Jan. 2019.

[40] C. Cancès, V. Ehrlacher, and L. Monasse. Finite volumes for the Stefan-Maxwell cross-diffusion system. arXiv:2007.09951 [cs, math], July 2020. arXiv: 2007.09951.

[41] C. Cancès, T. O. Gallouët, and L. Monsaingeon. The gradient flow structure for incompressible immiscible two-phase flows in porous media. Comptes Rendus Mathematique, 353(11):985–989, Nov. 2015.

[42] C. Cancès and B. Gaudeul. Entropy diminishing finite volume approximation of a cross-diffusion system. In FVCA9 - International Conference on Finite Volumes for Complex Applications IX, Bergen, Norway, 2020.

[43] C. Cancès, D. Matthes, and F. Nabet. A two-phase two-fluxes degenerate Cahn-Hilliard model as constrained Wasserstein gradient flow. Archive for Rational Mechanics and Analysis, 233(2):837–866, 2019. Publisher: Springer Verlag.

[44] C. Cancès and A. Zurek. A convergent finite volume scheme for dissipation driven models with volume filling constraint. 2021.

[45] J. A. Carrillo, F. Filbet, and M. Schmidtchen. Convergence of a finite volume scheme for a system of interacting species with cross-diffusion. Numerische Mathematik, 145(3):473–511, July 2020.

[46] J. A. Carrillo, D. Matthes, and M.-T. Wolfram. Lagrangian schemes for Wasserstein gradient flows. arXiv:2003.03803, 2020.

[47] C. Chainais-Hillairet. Entropy method and asymptotic behaviours of finite volume schemes. In Finite volumes for complex applications. VII. Methods and theoretical aspects, volume 77 of Springer Proc. Math. Stat., pages 17–35. Springer, Cham, 2014.

[48] C. Chainais-Hillairet, J.-G. Liu, and Y.-J. Peng. Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. ESAIM: Mathematical Modelling and Numerical Analysis, 37(2):319–338, Mar. 2003.

[49] W. Chen, C. Wang, X. Wang, and S. M. Wise. Positivity-preserving, energy stable numerical schemes for the Cahn-Hilliard equation with logarithmic potential. J. Comput. Phys.: X, 3:100031, 2019.

[50] F. H. Clarke. Optimization and nonsmooth analysis, volume 5 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990.

[51] R. Coehoorn, W. Pasveer, P. Bobbert, and M. Michels. Charge-carrier concentration dependence of the hopping mobility in organic materials with gaussian disorder. Physical Review B, 72(15):155206, 2005.

[52] Y. Coudière, J.-P. Vila, and P. Villedieu. Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. ESAIM: Mathematical Modelling and Numerical Analysis, 33(3):493–516, May 1999.

[53] E. S. Daus, A. Jüngel, and Zurek. Convergence of a finite-volume scheme for a degenerate-singular cross-diffusion system for biofilms. arXiv:2001.09544, 2020.

[54] P. G. de Gennes. Dynamics of fluctuations and spinodal decomposition in polymer blends. J. Chem. Phys., 72:4756–4763, 1980.

[55] S. R. De Groot and P. Mazur. Non-equilibrium thermodynamics. North-Holland, Amsterdam, 1962.

[56] K. Deimling. Nonlinear Functional Analysis. Springer-Verlag, Berlin Heidelberg, 1985.

[57] L. Dong, C. Wang, H. Zhang, and Z. Zhang. A positivity-preserving, energy stable and convergent numerical scheme for the Cahn-Hilliard equation with a Flory-Huggins-deGennes energy. Commun. Math. Sci., 17(4):921–939, 2019.

[58] DPHYMS. Co-evaporation. Université du Luxembourg, (https://wwwfr.uni.lu/recherche/fstm/dphyms/research/photovoltaics/research/co_evaporation), may 2021.

[59] W. Dreyer, C. Guhlke, and M. Landstorfer. A mixture theory of electrolytes containing solvation effects. Electrochemistry communications, 43:75–78, 2014.

[60] W. Dreyer, C. Guhlke, and R. Müller. Overcoming the shortcomings of the Nernst–Planck model. Physical Chemistry Chemical Physics, 15(19):7075–7086, Apr. 2013.

[61] J. Droniou. Finite volume schemes for diffusion equations: Introduction to and review of modern methods. Mathematical Models and Methods in Applied Sciences, 24(08):1575–1619, Jan. 2014.

[62] J. Droniou and R. Eymard. Study of the mixed finite volume method for stokes and navier-stokes equations. Numerical Methods for Partial Differential Equations, 25:137–171, 2009.

[63] J. Droniou and R. Eymard. The Asymmetric Gradient Discretisation Method. In C. Cancès and P. Omnes, editors, Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects, volume 199 of Springer Proceedings in Mathematics & Statistics, pages 311–319, Cham, 2017. Springer International Publishing.

[64] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. The Gradient Discretisation Method, volume 42 of Mathématiques et Applications. Springer International Publishing, July 2018.

[65] J. Droniou and N. Nataraj. Improved L^2 estimate for gradient schemes and super-convergence of the TPFA finite volume scheme. IMA J. Numer. Anal., 38(3):1254–1293, 2018.

[66] R. Eymard and T. Gallouët. H-Convergence and Numerical Schemes for Elliptic Problems. SIAM Journal on Numerical Analysis, 41(2):539–562, Jan. 2003.

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

[68] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In Handbook of Numerical Analysis, volume 7 of Solution of Equation in R (Part 3), Techniques of Scientific Computing (Part 3), pages 713–1018. Elsevier, Jan. 2000.

[69] P. Farrell, T. Koprucki, and J. Fuhrmann. Computational and analytical comparison of flux discretizations for the semiconductor device equations beyond Boltzmann statistics. J. Comput. Phys., 346:497–513, 2017.

[70] J. Fuhrmann. Comparison and numerical treatment of generalised Nernst–Planck models. Computer Physics Communications, 196:166–178, 2015.

[71] J. Fuhrmann. A Numerical Strategy for Nernst–Planck Systems with Solvation Effect. Fuel Cells, 16(6):704–714, 12 2016.

[72] J. Fuhrmann. UnipolarDriftDiffusion.jl - Numerical examples for finite volume schemes for unipolar drift-diffusion problems, 2019. DOI:10.5281/zenodo.3351467.

[73] J. Fuhrmann. VoronoiFVM.jl: Solver for coupled nonlinear partial differential equations based on the voronoi finite volume method, 2020. DOI:10.5281/zenodo.3529808.

[74] J. Fuhrmann and C. Guhlke. A finite volume scheme for Nernst-Planck-Poisson systems with ion size and solvation effects. In C. C. . P. Omnes, editor, Finite Volumes for Complex Applications VIII - Hyperbolic, Elliptic and Parabolic Problems, volume 200 of Springer Proceedings in Mathematics & Statistics, pages 497–505, Lille, France, 2017. Springer International Publishing.

[75] H. Gajewski. On the uniqueness of solutions to the drift-diffusion model of semiconductor devices. Math. Models Methods Appl. Sci., 4(1):121–133, 1994.

[76] T. Gallouët. Nonlinear methods for linear equations. In Proceedings of Tamtam'07 conference held in Tipaza, 2007.

[77] T. Gallouët, R. Herbin, D. Maltese, and A. Novotný. Error estimates for a numerical approximation to the compressible barotropic Navier-Stokes equations. IMA J. Numer. Anal., 36(2):543–592, 2016.

[78] T. Gallouët and J.-C. Latché. Compactness of discrete approximate solutions to parabolic PDEs - Application to a turbulence model. Communications on Pure & Applied Analysis, 11(6):2371–2391, 2012.

[79] T. Gallouët, D. Maltese, and A. Novotný. Error estimates for the implicit MAC scheme for the compressible Navier-Stokes equations. Numer. Math., 141(2):495–567, 2019.

[80] K. Gärtner and L. Kamenski. Why Do We Need Voronoi Cells and Delaunay Meshes? In V. A. Garanzha, L. Kamenski, and H. Si, editors, Numerical Geometry, Grid Generation and Scientific Computing, Lecture Notes in Computational Science and Engineering, pages 45–60, Cham, 2019. Springer International Publishing.

[81] B. Gaudeul and J. Fuhrmann. Entropy and convergence analysis for two finite volume schemes for a Nernst-Planck-Poisson system with ion volume constraints. Feb. 2021.

[82] N. Gavish and A. Yochelis. Theory of phase separation and polarization for pure ionic liquids. The journal of physical chemistry letters, 7(7):1121–1126, 2016.

[83] A. Gerstenmayer and A. Jüngel. Analysis of a degenerate parabolic cross-diffusion system for ion transport. J. Math. Anal. Appl., 461(1):523–543, 2018.

[84] A. Gerstenmayer and A. Jüngel. Comparison of a finite-element and finite-volume scheme for a degenerate cross-diffusion system for ion transport. Computational and Applied Mathematics, 38(3):108, May 2019.

[85] G. Giacomin and J. L. Lebowitz. Phase segregation dynamics in particle systems with long range interactions. I. Macroscopic limits. J. Stat. Phys., 87(1/2):37–61, 1997.

[86] A. Glitzky and J. A. Griepentrog. Discrete Sobolev-Poincaré inequalities for Voronoi finite volume approximations. SIAM J. Numer. Anal., 48:372–391, 2010.

[87] Y. Gu and J. Shen. Bound preserving and energy dissipative schemes for porous medium equation. J. Comput. Phys., 410:109378, 2020.

[88] W. Hackbusch. Elliptic Differential Equations: Theory and Numerical Treatment. Springer Series in Computational Mathematics 18. Springer-Verlag Berlin Heidelberg, 2nd edition, 2017.

[89] R. Herbin. An error estimate for a finite volume scheme for a diffusion–convection problem on a triangular mesh. Numerical Methods for Partial Differential Equations, 11(2):165–173, 1995.

[90] X. Huo, H. Liu, A. E. Tzavaras, and S. Wang. An energy stable and positivity-preserving scheme for the Maxwell-Stefan diffusion system. arXiv:2005.08062, 2020.

[91] A. Jüngel. The boundedness-by-entropy method for cross-diffusion systems. Nonlinearity, 28(6):1963–2001, 2015.

[92] A. Jüngel. Entropy methods for diffusive partial differential equations. SpringerBriefs in Mathematics. Springer, [Cham], 2016.

[93] A. Jüngel and O. Leingang. Convergence of an implicit Euler Galerkin scheme for Poisson-Maxwell-Stefan systems. Adv. Comput. Math., 45(3):1469–1498, 2019.

[94] A. Jüngel and I. V. Stelzer. Existence Analysis of Maxwell–Stefan Systems for Multicomponent Mixtures. SIAM Journal on Mathematical Analysis, 45(4):2421–2440, Jan. 2013. Publisher: Society for Industrial and Applied Mathematics.

[95] E. Keszei. Chemical thermodynamics: an introduction. Springer Science & Business Media, 2013.

[96] T. Koprucki and K. Gärtner. Discretization scheme for drift-diffusion equations with strong diffusion enhancement. Optical and Quantum Electronics, 45(7):791–796, 2013.

[97] J. Kou, S. Sun, and X. Wang. Linearly decoupled energy-stable numerical methods for multicomponent two-phase compressible flow. SIAM J. Numer. Anal., 56(6):3219–3248, 2018.

[98] J. Leray and J. Schauder. Topologie et équations fonctionnelles. Annales scientifiques de l'École Normale Supérieure, 51((3)):45–78, 1934.

[99] J.-L. Liu and B. Eisenberg. Poisson-nernst-planck-fermi theory for modeling biological ion channels. The Journal of chemical physics, 141(22):12B640_1, 2014.

[100] R. Maex. On the Nernst–Planck equation. Journal of Integrative Neuroscience, 16(1):73–91, Jan. 2017. Publisher: IOS Press.

[101] Y. Marcus. The standard partial molar volumes of ions in solution. part 4. ionic volumes in water at 0–100 °c. Journal of Physical Chemistry B, 113:10285–10291, 2009.

[102] J. C. Maxwell. On the Dynamical Theory of Gases. Philosophical Transactions of the Royal Society of London, 157:49–88, 1867. Publisher: The Royal Society.

[103] A. D. McNaught and A. Wilkinson, editors. IUPAC Compendium of Chemical Terminology (the "Gold Book"). Blackwell Scientific, Oxford, 1997. Online version (2019-) created by S. J. Chalk. https://doi.org/10.1351/goldbook.

[104] A. Mielke. A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. Nonlinearity, 24(4):1329–1346, 2011.

[105] A. Mielke, M. A. Peletier, and D. R. M. Renger. On the relation between gradient flows and the large-deviation principle, with applications to Markov chains and diffusion. Potential Anal., 41(4):1293–1327, 2014.

[106] T. J. Murphy and N. J. Walkington. Control volume approximation of degenerate two-phase porous flows. SIAM J. Numer. Anal., 57(2):527–546, 2019.

[107] L. Onsager. Reciprocal relations in irreversible processes. I. Physical Review, 37:405–426, 1931.

[108] L. Onsager. Reciprocal relations in irreversible processes. II. Physical Review, 38:2265–2279, 1931.

[109] B. Ostle and H. L. Terwilliger. A comparison of two means. 17:69–70, 1957.

[110] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. Comm. Partial Differential Equations, 26(1-2):101–174, 2001.

[111] F. Otto and W. E. Thermodynamically driven incompressible fluid mixtures. J. Chem. Phys., 107(23):10177–10184, 1997.

[112] G. Paasch and S. Scheinert. Charge carrier density of organics with gaussian density of states: analytical approximation for the gauss–fermi integral. Journal of Applied Physics, 107(10):104501, 2010.

[113] M. A. Peletier. Variational modelling: Energies, gradient flows, and large deviations. Lecture Notes, Würzburg. Available at http://www.win.tue.nl/∼mpeletie, Feb. 2014.

[114] M. A. Peletier, R. Rossi, G. Savaré, and O. Tse. Jump processes as Generalized Gradient Flows. arXiv:2006.10624 [math], June 2020. arXiv: 2006.10624.

[115] J. Revels, M. Lubin, and T. Papamarkou. Forward-mode automatic differentiation in julia. arXiv:1607.07892 [cs.MS], 2016.

[116] D. L. Scharfetter and H. K. Gummel. Large-signal analysis of a silicon Read diode oscillator. IEEE Transactions on Electron Devices, 16(1):64–77, Jan. 1969.

[117] S. Selberherr. Analysis and simulation of semiconductor devices. Springer, 2012.

[118] J. Shen and J. Xu. Unconditionally bound preserving and energy dissipative schemes for a class of Keller-Segel equations. submitted for publication.

[119] J. Shen, J. Xu, and J. Yang. A new class of efficient and robust energy stable schemes for gradient flows. SIAM Rev., 61(3):474–506, 2019.

[120] Silvaco International. Atlas user's manual, 2016. Santa Clara, CA.

[121] J. Stefan. Über das Gleichgewicht und die Bewegung, insbesondere die Diffusion von Gasgemengen. page 62.

[122] Z. Sun, J. A. Carrillo, and C.-W. Shu. An entropy stable high-order discontinuous Galerkin method for cross-diffusion gradient flow systems. Kinet. Relat. Models, 12(4):885–908, 2019.

[123] Synopsys, Inc. Sentaurus device userguide, 2010. Mountain View, CA.

[124] P. Vágner, C. Guhlke, V. Miloš, R. Müller, and J. Fuhrmann. A continuum model for yttria-stabilized zirconia incorporating triple phase boundary, lattice structure and immobile oxide ions. Journal of Solid State Electrochemistry, pages 1–20, 2019.

[125] S. Van Mensfoort and R. Coehoorn. Effect of gaussian disorder on the voltage dependence of the current density in sandwich-type devices based on organic semiconductors. Physical Review B, 78(8):085207, 2008.

[126] Z. Yu and R. Dutton. SEDAN III. www-tcad.stanford.edu/tcad/programs/sedan3.html, July 1988.

[127] M. Zagour. Modeling and mathematical analysis of complex systems: Kinetic and macroscopic approaches and applications in biology and vehicular traffic. Theses, Université Cadi Ayyad Marrakech (Maroc), Apr. 2019.

[128] S. Zhou, Y. Wang, X. Yue, and C. Wang. A second order numerical scheme for the annealing of metal-intermetallic laminate composite: A ternary reaction system. J. Comput. Phys., 374:1044–1060, 2018.

Entropic numerical approximations for cross-diffusion systems arising in physics

Résumé

Dans cette thèse, on s'intéresse à la modélisation et à l'analyse numérique de systèmes physiques à diffusion croisée. Les modèles considérés décrivent notamment les phénomènes à l'œuvre dans les batteries et lors de la fabrication de panneaux solaires.
Dans les deux premiers chapitres, on étudie différents schémas pour un modèle de diffusion des ions adapté aux concentrations élevées et proposé en 2013. Le premier chapitre concerne l'étude du cas simplifié à une inconnue de concentration. On y propose quatre schémas volumes finis à deux points pour lesquels on démontre l'existence de solutions physiquement réalistes, puis la convergence de ces solutions approchées vers une solution faible du problème continu pour deux des quatre schémas. Dans le deuxième chapitre, on s'attaque à une variante sans pression du problème multi-espèces de 2013 avec les deux schémas validés lors du premier chapitre. On y démontre à nouveau l'existence de solutions approchées. Sous réserve de non-disparition du solvant, on démontre enfin la convergence de ces solutions. Les phénomènes de diffusion croisée amènent à définir une notion de coercivité assez minimale.
Dans un second temps, on s'intéresse à d'autres mécanismes de dissipation de l'énergie libre. On se penche d'abord sur des variantes au problème étudié dans les ceux premiers chapitres. Ces variantes sont obtenues par des techniques de modélisation variationnelle. Ce choix de modélisation permet un traitement naturel de la pression lorsqu'elle fait partie du modèle. Quelques simulations numériques mettent en exergue les différences de comportements de ces modèles à hautes concentrations. Enfin, le dernier chapitre traite d'un modèle mathématique de diffusion à l'œuvre dans la fabrication de certains panneaux solaires. On y représente un système de diffusion croisée comme une perturbation de l'équation de la chaleur. Le traitement de la perturbation repose explicitement sur la construction d'une concentration d'interface, technique introduite dans les chapitres précédents et qui permet d'étendre des méthodes d'analyse pour les schémas centrés à des schémas plus généraux. Cette technique permet de démontrer la préservation de dérivations discrètes.

Keywords: cross-diffusion, partial differential equations, numerical analysis, modelisation

Laboratoire Paul Painlevé
Laboratoire Paul Painlevé – Université Lille 1 – Cité Scientifique – Bâtiments M2 et M3 – Villeneuve d'Ascq

Entropic numerical approximations for cross-diffusion systems arising in physics

Abstract

In this thesis, we are interested in the modeling and numerical analysis of physical systems with cross diffusion. The models considered describe in particular the phenomena at work in batteries and during the manufacturing of solar panels.
In the first two chapters, different schemes for an ion diffusion model suitable for high concentrations and proposed in 2013 are studied. The first chapter focuses on the study of the simplified case with only one concentration unknown. Four two-point finite volume schemes are proposed for which we prove the existence of physically realistic solutions, and the convergence of these approximate solutions to a weak solution of the continuous problem for two of the four schemes. In the second chapter, we tackle a pressureless variant of the 2013 multispecies problem with the two schemes validated in the first chapter. The existence of approximate solutions is proved again. Under the condition of non-disappearance of the solvent, we finally show the convergence of these solutions. The cross diffusion effects lead to define a rather minimal notion of coercivity.
In the two following chapters, we consider other mechanisms of free energy dissipation. We first focus on variants of the problem studied in the first part obtained by variational modeling techniques. This choice of modeling allows for a natural treatment of the pressure when part of the model. Some numerical simulations highlight the differences in the behavior of these models at high concentrations. Finally, the last chapter deals with a mathematical model of diffusion at work in the manufacture of some solar panels. We represent a system of cross diffusion as a perturbation of the heat equation. The treatment of the perturbation is explicitly based on the construction of an interface concentration, a technique introduced in previous chapters that allows the extension of analysis methods for centered schemes to more general schemes. This technique allows proving the preservation of discrete derivations.

Keywords: cross-diffusion, partial differential equations, numerical analysis, modelisation