Analysis of Algebraic Flux Correction Schemes for Transient Convection-Diffusion Equations

Master's thesis

submitted by

Paul Korsmeier

supervised by

Prof. Dr. Volker John
Dr. Petr Knobloch

Institute of Mathematics
Department of Mathematics and Computer Science
Freie Universität Berlin

Berlin, July 2018

Contents

Latin Letters

Notation	Description	Page
a	bilinear form associated to L	9
b	divergence-free convection field	9
$\tilde{b}$	$= (b, 1)$	27
$\mathcal{B}_i(\Sigma)$	boundary parts of $\mathcal{S}_\Sigma$, $i = 1, 2, 3$	30
$B_r(x)$	open norm ball of radius r around $x \in \mathbb{R}^d$	x
$C \in \mathbb{R}^{N\times N}$	full convection matrix	xii
$C > 0$	generic positive constant	
c	non-negative reaction term	9
$C_0^\infty(S)$	compactly supported smooth functions on an open set $S \subset \mathbb{R}^n$	x
C_a	bound for a	10
c_a	positive coercivity constant of a	10
C_i	i-th cell of the barycentric dual mesh	39, 42
$c_{ij}(\tau)$	auxiliary function in Section 5.4.3	71
C_{PF}	constant in Poincaré-Friedrichs inequality	10
D	diffusion matrix	xii
D_f	set of differentiability points of a function f	72
$\mathcal{D}_f(x;\cdot)$	a notion of generalised directional derivative of a function f at x	72
F_+, F_0	auxiliary functions in Chapter 3	31, 33
F	function in Section 5.4.3 to find solutions $h(\tau)$	71
$f \in L^2(\Omega_T)$	right-hand side (linear source term) (unless locally defined otherwise)	9
f	raw antidiffusive flux in Chapter 5 (unless locally defined otherwise)	68
f_{ij}	raw internodal flux from node j to node i	67, 69, 71
$\mathcal{G}$	backward exit locus	28

Notation	Description	Page
$\mathcal{G}_-$	backward exit locus in $\partial\Omega \times [0,T)$	28
$g_{ij}(\tau)$	auxiliary function in Section 5.4.3	71
$h \in \mathbb{R}^N(\mathbb{R}^M)$	solution update from time step n to $n+1$ in Chapter 5	71
$h > 0$	mesh width of $\mathcal{T}$	xi
h_T	diameter of d-simplex T	xi
$\mathcal{I}$	index set of interior nodes	xi
$K \in \mathbb{R}^{N\times N}$	$:= -(C + \epsilon D)$	xii
$K(i)$	index set of neighbours of node p_i	xi
K_i	positive constant independent of ϵ in Chapter 3	33
$L \in \mathbb{R}^{N\times N}$	$:= K + Y$	69
L	elliptic differential operator	9, 17
$M \in \mathbb{N}$	number of interior nodes of $\mathcal{T}$	xi
M_C	full consistent mass matrix	xii
M_L	full lumped mass matrix	xii
$N \in \mathbb{N}$	number of nodes of $\mathcal{T}$	xi
$\mathcal{N}$	node set of $\mathcal{T}$	xi
$\mathcal{N}^\circ$	interior node set of $\mathcal{T}$	xi
N_1, N_2	certain null sets	86
$P_i^\pm$	sum of non-negative (non-positive) antidiffusive fluxes into node/cell i	62, 71
$\mathbb{P}^k(S)$	polynomials of degree $\leq k$ on the set $S \subset \mathbb{R}^d$	xi
$PC^k(V, \mathbb{R}^m)$	piecewise C^k functions mapping $V \subset \mathbb{R}^n$ to $\mathbb{R}^m$	76
Pe	cell Péclet number	41
$Q_i^\pm$	admissible non-negative (non-positive) correction for node/cell i	62, 71
$R_i^\pm$	$\frac{Q_i^\pm}{P_i^\pm}$, cut off at 1	62, 71
$\mathcal{S}_\Sigma$	characteristic tube of a set $\Sigma \subset \mathcal{G}$	29
$T > 0$	a positive time	9
$T \in \mathcal{T}$	a d-simplex	xi
$\mathcal{T}$	triangulation of Ω	x
TV	discrete total variation	38
u	weak solution to convection-diffusion-reaction equation; in Chapter 3: solution to reduced probem	9, 13, 27

Notation	Description	Page
u_0	initial condition	9
u_ϵ	solution to convection-diffusion-reaction equation in Chapter 3	27
U_γ	neighbourhood of $\pi_x(\mathcal{B}_2(V))$	31
$V^k := V^k(\mathcal{T})$	continuous piecewise polynomial (of order $\leq k$) functions	xii
$V_0^k := V_0^k(\mathcal{T})$	$V^k(\mathcal{T}) \cap H_0^1(\Omega)$	xii
V_h^k	shorthand for $V^k(\mathcal{T}_h)$	xii
$V_{h,0}^k$	shorthand for $V_0^k(\mathcal{T}_h)$	xii
Var	continuous total variatiation	37
$\overline{x}_\pm$	forward/backward exit points of $\gamma_{\overline{x}}$	28
Y	upwinding/discrete diffusion matrix	46, 69

Greek Letters

Notation	Description	Page
α_{ij}	correction factor for antidiffusive fluxes	62, 70
δ, ε	generic small positive numbers	x
ϵ	diffusion coeefficient	9
$\gamma_{\overline{x}}$	characteristic associated to $\overline{x} \in \Omega \times (0,T)$	27
Γ_+	outflow boundary	29
Γ_-	inflow boundary	29
Γ_0	parabolic boundary	29
Γ_T	parabolic boundary of Ω_T	17
Ω	a domain	x
Ω_T	the cylider $\Omega \times (0,T]$	9
ρ_T	insphere radius of d-simplex T	xi
σ_T	shape factor of d-simplex T	xi
$\sigma_{\mathcal{T}}$	shape factor of $\mathcal{T}$	xi
τ	time-step length in a time-discretised scheme	24, 38, 70
$\tau_\pm(\overline{x})$	forward/backward exit times associated to $\overline{x} \in \Omega \times (0,T)$	27
θ	implicitness parameter in a θ-stepping	24
φ_i	i-th piecewise linear nodal basis function	xii

Other Symbols

Notation	Description	Page
x^+	$= \max(0, x) \geq 0$	
x^-	$= \min(0, x) \leq 0$	
$\lvert\cdot\rvert$	Hausdorff measure, absolute value, Euclidean norm or induced matrix norm	x
S°	interior of a set $S \subset \mathbb{R}^d$	
$\overline{S}$	closure of a set $S \subset \mathbb{R}^d$	
X^*	topological dual of a Banach space X	
$\langle \cdot, \cdot \rangle := \langle \cdot, \cdot \rangle_{X^* \times X}$	duality pairing associated to a Banach space X	11
$(\cdot, \cdot)$	L^2 product on Ω	x
A°	restriction of a matrix $A \in \mathbb{R}^{N \times N}$ to interior nodes	xii
$\partial_B f(x)$	Bouligand subdifferential of a function f at point x	72
$\partial f(x)$	Clarke's generalised Jacobian of a function f at point x	72

Notation and Nomenclature

The notation in this work is certainly standard, but we introduce at this point some frequently used notation and nomenclature to preclude ambiguity and for later reference.

Definition 0.1 (General nomenclature and notation)**.** Let $d \in \mathbb{N}$.

(i) By a *domain* $\Omega \subset \mathbb{R}^d$ we mean an open connected set.

(ii) Whenever we say *smooth*, we mean infinitely many times differentiable.

(iii) For an open set $S \subset \mathbb{R}^n$ with $n \in \mathbb{N}$, we denote by $C_0^\infty(S)$ the set of smooth real-valued functions with compact support in S.

(iv) $(\cdot, \cdot)$ denotes the L^2 scalar product on Ω.

(v) $|\cdot|$ is defined depending on the context:

- For $x \in \mathbb{C}$, $|x|$ denotes the absolute value of x.
- For a vector $v \in \mathbb{R}^d$, if not stated otherwise, $|v|$ denotes the Euclidean 2-norm.
- If $S \subset \mathbb{R}^d$ is of Hausdorff dimension d', then by $|S|$ we mean the d'-dimensional Hausdorff measure of S.

(vi) For $x \in \mathbb{R}^d$ and $r > 0$, $B_r(x)$ denotes the open Euclidean norm ball of radius r around x.

(vii) Due to the symbol ϵ being reserved as the diffusion coefficient (see (1.1)), we will use δ, ε to be able to carry out calculus in the usual notation. No confusion should arise from this.

(viii) $C > 0$ is a generic positive constant whose value is allowed to change even between two usages in the same line.

△

Definition 0.2 (Triangulations)**.** Let $\Omega \subset \mathbb{R}^d$ be a domain.

(i) A *triangulation* or *simplicial partition* or *simplicial mesh* $\mathcal{T}$ on Ω is a collection of closed d-simplices T such that $T^\circ \cap T'^\circ = \emptyset$ for all distinct $T, T' \in \mathcal{T}$ and

$$\bigcup_{T \in \mathcal{T}} = \overline{\Omega}. \tag{0.1}$$

Even when not explicitly stated, we shall always assume that the triangulation is *regular*, i.e.

$$T \cap T' = \emptyset \text{ or } T \cap T' \text{ is a subsimplex of both } T \text{ and } T'. \tag{0.2}$$

(ii) Equation (0.1) implies that a triangulated domain Ω is *polyhedral*. Actually, we take the existence of a triangulation as the defining property of a polyhedral domain.

(iii) Let $\mathcal{T}$ be a regular triangulation of Ω. We do not differentiate between vertices and nodes and define

(a) $\mathcal{N} := \{p : p \text{ is a vertex of a simplex } T \in \mathcal{T}\}$ (the *node set*)
(b) $N := \#(\mathcal{N})$.
(c) $\mathcal{N}^\circ := \mathcal{N} \cap \Omega$ (the *interior node set*)
(d) $\mathcal{I} := \{i \in \{1, \dots, N\} : p_i \in \mathcal{N}^\circ\}$ (the indices of interior nodes)
(e) $M := \#(\mathcal{I})$.

(iv) For a closed d-simplex T, we define

(a) $h_T := \max\{|x - y| : x, y \in T\}$ (its *diameter*)
(b) $\rho_T := \max\{\rho > 0 : \exists\, x \in T : B_r(x) \subset T\}$ (the *insphere radius*)
(c) $\sigma_T := h_T/\rho_T$ (the *shape factor*)
(d) a *side* or *facet* of T to be a $(d-1)$-subsimplex.

(v) For a triangulation $\mathcal{T}$ define

(a) $h := \max_{T \in \mathcal{T}} h_T$ (its *mesh width*)
(b) $\sigma_{\mathcal{T}} := \max_{T \in \mathcal{T}} \sigma_T$ (its *shape factor*)

A family of triangulations $(\mathcal{T}_h)_{h \in I}$ (usually formally indexed by its mesh width) is called *shape-regular* if $\sigma := \sup_{h \in I} \sigma_{\mathcal{T}_h} < \infty$. It is called *quasi-uniform* if there exists a constant $\tau > 0$ such that $\min_{T \in \mathcal{T}_h} \geq \tau h$ for all h.

(vi) For an enumerated node set $\mathcal{N} = \{p_i : i = 1, \dots, N\}$ and a node $p_i \in \mathcal{N}$ we say that a node p_j, $j \neq i$ is a *neighbour* of p_i if there exists $T \in \mathcal{T}$ with $p_i, p_j \in T$. We define the *set of neighbour indices* of node p_i as

$$K(i) := \{j \in \{1, \dots, N\} : p_j \text{ is a neighbour of } p_i\}. \tag{0.3}$$

△

Definition 0.3 (Finite Element Spaces)**.** Let $k \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$ a domain partitioned by a regular triangulation $\mathcal{T}$.

(i) For any set $S \subset \mathbb{R}^d$ of sufficient cardinality we define

$$\mathbb{P}^k(S) := \left\{ f : S \to \mathbb{R} : f(x) = \sum_{\alpha \in \mathbb{N}_0^d : |\alpha| \leq k} c_\alpha x^\alpha \right\} \tag{0.4}$$

with the usual multi-index notation.

(ii) The space $V^k := V^k(\mathcal{T})$ denotes the *continuous piecewise k-polynomial elements*

$$V^k(\mathcal{T}) := \{f \in C^0(\Omega) : f|_T \in \mathbb{P}^k(T) \text{ for all } T \in \mathcal{T}\} \subset H^1(\Omega) \tag{0.5}$$

and $V_0^k := V_0^k(\mathcal{T}) := V^k(\mathcal{T}) \cap H_0^1(\Omega)$ is its counterpart with vanishing boundary values.

(iii) If $\mathcal{T} = \mathcal{T}_h$, the shorthands V_h^k and $V_{h,0}^k$ can be convenient.

(iv) For an enumerated node set $\mathcal{N} = \{p_1, \ldots, p_N\}$ and $i \in \{1, \ldots, N\}$ define the *i-th nodal basis function* or *i-th hat function* $\varphi_i \in V^1(\mathcal{T})$ by $\varphi_i(p_j) = \delta_{ij}$.

△

Definition 0.4 (Matrices of the standard Galerkin method)**.** Let $\mathcal{T}$ be a triangulation of Ω with nodes p_i, $i = 1, \ldots, N$. Then we define the associated *full consistent mass matrix* M_C, the *full diffusion matrix* D and the *full convection matrix* C by

$$M_C \in \mathbb{R}^{N\times N} \qquad m_{ij} := (\varphi_j, \varphi_i), \tag{0.6}$$

$$D \in \mathbb{R}^{N\times N} \qquad d_{ij} := (\nabla\varphi_j, \nabla\varphi_i) \tag{0.7}$$

$$C \in \mathbb{R}^{N\times N} \qquad c_{ij} := (b \cdot \nabla\varphi_j, \varphi_i), \tag{0.8}$$

respectively, the *stiffness matrix* or *negative transport operator* K as

$$K := -(C + \epsilon D) \tag{0.9}$$

and the *full lumped mass matrix* $M_L \in \mathbb{R}^{M\times M}$ by

$$M_L := \operatorname{diag}(m_1, \ldots, m_M) \qquad m_i := \sum_{j=1}^{N} m_{ij}. \tag{0.10}$$

For $A \in \{M_C, M_L, D, C, K\}$ define the *restricted* counterpart $A^\circ := A_{\mathcal{I}\mathcal{I}} := (a_{ij})_{i,j\in\mathcal{I}}$. △

1. Introduction

In this thesis we will be interested in the numerical solution of time-dependent (also called unsteady or transient) convection-diffusion equations and a method called "Finite Element Flux Corrected Transport" (FEM-FCT) that is expected to alleviate or resolve a severe problem inherited in the standard V_0^1 finite element treatment thereof: the emergence of spurious (i.e. unphysical) oscillations in the vicinity of steep gradients in the approximate solutions. These occur when the diffusion part of the differential operator is dominated by its convective part.

Specifically, let $\Omega \subset \mathbb{R}^d$ be a domain, $T > 0$ some positive time, $\Omega_T := \Omega \times (0, T]$ the *cylinder*, $b : \Omega_T \to \mathbb{R}^d$ a given vector field satisfying $\operatorname{div}(b) = 0$ and $0 < \epsilon \ll \|b\|_{L^\infty(\Omega_T)}$.
Then for some initial datum $u_0 : \Omega \to \mathbb{R}$ and some $f : \Omega_T \to \mathbb{R}$ the differential formulation of the problem of interest is to find $u : \overline{\Omega_T} \to \mathbb{R}$ satisfying

$$\begin{cases} u_t - \epsilon \Delta u + b \cdot \nabla u = f & \text{in } \Omega_T \\ u = 0 & \text{on } \partial\Omega \times [0, T] \\ u = u_0 & \text{on } \Omega \times \{t = 0\}. \end{cases} \tag{1.1}$$

We will generally restrict ourselves to homogeneous Dirichlet boundary conditions on $\partial\Omega \times [0, T]$.

In order to witness in a simple case the spurious effects of applying a standard finite element method to such a problem, let us contrast our expectations of the true solution and the reality of the discrete solution for the one-dimensional example problem

$$\begin{cases} u_t - \epsilon u_{xx} + u_x = 0 & \text{in } (0, 4) \times (0, T] \\ u = 0 & \text{on } \{0, 4\} \times [0, T] \\ u = \chi_{[1,2]} & \text{on } (0, 4) \times \{t = 0\}, \end{cases} \tag{1.2}$$

where $\chi_{[1,2]}$ is the characteristic function of the interval $[1, 2]$.

We expect the solution u to display only the two effects of convection and diffusion; for instance if we interpret $u_0 = \chi_{[1,2]}$ as an initial spatial distribution of a substance's concentration in the one-dimensional container $\Omega = (0, 4)$ in which the fluid is transported with constant velocity $b = 1$ (the container's walls are no obstacle to the fluid flow), we expect $u(t)$ for $0 < t < 2$ to be a (slightly, since $\epsilon \ll b$) smoothed version of the shifted initial profile $\chi_{[1+t,2+t]}$. For $t > 2$ the homogeneous

Dirichlet condition at $x = 4$ will steer the profile down rapidly to the value 0 at that point.

Neglecting for a moment the boundary conditions and pretending the domain were $\Omega = \mathbb{R}$, (1.2) describes pure diffusion in a shifting coordinate system. Since it is the nature of the diffusion (= heat) equation on $\mathbb{R}$ that features of the initial profile flatten out, become blurred (convolved with a Gaussian kernel of increasing width, in fact) and less extreme over time, we can expect the values of $u(\cdot, t)$ to always remain within the interval $[0, 1]$ and $u(\cdot, t)$ to be of decreasing total variation as t increases. These two principles will be reflected in the *parabolic weak maximum principle* (see Theorem 2.16) and Proposition 4.2, respectively. A good numerical approximation should have these properties in a discrete sense, too.

If, however, we consider the discretisation using the finite element space $V^1_{h,0}$ over the triangulation $\mathcal{T}_h$ given by the equidistant grid $0 = x_0 < x_1 < \cdots < x_N = 4$ dividing $\Omega = (0, 4)$ into N intervals of length $h = 4/N$ and θ-stepping with a constant time-step $\tau > 0$ for the time-discretisation, then the results violate these two requirements. For our example, we fix $\epsilon := 10^{-3} \ll 1 = b$.

Let u_i^n denote the discrete solution at grid-point $(ih, n\tau)$. Then this discretisation gives the implicit scheme

$$\frac{2}{3}\delta_t u_i^{n+1} + \frac{1}{3}(\delta_t u_{i+1}^{n+1} - \delta_t u_{i-1}^{n+1}) = \theta\left(\epsilon L_1 u_i^{n+1} + L_0^c u_i^{n+1}\right) + (1-\theta)\left(\epsilon L_1 u_i^n + L_0^c u_i^n\right), \tag{1.3}$$

where

$$\delta_t u_i^{n+1} := \frac{u_i^{n+1} - u_i^n}{\tau}, \qquad L_1 u_i^n := \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}, \qquad L_0^c u_i^n := -\frac{u_{i+1}^n - u_{i-1}^n}{2h} \tag{1.4}$$

and $\theta \in [0, 1]$. Setting $u_0^n = u_N^n = 0$ for all $n \in \mathbb{N}$ realises the homogeneous Dirichlet boundary condition. For $\theta = 0, 0.5, 1$ this is the *forward Euler*, *Crank-Nicolson* and *backward Euler* method, respectively.

For instance, let us choose $N = 100$ and plot the solutions with $\theta = 0$, 0.5, 1 at four different times t. We set $\tau = h^2$ to make sure that the forward Euler method is stable (for values significantly larger than that, this method produces wild oscillations resulting in numerical overflow). The results are shown in Figure 1.1. Oscillations lead to values outside the initial function range $[0, 1]$ and cause an increase in total variation. The sharp drop near the boundary point $x = 4$ causes particularly severe ones. Since this effect occurs not only for $\theta = 0$ but also for the unconditionally stable Crank-Nicolson and backward Euler method, we see that it cannot be attributed to a possible instability of the time-discretisation, but rather that it is an inherent deficiency in the finite element space-discretisation.

Hence let us now focus on the forward Euler method. In order to make this method truly explicit, a technique called *mass lumping* is commonly used. In the considered 1D case, this diagonalisation of the mass-matrix $M = (\varphi_j, \varphi_i)_{i,j=0,\dots N}$ in conjunction with setting $\theta = 0$ yields the scheme

$$\delta_t u_i^{n+1} = \epsilon L_1 u_i^n + L_0^c u_i^n. \tag{1.5}$$

The explicitness of this method makes it computationally cheaper, but more importantly there are now conditions on h, τ that, when complied with, ensure that the resulting method no longer

produces the above negative effects. In Section 4.1.3 we will learn that the crucial numbers here are

$$Pe := \frac{bh}{\epsilon} \qquad \text{and} \qquad \gamma := \tau \frac{2\epsilon}{h^2}, \tag{1.6}$$

where Pe is the so-called *cell Péclet number* that characterises how much b dominates ϵ in a cell of size h.

The following table summarises an experiment in which N and τ are varied independently of each other so that Pe and γ are in a range around certain threshold values. A green field symbolises that, at time $t = 0.01$ (actually, $t = n\tau$ for $n = \lceil 0.01/\tau \rceil$), the numerical solution's range is contained in $[0, 1]$, whereas a red field represents the case that this range is exceeded.

γ \ Pe	3.0	2.5	2.2	2.1	2.0	1.9	1.8	1.5	1.0
1.30									
1.20									
1.10									
1.05									
1.05									
1.00									
0.95									
0.90									
0.80									
0.70									

This indicates that the mass-lumped explicit Euler scheme does not show oscillations if and only if $Pe \leq 2$ and $\gamma \leq 1$. That this condition is indeed sufficient will be shown in Section 4.1.3.

If we change the definitions of Ω and u_0 slightly in order to march the profile to the right boundary more quickly, we can compare for $Pe = 1$ and $\tau = h^2/10\epsilon = 10^{-4}$ the lumped and non-lumped scheme at $t = 10^{-3}$ and $t = 0.2$. It is seen that, as predicted by the above table, the lumped version is free of oscillations for these values h, τ. The non-lumped version no longer shows the terrible boundary oscillation of before, but still some oscillations at the first time-steps, see Figure 1.2 and Figure 1.3.

Now, the conditions $Pe = bh/\epsilon \leq 1$ and $\tau \leq h^2/2\epsilon$ are extremely restrictive for a large ratio $|b|/\epsilon$ and likely to make also the mass-lumped method (1.5) unusable due to excessive computational effort. In order to make the explicit Euler scheme non-oscillatory with a reasonable time-step constraint, we need to replace the central difference approximation L_0^c by the one-sided difference

$$L_0^u u_i^n := -\frac{u_i^n - u_{i-1}^n}{h}, \tag{1.7}$$

where the superscript u stands for *upwind*. Since the "wind" $b > 0$ is directed towards the right, this uses only information located upwind from (i.e. to the left of) node i to compute the advection

contribution to u_i^{n+1}. With this modification, the time-step threshold for stability will turn out to be

$$\tau_u := \left(\frac{b}{h} + \frac{2\epsilon}{h^2} \right)^{-1}. \tag{1.8}$$

A quick experiment with $N = 400$ and a range of for values τ near τ_u verifies this, see Figure 1.4. For $\tau \leq \tau_u$, the scheme shows no spurious oscillations or out-of-range values, regardless of the size of the Péclet number Pe, but as can be seen from a comparison with the exact and central scheme solutions in Figure 1.5, it adds a large amount of artificial diffusion.

In short, the subject-matter of this thesis is to understand this behaviour in the one-dimensional case, find a generalisation of upwinding to the case of domains $\Omega \subset \mathbb{R}^d$ for $d \geq 2$ and then analyse the method of flux-corrected transport (FCT), which is an attempt to blend the upwinded and the standard Galerkin method in such a way that their respective strengths are maximised and their weaknesses minimised, i.e. such that both spurious oscillations and excessive numerical diffusion are kept at a minimum.

Chapter 2 covers some standard results on the existence and uniqueness of weak solutions of solutions to the linear second order parabolic problem on domains $\Omega \subset \mathbb{R}^d$ with homogeneous Dirichlet boundary conditions, as well as on the stability and convergence of the discretisation by the method of lines (first in space, then in time) using $V_{h,0}^1$ finite elements and θ-stepping. With the weak maximum principle, an important property of classical solutions will be introduced that will later give rise to the concept of *local extremum diminishing* (LED) semi-discrete schemes.

Chapter 3 is of little relevance to the following chapters but it is an interesting application of the (strong) maximum principle. It studies the uniform convergence of the convection-diffusion solutions to the *reduced problem's* solution, i.e. the problem without a diffusion term, under certain – partly natural and partly technical – conditions on the considered domain on which this convergence takes place.

In Chapter 4 we introduce the notion of total variation and *total variation diminishing* (TVD) schemes in one dimension and show that the latter is a natural property that the upwinded explicit mass-lumped scheme has. We then generalise the upwinding procedure of the convective part of the stiffness matrix to the case $d = 2$ and show that the resulting scheme can be seen as a member of the class of upwind finite element methods of Baba and Tabata. Finally, the LED principle is introduced and conditions are given such that the scheme thus upwinded is LED.

Chapter 5, finally, presents the original FCT method for conservation laws in multiple space dimensions devised by Zalesak in 1979 and Kuzmin's approach from [Kuz10] to apply this framework in a non-linear blend of the upwinded and standard finite element method. Conditions for well-posedness of the arising non-smooth non-linear problem and for local convergence of a semi-smooth Newton method are investigated.

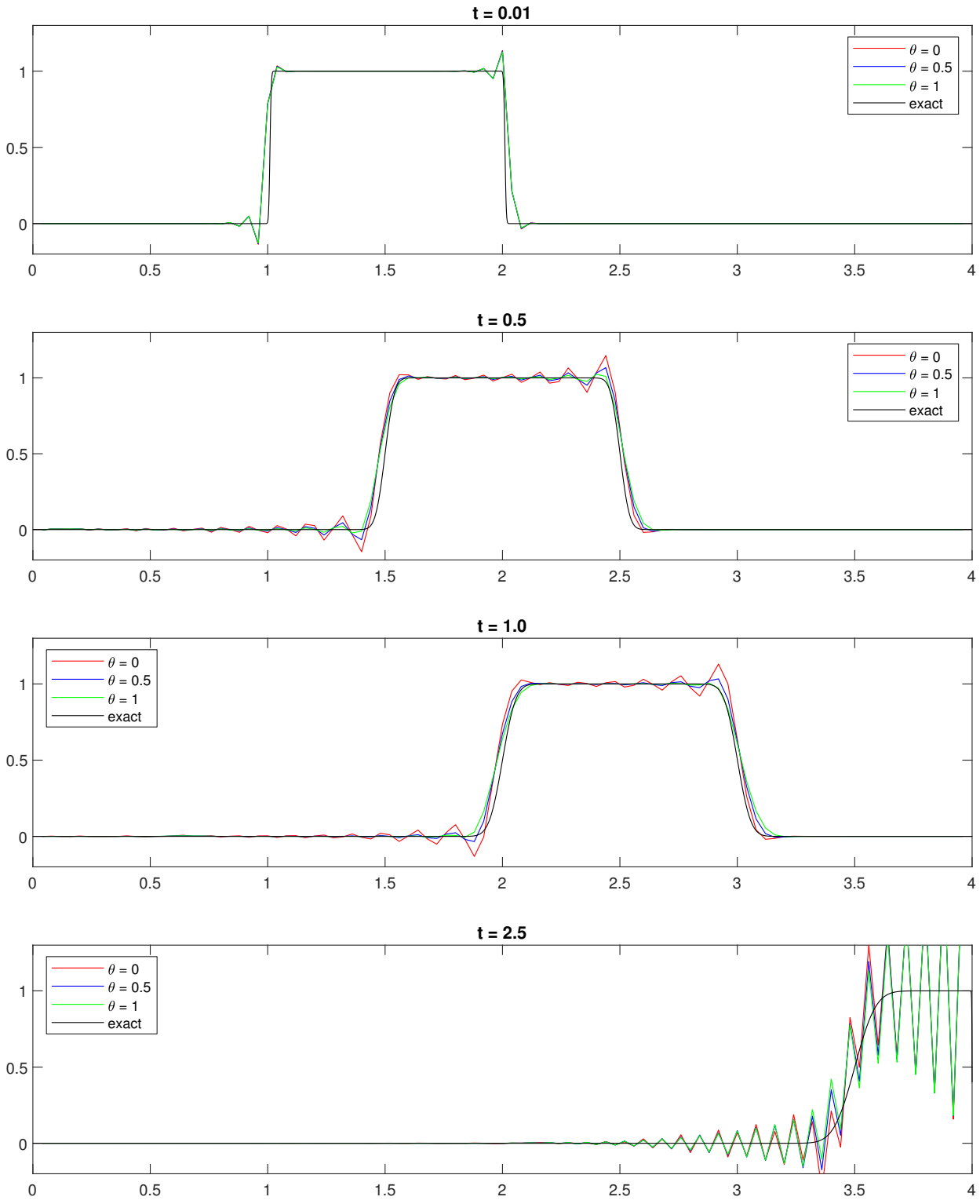

Figure 1.1: Results for $N = 100$ and $\theta = 0$, 0.5 and 1 at four different times t in comparison with the exact solution obtained by a very fine discretisation

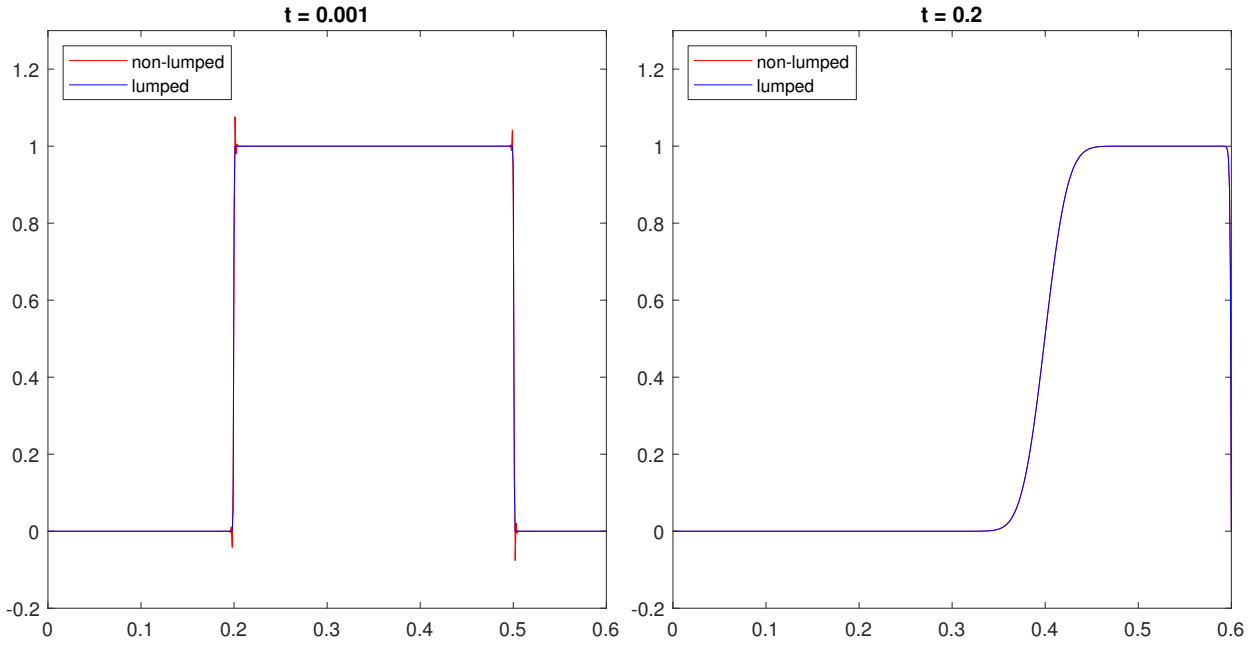

Figure 1.2: Comparison of the non-lumped scheme (1.3) and the lumped scheme (1.5) for $\Omega = (0, 0.6)$, $u_0 = \chi_{[0.2,0.5]}$, $Pe = 1$ and $\tau = h^2/10\epsilon = 10^{-4}$

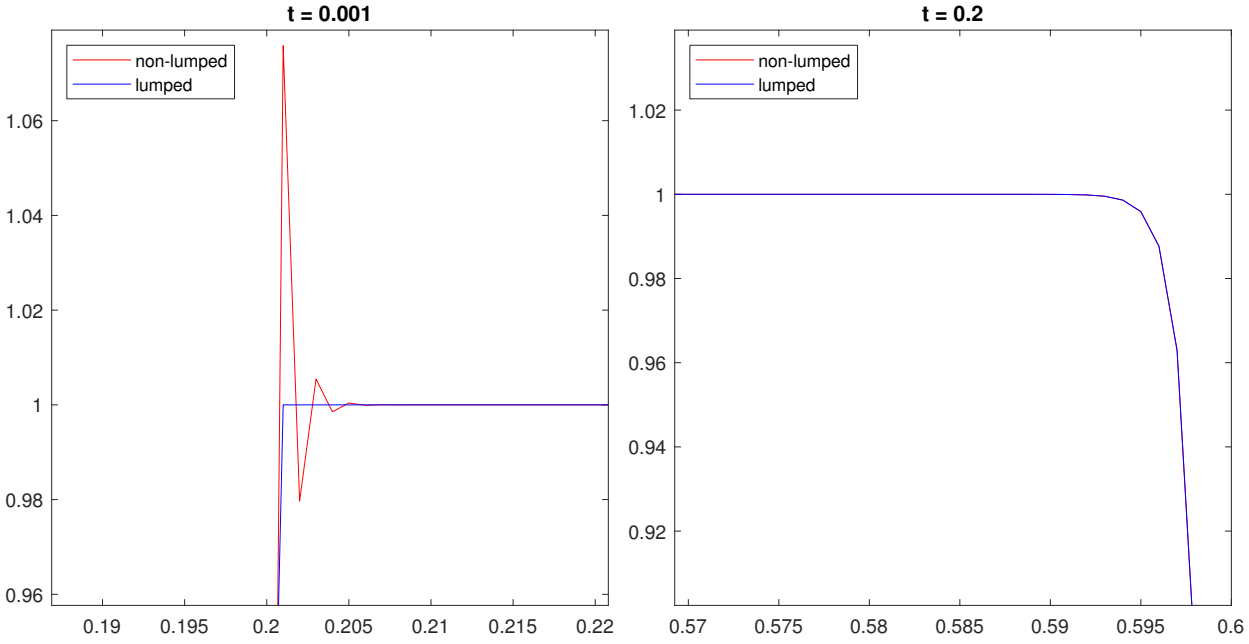

Figure 1.3: Blow-up of the interesting regions Figure 1.2

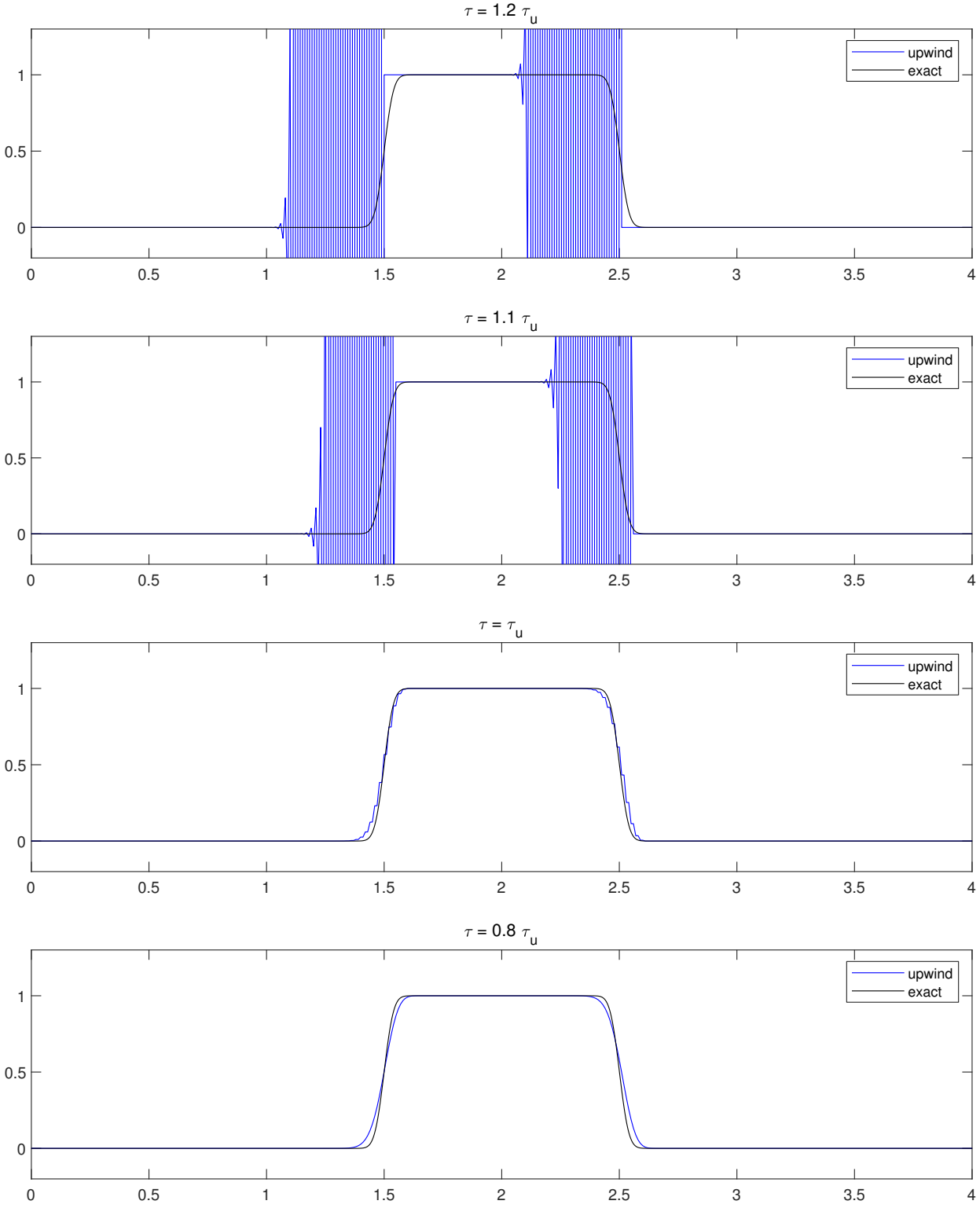

Figure 1.4: The upwinded scheme with $N = 400$ for values τ near the critical value τ_u

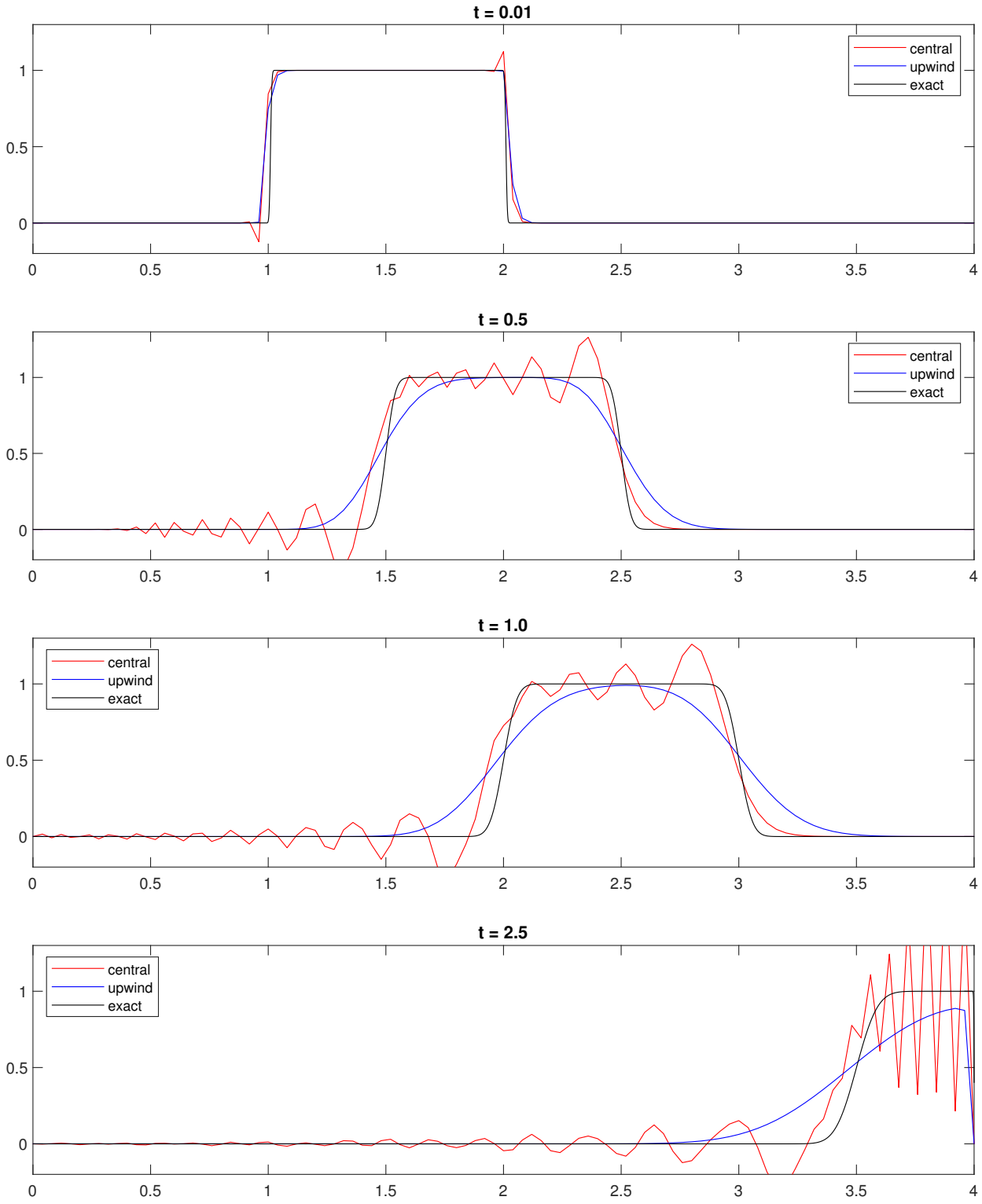

Figure 1.5: Comparison of the explicit central, explicit upwind scheme and exact solutions for $N = 100$ and $\tau = h^2$

2. Existence, Uniqueness, Maximum Principle and Standard Galerkin Approximation

Let $T > 0$ be a time and $\Omega_T := \Omega \times (0, T]$ the *cylinder*. We want to prove existence and uniqueness of a weak solution $u : \overline{\Omega_T} \to \mathbb{R}$ to

$$\begin{cases} u_t + Lu = f & \text{in } \Omega_T \\ u = 0 & \text{on } \partial\Omega \times [0, T] \\ u = u_0 & \text{on } \Omega \times \{t = 0\}, \end{cases} \tag{2.1}$$

where $Lu := -\epsilon\Delta u + b \cdot \nabla u + cu$ is the elliptic (spatial) differential operator and $\Omega \subset \mathbb{R}^d$ is a domain with Lipschitz boundary.

Assumption 2.1. We fix the following assumptions about the functions constituting the data:

$$b \in W^{1,\infty}(\Omega_T, \mathbb{R}^d) \text{ with } \operatorname{div} b = 0 \tag{2.2a}$$
$$c \in L^\infty(\Omega_T, \mathbb{R}) \text{ with } c \geq 0 \tag{2.2b}$$
$$f \in L^2(\Omega_T) \tag{2.2c}$$
$$u_0 \in L^2(\Omega). \tag{2.2d}$$

Furthermore, for the sake of simplicity, we assume the operator L and therefore the functions b and c to be *independent* of t! △

Remark 2.2. Equation (2.2a) is actually equivalent to b being Lipschitz continuous with Lipschitz constant $\|b\|_{W^{1,\infty}(\Omega_T, \mathbb{R}^d)}$ (see [Alt16, Theorem 10.5]) △

Definition 2.3. Define the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ associated to the elliptic operator L by

$$a(u, v) := \epsilon(\nabla u, \nabla v) + (b \cdot \nabla u, v) + (cu, v). \tag{2.3}$$

△

The following two results are well-known from the theory of elliptic partial differential equations:

Lemma 2.4 (Poincaré-Friedrichs inequality)**.** *There exists a constant $C_{PF}(\Omega) > 0$ such that for all $v \in H_0^1(\Omega)$*

$$\|v\|_{L^2(\Omega)} \leq C_{PF} \|\nabla v\|_{L^2(\Omega)}. \tag{2.4}$$

△

Lemma 2.5 (Boundedness and coercivity of a)**.** *The bilinear form a is bounded by*

$$C_a := \epsilon + \|b\|_{L^\infty(\Omega)} + \|c\|_{L^\infty(\Omega)}. \tag{2.5}$$

If we assume in addition that that $c - \frac{1}{2}\operatorname{div}(b) \geq 0$, then a is coercive with constant

$$c_a := (C_{PF}^2 + 1)^{-1}\epsilon. \tag{2.6}$$

△

Proof. The first part of the statement is trivial. For the coercivity, note that

$$(b \cdot \nabla u + cu, u) = \left(\frac{1}{2}b, \nabla(u^2)\right) + (cu, u) = -\left(\frac{1}{2}\operatorname{div}(b), u^2\right) + (c, u^2) \geq c_0 \|u\|_{L^2(\Omega)}^2 \geq 0, \tag{2.7}$$

so that

$$a(u,u) \geq \epsilon \|\nabla u\|_{L^2(\Omega)}^2 \geq (C_{PF}^2 + 1)^{-1}\epsilon \|u\|_{H^1(\Omega)}^2. \tag{2.8}$$

□

2.1 Weak Solution

Definition 2.6 (Abstract function spaces)**.** Let $(X, \|\cdot\|)$ be a real Banach space and $T > 0$.

(i) For $p \in [1, \infty]$ we denote by $L^p(0,T;X)$ the Banach space of Bochner measurable functions $f : [0,T] \to X$ such that

$$\|f\|_{L^p(0,T;X)}^p := \int_0^T \|f(t)\|^p \, dt < \infty \tag{2.9}$$

for $p \in [1, \infty)$ and

$$\|f\|_{L^\infty(0,T;X)} := \operatorname*{ess\,sup}_{[0,T]} \|f\| < \infty. \tag{2.10}$$

(ii) The space $C([0,T];X)$ is the space of continuous functions $f : [0,T] \to X$ with the norm

$$\|f\|_{C[0,T];X)} := \max_{[0,T]} \|f\| < \infty. \tag{2.11}$$

△

Definition 2.7 (Weak derivatives of abstract functions)**.** Let X, T be as before. For $u \in L^1(0,T;X)$ we say that $v \in L^1(0,T;X)$ is a *weak time derivative* of u if

$$\int_0^T \varphi'(t)u(t) \, dt = -\int_0^T \varphi(t)v(t) \, dt \tag{2.12}$$

for all $\varphi \in C_0^\infty((0,T), \mathbb{R})$. We write $u' := v$, since by a fundamental lemma of calculus of variations type argument there exists at most one such weak time derivative. △

Note that – as usual – $L^p(0,T;X) \subset L^1(0,T;X)$ for $p \in [1,\infty]$ by Hölder's inequality, since $[0,T]$ has finite Lebesgue measure in $\mathbb{R}$.

Definition 2.8. For X, T as before, the Sobolev space $W^{1,p}(0,T;X)$ is defined as

$$W^{1,p}(0,T;X) := \{v \in L^p(0,T;X) : v' \text{ exists and } v' \in L^p(0,T;X)\} \tag{2.13}$$

with norm

$$\|v\|_{W^{1,p}(0,T;X)} := \begin{cases} \left(\int_0^T \|v(t)\|^p + \|v'(t)\|^p \ dt\right)^{1/p} & \text{for } p \in [1,\infty) \\ \operatorname{ess\,sup}_{[0,T]}(\|v\| + \|v'\|) & \text{for } p = \infty. \end{cases} \tag{2.14}$$

$H^1(0,T;X)$ is used as an abbreviation for $W^{1,2}(0,T;X)$ and hints at the fact that this is a Hilbert space. △

Remark 2.9. We will encounter the case $u \in L^2(0,T;H_0^1(\Omega))$ and $u' \in L^2(0,T;H^{-1}(\Omega))$. Definition 2.7 does not immediately give sense to this combination of expressions, since $H_0^1(\Omega)$ and its dual $H^{-1}(\Omega)$ are different spaces. To make sense of this, we need to regard $H_0^1(\Omega)$ as a subspace of $H^{-1}(\Omega)$ by means of the embedding $\iota_2 \circ \iota_1$, where

$$H_0^1(\Omega) \underset{\iota_1}{\hookrightarrow} L^2(\Omega) \underset{\iota_2}{\hookrightarrow} H^{-1}(\Omega). \tag{2.15}$$

Here, ι_1 is the inclusion and ι_2 maps $f \in L^2(\Omega)$ to the functional $v \mapsto \int_\Omega fv \ dx$ in $H^{-1}(\Omega)$. Note that ι_2 is indeed injective, because $H_0^1(\Omega)$ is dense in $L^2(\Omega)$. △

Theorem 2.10 ([Eva10, Theorem 5.9.2.3])**.** *Let $u \in L^2(0,T;H_0^1(\Omega))$ and $u' \in L^2(0,T;H^{-1}(\Omega))$. Then*

(i) $u \in C([0,T];L^2(\Omega))$ and

(ii) $t \mapsto \|u(t)\|_{L^2(\Omega)}^2$ is absolutely continuous and differentiable a.e. on $[0,T]$ with

$$\frac{d}{dt}\|u(t)\|_{L^2(\Omega)}^2 = 2\langle u'(t), u(t)\rangle. \tag{2.16}$$

Here, $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_{X^*\times X}$ denotes the duality pairing of a Banach space X, in this case $X = H_0^1(\Omega)$ and $X^* = H^{-1}(\Omega)$. △

Lemma 2.11 (Characterisation of weak time derivatives)**.** *Let X be a Banach space and $T > 0$. For two functions $u, v \in L^1(0,T;X)$ the following assertions are equivalent:*

(i) $v = u'$.

(ii) $t \mapsto \langle f, v(t)\rangle$ is the weak time derivative of $t \mapsto \langle f, u(t)\rangle$ for all $f \in X^*$. △

Proof. Let $\varphi \in C_0^\infty((0,T),\mathbb{R})$. Then

$$\int_0^T \varphi'(t)u(t) \ dt = -\int_0^T \varphi(t)v(t) \ dt \tag{2.17}$$

is equivalent to

$$\langle f, \int_0^T \varphi'(t)u(t)\ dt\rangle = \langle f, -\int_0^T \varphi(t)v(t)\ dt\rangle \quad \forall\ f \in X^* \tag{2.18}$$

by Hahn-Banach's theorem. But bounded linear functionals can be shifted under the Bochner integral, so we obtain equivalence of (2.17) to

$$\int_0^T \varphi'(t)\langle f, u(t)\rangle\ dt = -\int_0^T \varphi(t)\langle f, v(t)\rangle\ dt \quad \text{for all} \quad f \in X^*. \tag{2.19}$$

□

We now follow [Eva10, page 373] with some added details in devising a weak formulation for (2.1). Let us proceed formally by supposing that this problem has a smooth solution. Then we can multiply the differential equation by a test function $v \in H_0^1(\Omega)$, integrate over Ω and apply integration by parts to obtain

$$(u_t, v) + a(u, v) = (f, v) \tag{2.20}$$

with the bilinear form a from Definition 2.3.

Regarding u and f as abstract functions $u : [0, T] \to H_0^1(\Omega)$ and $f : [0, T] \to L^2(\Omega)$ and abusing notation slightly, let us argue that we can equate $(u'(t))(x) = (u_t(t))(x) := u_t(t, x)$ for almost every $x \in \Omega$. Because of the assumed smoothness of the solution, we have

$$u, u_t \in L^1(0, T; L^2(\Omega))$$

and therefore also

$$\int_0^T u_t(t)\varphi(t)\ dt - \int_0^T u(t)\varphi'(t)\ dt \in L^2(\Omega) \tag{2.21}$$

with arbitrary $\varphi \in C_0^\infty((0, T), \mathbb{R})$. For $x \in \Omega$, let $B := B_\varepsilon(x)$ be a ball around x entirely contained in Ω. Then $\chi_B \in L^2(\Omega)$ defines a linear functional on $L^2(\Omega)$ via

$$v \mapsto (\chi_B, v) = \int_B v\ dx \tag{2.22}$$

and we obtain

$$\begin{aligned}
\left(\int_0^T u_t(t)\varphi(t)\ dt\right)(x) &= \lim_{\varepsilon\to 0}\left(\frac{\chi_{B_\varepsilon(x)}}{|\chi_{B_\varepsilon(x)}|}, \int_0^T u_t(t)\varphi(t)\ dt\right) = \lim_{\varepsilon\to 0}\int_0^T \left(\frac{\chi_{B_\varepsilon(x)}}{|\chi_{B_\varepsilon(x)}|}, u_t(t)\right)\varphi(t)\ dt \\
&= \int_0^T \lim_{\varepsilon\to 0}\left(\frac{\chi_{B_\varepsilon(x)}}{|\chi_{B_\varepsilon(x)}|}, u_t(t)\right)\varphi(t)\ dt = \int_0^T u_t(x, t)\varphi(t)\ dt \\
&= -\int_0^T u(x, t)\varphi'(t)\ dt = -\lim_{\varepsilon\to 0}\int_0^T \left(\frac{\chi_{B_\varepsilon(x)}}{|\chi_{B_\varepsilon(x)}|}, u(x, t)\right)\varphi'(t)\ dt \\
&= \left(-\int_0^T u(t)\varphi'(t)\ dt\right)(x)
\end{aligned} \tag{2.23}$$

for almost every $x \in \Omega$. Hence we can replace u_t by u' in (2.20).

As is made plausible in [Eva10], a weak solution should be sought such that $u'(t) \in H^{-1}(\Omega)$ for almost all $t \in [0,T]$, which is why we we replace (u', v) by $\langle u', v\rangle$. Our weak formulation of (2.1) is therefore to find $u \in L^2(0,T;H_0^1(\Omega))$ with $u' \in L^2(0,T;H^{-1}(\Omega))$ satisfying

$$\begin{cases} \langle u'(t), v\rangle + a(u(t), v) = (f(t), v) & \text{for all } v \in H_0^1(\Omega), \text{ for a.e. } t \in [0,T] \\ u(0) = u_0. \end{cases} \tag{2.24}$$

Since $H_0^1(\Omega)$ is reflexive, this is equivalent to finding $u \in L^2(0,T;H_0^1(\Omega))$ with $u' \in L^2(0,T;H^{-1}(\Omega))$ satisfying

$$\begin{cases} \langle u(t), v\rangle' + a(u(t), v) = (f(t), v) & \text{for all } v \in H_0^1(\Omega), \text{ for a.e. } t \in [0,T] \\ u(0) = u_0 \end{cases} \tag{2.25}$$

(due to Lemma 2.11), where $\cdot'$ denotes the weak time derivative of the real-valued function $t \mapsto \langle u(t), v\rangle$; but then Theorem 2.10 allows us to replace the duality pairing by the L^2 inner product:

Find $u \in L^2(0,T;H_0^1(\Omega))$ with $u' \in L^2(0,T;H^{-1}(\Omega))$ such that

$$\begin{cases} (u(t), v)' + a(u(t), v) = (f(t), v) & \text{for all } v \in H_0^1(\Omega), \text{ for a.e. } t \in [0,T] \\ u(0) = u_0. \end{cases} \tag{2.26}$$

Remark 2.12. The condition $u(0) = u_0$ makes sense if $u \in L^2(0,T;H_0^1(\Omega))$, $u' \in L^2(0,T;H^{-1}(\Omega))$, since then we have by Theorem 2.10 that u is uniformly continuous on $[0,T]$ when understood as an $L^2(\Omega)$-valued function. △

Now we are in a position to prove existence and uniqueness for problem (2.26). This is the content of [QV94, Theorem 11.1.1], whose proof we follow while also adding details.

Theorem 2.13 (Existence, uniqueness and energy estimate)**.** *Let Assumption 2.1 hold. Then there exists a unique solution $u \in C([0,T];L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ with $u' \in L^2(0,T;H^{-1}(\Omega))$ to the weak problem* (2.26) *and the energy estimate*

$$\|u(t)\|_{L^2(\Omega)}^2 + c_a \int_0^t \|u(s)\|_{H^1(\Omega)}^2 \, dt \le \|u_0\|_{L^2(\Omega)}^2 + \frac{1}{c_a} \int_0^t \|f(s)\|_{L^2(\Omega)}^2 \, ds \tag{2.27}$$

holds for all $t \in [0,T]$. △

Proof. The proof's core is a semi-discretisation in space by the so called Faedo-Galerkin method. Let $(\phi_j)_{j\in\mathbb{N}}$ be a sequence in $H_0^1(\Omega)$ forming an orthonormal basis with respect to the $H^1(\Omega)$ scalar product $\langle\cdot,\cdot\rangle_{H^1(\Omega)} = (\cdot,\cdot) + (\nabla\cdot, \nabla\cdot)$ and set $V^N := \operatorname{span}\{\phi_1, \dots, \phi_N\}$. Then we consider the finite-dimensional evolution problem:

Find $u^N : [0,T] \to V^N$ such that

$$\begin{cases} (u^N(t), \phi_j)' + a(u^N(t), \phi_j) = (f(t), \phi_j) & \text{for } j \in \{1, \dots, N\}, \text{ for a.e. } t \in (0,T) \\ u^N(0) = u_0^N := \Pi_N(u_0), \end{cases} \tag{2.28}$$

where $\Pi_N : H_0^1(\Omega) \to V^N$ is the orthogonal projection onto V^N with respect to the L^2 inner product $(\cdot,\cdot)$. We introduce the mass matrix M, stiffness matrix A, right-hand side $F(t)$, initial value c_0^N and unknown vector c^N:

$$\begin{aligned} m_{ij} &:= (\phi_j, \phi_i) && \text{for } i,j \in \{1,\dots,N\} \\ a_{ij} &:= a(\phi_j, \phi_i) && \text{for } i,j \in \{1,\dots,N\} \\ F_i(t) &:= (f(t), \phi_i) && \text{for } i \in \{1,\dots,N\} \\ \tilde{c}_{0,i}^N &:= (\Pi_N(u_0), \phi_i) = (u_0, \phi_i) && \text{for } i \in \{1,\dots,N\} \\ c_i^N(t) &:= \langle u^N(t), \phi_i \rangle_{H^1(\Omega)} && \text{for } i \in \{1,\dots,N\}. \end{aligned} \tag{2.29}$$

Given that M is symmetric and positive definite, (2.28) can be equivalently written as

$$\begin{cases} (c^N)'(t) = M^{-1}\left(F(t) - Ac^N(t)\right) & \text{for a.e. } t \in (0,T) \\ c^N(0) = c_0^N := M^{-1}\tilde{c}_0^N. \end{cases} \tag{2.30}$$

We see that this is a finite-dimensional ordinary differential equation with a right hand side affine in c^N but not necessarily continuous in t.

Let us assert the conditions for the application of Carathéodory's local existence and uniqueness theorem (Theorem B.3). Set $\hat{F}(t) := M^{-1}F(t)$, $\hat{A} := M^{-1}A$.

(i) The function $g(t, c^N) := \hat{F}(t) - \hat{A}c^N$ is defined on $[0,T] \times \mathbb{R}^N$, measurable in t for each fixed $c^N \in \mathbb{R}^N$ and Lipschitz continuous in c^N with the time-independent Lipschitz constant $|\hat{A}|$ for each fixed $t \in [0,T]$.

(ii) For $R \in \mathbb{R}_{>0}$ arbitrarily large, it holds that

$$|g(t, c^N)| \le H(t) := |\hat{F}(t)| + |\hat{A}|R \quad \text{for all } c^N \in B_R(0)$$

and $H \in L^1([0,T])$:

$$\begin{aligned} \left(\int_0^T |\hat{F}(t)|\, dt\right)^2 &\le T\int_0^T |\hat{F}(t)|^2\, dt \le T|M^{-1}|^2 \int_0^T |F(t)|^2\, dt = T|M^{-1}|^2 \int_0^T \sum_{i=1}^N (f(t),\phi_i)^2\, dt \\ &\le T|M^{-1}|^2 \sum_{i=1}^N \|\phi_i\|_{L^2(\Omega)}^2 \underbrace{\int_0^T \|f(t)\|_{L^2(\Omega)}^2\, dt}_{=\|f\|_{L^2(\Omega_T)}^2} < \infty. \end{aligned} \tag{2.31}$$

By Theorem B.3 there exists $\delta > 0$ and a unique absolutely continuous function $c^N : [0,\delta] \to \mathbb{R}^N$ satisfying

$$c^N(t) = c_0^N + \int_0^t g(s, c^N(s))\, ds. \tag{2.32}$$

We need the solution to exists globally on $[0,T]$. If the maximally extended solution were only defined on $[0,T')$ for $T' < T$, then $c^N(t)$ would have to approach the boundary of $B_R(0)$. Remember

that we can choose $R > 0$ arbitrarily large; it suffices therefore to show that the maximally extended solution remains bounded uniformly with respect to $T' < T$. We have

$$c^N(t) = c_0^N + \int_0^t \hat{F}(t) - \hat{A}c^N(t)\ dt = c_0^N + \int_0^t \hat{F}(t)\ dt - \hat{A}\int_0^t c^N(t)\ dt$$

and therefore

$$|c^N(t)| \leq \underbrace{|c_0^N| + \int_0^t |\hat{F}(t)|\ dt}_{=:h(t)} + |\hat{A}| \int_0^t |c^N(t)|\ dt.$$

The function h is monotonically increasing, so Gronwall's lemma in integral form (Theorem B.1) yields

$$|c^N(t)| \leq \exp(t|\hat{A}|)\left(|c_0^N| + \int_0^t |\hat{F}|\ ds\right) \leq \exp(T|\hat{A}|)\left(|c_0^N| + \int_0^T |\hat{F}|\ ds\right) < \infty.$$

This is the uniform bound. Remebering that the solution (2.32), which has just been shown to exist on $[0, T)$, is both weakly and almost everywhere strongly differentiable with $(c^N)'(t) = g(t, c^N(t))$, we see that (2.30) is solved.
The boundedness of c^N and the fact that $F \in L^2(0, T; \mathbb{R}^N)$ (see (2.31)) gives $c^N \in H^1(0, T; \mathbb{R}^N)$ and, by H^1 orthogonality of the basis $(\phi_1, \ldots, \phi_N)$, that $u^N \in H^1(0, T; H_0^1(\Omega))$ with $\left\|u^N\right\|_{H^1(0,T;H_0^1(\Omega))} = \left\|c^N\right\|_{H^1(0,T;\mathbb{R}^N)} < \infty$. For any fixed $t \in (0, T)$ we can now test with $v := u^N(t)$:

$$\left((u^N)'(t), u^N(t)\right) + a(u^N(t), u^N(t)) = (f(t), u^N(t)) \tag{2.33}$$

and apply Theorem 2.10, coercivity of a, Hölder's and Young's inequalities to obtain for almost every $t \in [0, T]$:

$$\frac{1}{2}\frac{d}{dt}\left\|u^N(t)\right\|_{L^2(\Omega)}^2 + c_a \left\|u^N(t)\right\|_{H^1(\Omega)}^2 \leq \|f(t)\|_{L^2(\Omega)} \left\|u^N(t)\right\|_{L^2(\Omega)} \leq \frac{1}{2c_a}\|f(t)\|_{L^2(\Omega)}^2 + \frac{c_a}{2}\left\|u^N(t)\right\|_{L^2(\Omega)}^2. \tag{2.34}$$

Remember that $\|\cdot\|_{H^1(\Omega)} \geq \|\cdot\|_{L^2(\Omega)}$ and multiply by 2:

$$\frac{d}{dt}\left\|u^N(t)\right\|_{L^2(\Omega)}^2 + c_a \left\|u^N\right\|_{H^1(\Omega)}^2 \leq \frac{1}{c_a}\|f(t)\|_{L^2(\Omega)}^2. \tag{2.35}$$

Recall further that, by Theorem 2.10, the function $t \mapsto \left\|u^N(t)\right\|_{L^2(\Omega)}^2$ is absolutely continuous. Therefore we may integrate over $(0, \tau), \tau \in (0, T]$ and apply the second fundamental theorem of calculus for absolutely continuous functions in order to obtain the energy estimate (2.27) for the Galerkin solution:

$$\left\|u^N(\tau)\right\|_{L^2(\Omega)}^2 + c_a \int_0^\tau \left\|u^N(t)\right\|_{H^1(\Omega)}^2\ dt \leq \underbrace{\left\|u_0^N\right\|_{L^2(\Omega)}^2}_{\leq \|u_0\|_{L^2(\Omega)}^2} + \frac{1}{c_a}\int_0^\tau \|f(t)\|_{L^2(\Omega)}^2\ dt, \tag{2.36}$$

which tells us that $(u^N)_{N\in\mathbb{N}}$ is a bounded sequence in $L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$.

Now we make use of the fact that $L^2(0, T; H_0^1(\Omega))$ and $L^2(0, T; H^{-1}(\Omega))$ are Hilbert spaces and therefore reflexive and that $L^\infty(0, T; L^2(\Omega))$ is (via the usual isomorphism) isomorphic to the dual

space of $L^1(0,T;L^2(\Omega))$, which is a separable Banach space. It follows that there exists a subsequence of $(u^N)_{N\in\mathbb{N}}$ (we do not denote the fact that it is a subsequence) and

$$u \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega)) \tag{2.37}$$

with

$$u^N \overset{*}{\rightharpoonup} u \text{ in } L^\infty(0,T;L^2(\Omega)) \qquad \text{and} \qquad u^N \rightharpoonup u \text{ in } L^2(0,T;H_0^1(\Omega)), \tag{2.38}$$

where $\overset{*}{\rightharpoonup}$ and $\rightharpoonup$ denote weak* and weak convergence, respectively.

Fix $j \in \mathbb{N}$ and multiply for any $N \geq j$ the first line of (2.28) by some $\Psi \in C^1([0,T])$ with $\Psi(T) = 0$, integrate over $[0,T]$ and perform integration by parts to see

$$-\int_0^T (u^N(t),\phi_j)\Psi'(t)\ dt - (u_0^N,\phi_j)\Psi(0) + \int_0^T a(u^N(t),\phi_j)\Psi(t)\ dt = \int_0^T (f(t),\phi_j)\Psi(t)\ dt. \tag{2.39}$$

Note that $\phi_j\Psi' \in L^1(0,T;L^2(\Omega))$ and $a(\cdot,\phi_j\Psi) \in L^2(0,T;H^{-1}(\Omega))$ by the boundedness of a and remeber that u_0^N was defined to be the L^2 projection onto V^N. We thus obtain by letting $N \to \infty$:

$$-\int_0^T (u(t),\phi_j)\Psi'(t)\ dt - (u_0,\phi_j)\Psi(0) + \int_0^T a(u(t),\phi_j)\Psi(t)\ dt = \int_0^T (f(t),\phi_j)\Psi(t)\ dt. \tag{2.40}$$

Since $j \in \mathbb{N}$ can be chosen arbitrarily large and $\operatorname{span}\{\phi_j : j \in \mathbb{N}\}$ is dense in $H_0^1(\Omega)$, we can extend the result to arbitrary $v \in H_0^1(\Omega)$:

$$-\int_0^T (u(t),v)\Psi'(t)\ dt - (u_0,v)\Psi(0) + \int_0^T a(u(t),v)\Psi(t)\ dt = \int_0^T (f(t),v)\Psi(t)\ dt, \tag{2.41}$$

which gives us the first line of (2.26) by testing with all $\Psi \in C_0^\infty((0,T))$. In addition, notice that $u'(t) = f(t) - a(u(t),\cdot) \in L^2(0,T;H^{-1}(\Omega))$, so that the regularity requirement of a weak solution is met, too.

It remains to prove that the initial value u_0 is assumed in the L^2 sense to finish the existence proof. To this end, repeat the integration by parts argument with u instead of u^N and require in addition that $\Psi(0) = 1$:

$$-\int_0^T (u(t),v)\Psi'(t)\ dt - (u(0),v) + \int_0^T a(u(t),v)\Psi(t)\ dt = \int_0^T (f(t),v)\Psi(t)\ dt \tag{2.42}$$

for all $v \in H_0^1(\Omega)$. Comparing with (2.41) gives

$$(u(0),v) = (u_0,v) \quad \text{for all } v \in H_0^1(\Omega) \tag{2.43}$$

and therefore $u(0) = u_0$ owing to the density of $H_0^1(\Omega)$ in $L^2(\Omega)$.

The energy estimate follows by testing with $v = u(t)$ and applying Theorem 2.10:

$$\underbrace{\langle u'(t),u(t)\rangle}_{=\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)}} + a(u(t),u(t)) = (f(t),u(t)) \quad \text{for a.e. } t \in [0,T] \tag{2.44}$$

and then proceeding analogously as from (2.33) to (2.36).

Finally, we show uniqueness. Set $f = 0$ and $u_0 = 0$; then we only need to show that $u = 0$ is the unique solution in this case. But from the energy estimate

$$\frac{d}{dt} \|u(t)\|^2_{L^2(\Omega)} \leq 0 \tag{2.45}$$

and Gronwall's lemma for absolutely continuous functions in differential form (Theorem B.2) it follows that $\|u(t)\|^2_{L^2(\Omega)} = \|u(t)\|^2_{L^2(\Omega_T)} = 0$ for all $t > 0$; hence $u = 0$ almost everywhere on Ω_T. □

By merely introducing the additional condition $u_0 \in H^1_0(\Omega)$ to ensure some compatibility of the data on $\partial\Omega \times [0,T]$, a higher regularity of the weak solution can be shown.

Assumption 2.14. Let Assumption 2.1 and $u_0 \in H^1_0(\Omega)$ hold. △

Proposition 2.15 ([QV94, Proposition 11.1.1])**.** *Let Assumption 2.14 hold. Then the weak solution to problem* (2.26) *belongs to* $L^\infty(0,T;H^1_0(\Omega)) \cap H^1(0,T;L^2(\Omega))$ *with the energy estimate*

$$\operatorname*{ess\,sup}_{t\in[0,T]} \|u(t)\|^2_{H^1(\Omega)} + \int_0^T \left\|u'(t)\right\|^2_{L^2(\Omega)} \, dt \leq C \left(\|u_0\|^2_{H^1(\Omega)} + \int_0^T \|f(t)\|^2_{L^2(\Omega)} \, dt \right), \tag{2.46}$$

$C = C(c_a)$ *being a constant independent of* T. △

2.2 The Weak Maximum Principle

The weak maximum principle for classical solutions of the convection-diffusion equation bounds the solution on Ω_T by its initial-boundary values on $\Gamma_T := \overline{\Omega_T} \setminus \Omega_T = \Omega \times \{0\} \cup \partial\Omega \times [0,T]$ and will play an important role in the following chapters.

Theorem 2.16 (Parabolic weak maximum principle, [Eva10, Theorem 7.1.4.8])**.** *Let* $\Omega \subset \mathbb{R}^d$ *be a domain and*

$$Lu := -\sum_{i,j=1}^{d} a_{ij} u_{x_i x_j} + \sum_{i=1}^{d} b_i u_{x_i} \tag{2.47}$$

with the coefficients continuous on $\overline{\Omega_T}$ *and* $A := (a_{ij})_{i,j=1,\dots,d}$ *satisfying an ellipticity (i.e. symmetric positive definiteness) property uniformly in* (x,t)*. Let* $u \in C^2_1(\Omega_T) \cap C(\overline{\Omega_T})$ *(twice (once) continuously differentiable in space (time)).*

(i) If $u_t + Lu \leq 0$ *on* Ω_T*, then* $\max_{\overline{\Omega_T}} u = \max_{\Gamma_T} u$*.*

(ii) If $u_t + Lu \geq 0$ *on* Ω_T*, then* $\min_{\overline{\Omega_T}} u = \min_{\Gamma_T} u$*.* △

The proof relies strongly on the proof of the weak maximum principle for the elliptic case. Because we need a result from the proof of this latter assertion, we state it here including its proof from [Eva10, Theorem 6.4.1.1].

Theorem 2.17 (Elliptic weak maximum principle)**.** *Let* L *be unchanged and* $u \in C^2(\Omega) \cap C(\overline{\Omega})$*.*

(i) If $Lu \le 0$ in Ω, then $\max_{\overline{\Omega}} u = \max_{\partial\Omega} u$.

(ii) If $Lu \ge 0$ in Ω, then $\min_{\overline{\Omega}} u = \min_{\partial\Omega} u$. △

Proof. 1.) The important point is that, wherever u attains a local maximum in some point $x_0 \in \Omega$, there holds $Lu(x_0) \ge 0$. Indeed, the smoothness assumptions on u imply that the first derivative vanishes and the Hessian form is negative semi-definite:

$$Du(x_0) = 0 \tag{2.48}$$

$$D^2u(x_0) \preceq 0. \tag{2.49}$$

The matrix $A := (a_{ij}(x_0))_{i,j=1,\dots,d}$ is symmetric and positive definite, hence we find an orthogonal matrix O such that

$$OAO^T = \operatorname{diag}(d_1, \dots, d_d) \tag{2.50}$$

with $d_1, \dots, d_d > 0$ and we can use the affine transformation of variables $y = x_0 + O(x - x_0)$, or $x - x_0 = O^T(y - x_0)$. With this transformation, it holds that

$$u_{x_i} = \sum_{k=1}^{d} u_{y_k} \frac{dx_i}{dy_k} = \sum_{k=1}^{d} u_{y_k} o_{ki} \qquad \text{and} \qquad u_{x_i x_j} = \sum_{k,l=1}^{d} u_{y_k y_l} o_{ki} o_{lj} \tag{2.51}$$

and therefore

$$\sum_{i,j=1}^{d} a_{ij} u_{x_i x_j} = \sum_{k,l=1}^{d} \sum_{i,j=1}^{d} a_{ij} u_{y_k y_l} o_{ki} o_{lj} = \sum_{k=1}^{d} d_k u_{y_k y_k} \le 0. \tag{2.52}$$

The vanishing first derivative at x_0 then yields $Lu(x_0) \ge 0$.

2.) The argument just made shows that, if the strict inequality $Lu < 0$ holds on Ω, then local maxima inside Ω are impossible and therefore part (i) of the theorem holds in this case.

3.) If only $Lu \le 0$, define for $\lambda, \varepsilon > 0$

$$u_\varepsilon(x) := u(x) + \varepsilon e^{\lambda x_1} \quad \text{for } x \in \Omega. \tag{2.53}$$

Then, with a uniform ellipticity constant $\theta > 0$,

$$Lu_\varepsilon = Lu + \varepsilon L(e^{\lambda x_1}) \le \varepsilon e^{\lambda x_1}(-\lambda^2 a_{11} + \lambda b_1) \le \varepsilon e^{\lambda x_1}(-\lambda^2\theta + \lambda \|b\|_{L^\infty(\Omega)}) < 0 \tag{2.54}$$

for large $\lambda > 0$, since then the term in brackets is negative independently of ε. Letting $\varepsilon \to 0$ and using step 2.), we find that

$$\max_{\overline{\Omega}} u = \lim_{\varepsilon\to 0} \max_{\overline{\Omega}} u_\epsilon = \lim_{\varepsilon\to 0} \max_{\partial\Omega} u_\varepsilon = \max_{\partial\Omega} u \tag{2.55}$$

and part (i) is proven. Part (ii) follows by considering $-u$. □

Proposition 2.18 (Interior local extrema diminish)**.** *Let the operator L be as in Theorem 2.16, $u \in C_1^2(\Omega_T) \cap C(\overline{\Omega_T})$, $f = u_t + Lu$ and $(x_0, t_0) \in \Omega \times (0, T)$.*

(i) If u has a local maximum with respect to x at (x_0, t_0) and if $f(x_0, t_0) \le 0$, then $u_t(x_0, t_0) \le 0$.

(ii) If u has a local minimum with respect to x at (x_0, t_0) and if $f(x_0, t_0) \geq 0$, then $u_t(x_0, t_0) \geq 0$.

The inequalities are strict if so are the local extrema or if $f(x_0, t_0) < 0$ (> 0). △

Proof. This simply follows from $u_t = f - Lu$, the assumption on $f(x_0, t_0)$ and $Lu(x_0, t_0) \geq 0$ at interior local maxima, which was shown in step 1.) of the proof of Theorem 2.17. The strictness assertion holds because the inequalities (2.49) and thus (2.52) are strict for strict local maxima. □

2.3 Semi-Discretisation in Space by Finite Elements

This section is based on Chapter 11.2 of [QV94]. For the entire section, let Assumption 2.14 hold. We revisit the Faedo-Galerkin idea used in the proof of Theorem 2.13 with a practical choice for the finite subspace, and instead of just passing to infinite dimension, we are interested in error estimates for finite dimension.

Let $V_h \subset H_0^1(\Omega)$ be an N-dimensional subspace with basis $(\varphi_1, \ldots, \varphi_N)$ and set $V := H_0^1(\Omega)$ for brevity. Then the semi-discrete Galerkin problem is

$$\begin{cases} (u_h'(t), v_h) + a(u_h(t), v_h) = (f(t), v_h) \quad \text{for all } v_h \in V_h, \text{ for a.e. } t \in (0, T) \\ u_h(0) = u_{0,h}, \end{cases} \tag{2.56}$$

where $u_{0,h} \in V_h$ is some approximation to $u_0 \in L^2(\Omega)$. Define the N-dimensional component vectors c and c_0 by

$$u_h(t) = \sum_{j=1}^{N} c_j(t)\varphi_j, \qquad\qquad u_{0,h} = \sum_{j=1}^{N} c_{0,j}\varphi_j \tag{2.57}$$

and define $M, A \in \mathbb{R}^{N \times N}$ and $F : [0, T] \to \mathbb{R}^N$ by

$$m_{ij} := (\varphi_j, \varphi_i), \qquad a_{ij} := a(\varphi_j, \varphi_i), \qquad F_i(t) := (f(t), \varphi_i). \tag{2.58}$$

Then (2.56) can once again be cast in the shape of an ordinary differential equation:

$$\begin{cases} c'(t) = M^{-1}\left(F(t) - Ac(t)\right) \quad \text{for a.e. } t \in (0, T) \\ c(0) = c_0 \end{cases} \tag{2.59}$$

and a unique global absolutely continuous solution in the sense of Carathéodory's theorem exists; similarly as in the proof of Theorem 2.13 (without the orthogonality of basis) we obtain $u_h \in H^1(0, T; V)$ and the energy estimate

$$\|u_h(\tau)\|_{L^2(\Omega)}^2 + c_a \int_0^\tau \|u_h(t)\|_{H^1(\Omega)}^2 \, dt \leq \|u_{h,0}\|_{L^2(\Omega)}^2 + \frac{1}{c_a} \int_0^\tau \|f(t)\|_{L^2(\Omega)}^2 \, dt, \tag{2.60}$$

which shows stability of the semi-discrete solution in the norms of $C([0, T]; L^2(\Omega))$ and $L^2(0, T; V)$.

Let us turn to the question of convergence of the semi-discrete method for the choice $V_h = V_{h,0}^k$ with $k \in \mathbb{N}$. The first proposition deals with the case $k = 1$, whereas the second proposition allows for arbitrary $k \in \mathbb{N}$ and gives a higher order of convergence even for $k = 1$ as long as higher regularity assumptions than Assumption 2.14 hold.

2.3.1 First Order Semi-Discrete Convergence

Lemma 2.19 ([QV94, Corollary 11.1.1])**.** *Assume that the solution u to (2.26) satisfies $u(t) \in H^2(\Omega)$ and*

$$\|u(t)\|^2_{H^2(\Omega)} \leq C(\|Lu(t)\|^2_{L^2(\Omega)} + \|u(t)\|^2_{H^1(\Omega)}) \tag{2.61}$$

for a.e. $t \in [0,T]$. Then $u \in L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega)) \cap C^0([0,T];V)$ and

$$\begin{aligned} \max_{t\in[0,T]} \|u(t)\|^2_{H^1(\Omega)} + \int_0^T \left(\left\|u'(t)\right\|^2_{L^2(\Omega)} + \|u(t)\|^2_{H^2(\Omega)}\right)\, dt \\ \leq C(c_a)\left(\|u_0\|^2_{H^1(\Omega)} + \int_0^T \|f(t)\|^2_{L^2(\Omega)}\, dt\right). \end{aligned} \tag{2.62}$$

△

Remark 2.20. (2.61) is guaranteed for any $d \in \mathbb{N}$ if Ω is a domain with C^2 boundary and for plane convex polygonal domains $\Omega \subset \mathbb{R}^2$. △

Proposition 2.21. *Let u be the solution to (2.26), $(\mathcal{T}_h)_h$ a shape-regular family of triangulations of $\Omega \subset \mathbb{R}^d$ for $d \leq 3$ and u_h the solution to (2.56) for the space $V_h := V^1_{h,0}$. Assume (2.61). Then*

$$\begin{aligned} &\|u(t) - u_h(t)\|^2_{L^2(\Omega)} + c_a \int_0^t \|u(\tau) - u_h(\tau)\|^2_{H^1(\Omega)}\, d\tau \\ &\leq \|u_0 - u_{0,h}\|^2_{L^2(\Omega)} + C(c_a, C_a)h^2\left(\|u_{0,h}\|^2_{H^1(\Omega)} + \|u_0\|^2_{H^1(\Omega)} + \int_0^t \|f(\tau)\|^2_{L^2(\Omega)}\, d\tau\right) \end{aligned} \tag{2.63}$$

for each $t \in [0,T]$. △

Proof. We take the proof of [QV94, Proposition 11.2.1] with some added details. Recall that we assume $u_0 \in V$ by requiring Assumption 2.14 to hold. Set $e(t) := u(t) - u_h(t)$. Then

$$(e'(t), v) + a(e(t), v) = 0 \quad \text{for all } v \in V_h \tag{2.64}$$

holds for almost every $t \in [0,T]$. For any such t, we choose $v(t) := u_h(t) - w(t)$ for $w(t) \in V_h$ to be defined momentarily. Then for each $\varepsilon > 0$ we find using Young's inequality in the last step that

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}(e(t), e(t)) + a(e(t), e(t)) &= \big(e'(t), u(t) - w(t)\big) + a(e(t), u(t) - w(t)) \\ &\leq \left\|e'(t)\right\|_{L^2(\Omega)} \|u(t) - w(t)\|_{L^2(\Omega)} + C_a \|e(t)\|_{H^1(\Omega)} \|u(t) - w(t)\|_{H^1(\Omega)} \\ &\leq \left\|e'(t)\right\|_{L^2(\Omega)} \|u(t) - w(t)\|_{L^2(\Omega)} + \frac{C_a^2}{4\varepsilon} \|u(t) - w(t)\|^2_{H^1(\Omega)} + \varepsilon \|e(t)\|^2_{H^1(\Omega)} \end{aligned} \tag{2.65}$$

for a.e. $t \in [0,T]$. Assuming (2.61) and using Sobolev's embedding we see that $u(t) \in H^2(\Omega) \subset C(\overline{\Omega})$ for almost every $t \in [0,T]$ and that we may choose $w(t) \in V_h$ to be the piecewise linear nodal interpolator of $u(t)$. Employing Lemma A.3 leaves us with the estimate

$$\|u(t) - w(t)\|^2_{L^2(\Omega)} + h^2 \|u(t) - w(t)\|^2_{H^1(\Omega)} \leq Ch^4 \|u(t)\|^2_{H^2(\Omega)}. \tag{2.66}$$

Let us choose $\varepsilon = c_a/2$, multiply (2.65) by 2, subtract $c_a \|e(t)\|^2_{H^1(\Omega)}$ and use Young's inequality:

$$\begin{aligned}\frac{d}{dt}\|e(t)\|^2_{L^2(\Omega)} + c_a\|e(t)\|^2_{H^1(\Omega)} &\le \frac{d}{dt}\|e(t)\|^2_{L^2(\Omega)} + a(e(t),e(t)) \\ &\le C\left\|e'(t)\right\|_{L^2(\Omega)} h^2 \|u(t)\|_{H^2(\Omega)} + C\frac{C_a^2}{c_a}h^2\|u(t)\|^2_{H^2(\Omega)} \\ &\le C(c_a,C_a)\left(\left\|e'(t)\right\|^2_{L^2(\Omega)} + \|u(t)\|^2_{H^2(\Omega)}\right)h^2\end{aligned} \tag{2.67}$$

for almost every $t \in [0,T]$. Integrating this on $(0,t)$ gives

$$\begin{aligned}\|e(t)\|^2_{L^2(\Omega)} + c_a\int_0^t \|e(\tau)\|^2_{H^1(\Omega)}\ d\tau &\le \|u_0 - u_{0,h}\|^2_{L^2(\Omega)} \\ &+ C(c_a,C_a)h^2\int_0^t \left(\left\|u'(\tau)\right\|^2_{L^2(\Omega)} + \left\|u_h'(\tau)\right\|^2_{L^2(\Omega)} + \|u(\tau)\|^2_{H^2(\Omega)}\right)\ d\tau.\end{aligned} \tag{2.68}$$

Similarly to Proposition 2.15, one can show

$$\int_0^t \left\|u_h'(\tau)\right\|^2_{L^2(\Omega)}\ d\tau \le C(c_a)\left(\|u_{0,h}\|^2_{H^1(\Omega)} + \int_0^t \|f(\tau)\|^2_{L^2(\Omega)}\ d\tau\right), \tag{2.69}$$

which leads straight to (2.63) when combined with (2.62). □

2.3.2 Higher Order Semi-Discrete Convergence

A better order of convergence can be obtained when assuming higher regularity of the true solution. The proof of this claim builds on approximation properties of the "elliptic projection operator" which necessitate the property of *adjoint regularity*. Let us recall some results from the theory of Galerkin methods for elliptic equations.

Definition 2.22 (Adjoint regularity)**.** Let $\Omega \subset \mathbb{R}^d$ be a domain, $V := H_0^1(\Omega)$, $a : V \times V \to \mathbb{R}$ a bounded coercive bilinear form with constants C_a and c_a, respectively, and $f \in V^*$. Then we call the problem

$$u \in V : a(u,v) = \langle f, v\rangle \quad \text{for all } v \in V \tag{2.70}$$

adjoint regular if the solution $\varphi(r)$ of the adjoint problem

$$\varphi \in V : a(v,\varphi) = (r,v) \quad \text{for all } v \in V \tag{2.71}$$

lies in $H^2(\Omega)$ for all $r \in L^2(\Omega)$. △

Remark 2.23. Adjoint regularity is guaranteed for any $d \in \mathbb{N}$ if Ω is a domain with C^2 boundary and for plane convex polygonal domains $\Omega \subset \mathbb{R}^2$. △

Theorem 2.24 (Céa's Lemma)**.** *Let $\Omega \subset \mathbb{R}^d$ be a domain, $m \in \mathbb{N}$, $V := H_0^1(\Omega)$, $a : V \times V \to \mathbb{R}$ a bounded coervice bilinear form with constants C_a and c_a, respectively. Let furthermore $f \in V^*$ and V_h be a finite-dimensional subspace of V and u, u_h the solutions defined by*

$$u \in V : a(u,v) = \langle f, v\rangle \quad \textit{for all } v \in V \tag{2.72}$$

$$u_h \in V_h : a(u_h,v) = \langle f, v\rangle \quad \textit{for all } v \in V_h. \tag{2.73}$$

Then it holds that

$$\|u - u_h\|_V \leq \frac{C_a}{c_a} \inf_{v \in V_h} \|u - v\|_V . \tag{2.74}$$

△

Theorem 2.25 (Aubin-Nitsche trick, [QV94, Proposition 6.2.2])**.** *Let $\Omega \subset \mathbb{R}^d$ for $d \leq 3$ be triangulated by a shape-regular family of triangulations $(\mathcal{T}_h)_h$ and let problem (2.70) be adjoint regular and $u \in V \cap H^s(\Omega)$, $s \in \mathbb{N}$, its solution. Set $V_h := V_{0,h}^k$ and let $u_h \in V_h$ be the solution of the discrete problem*

$$a(u_h, v) = \langle f, v \rangle \quad \textit{for all } v \in V_h \tag{2.75}$$

satisfying

$$\|u - u_h\|_{H^1(\Omega)} \leq C_* h^l \|u\|_{H^{l+1}(\Omega)} , \tag{2.76}$$

for $l := \min(k, s-1)$. If $u \in H^s(\Omega)$ for $s \geq 2$, then

$$\|u - u_h\|_{L^2(\Omega)} \leq C h^{l+1} \|u\|_{H^{l+1}(\Omega)} . \tag{2.77}$$

△

Proof. We add some details to Quarteroni and Valli's proof. A duality argument is applied in order to rewrite the L^2 norm:

$$\begin{aligned} \|u - u_h\|_{L^2(\Omega)} &= \sup_{r \in L^2(\Omega)\setminus\{0\}} \frac{(r, u - u_h)}{\|r\|_{L^2(\Omega)}} = \sup_{r \in L^2(\Omega)\setminus\{0\}} \frac{a(u - u_h, \varphi(r))}{\|r\|_{L^2(\Omega)}} \\ &\leq \sup_{r \in L^2(\Omega)\setminus\{0\}} C_a \|u - u_h\|_{H^1(\Omega)} \frac{\|\varphi(r) - \psi(r)\|_{H^1(\Omega)}}{\|r\|_{L^2(\Omega)}}, \end{aligned} \tag{2.78}$$

using that Galerkin "orthogonality" gives $a(u-u_h, \varphi(r)) = a(u-u_h, \varphi(r)-\psi(r))$ for any $\psi(r) \in V_h$. Since $d \leq 3$, $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$ by Sobolev's embedding theorem; we may therefore take $\psi(r)$ to be the nodal interpolator of $\varphi(r)$ satisfying $\|\varphi(r) - \psi(r)\|_{H^1(\Omega)} \leq Ch \|\varphi(r)\|_{H^2(\Omega)}$. Now we only need that $\|\varphi(r)\|_{H^2(\Omega)} \leq C \|r\|_{L^2(\Omega)}$ (i.e. $\varphi : L^2(\Omega) \to H^2(\Omega)$ is bounded), for then we can infer

$$\|u - u_h\|_{L^2(\Omega)} \leq C C_a \|u - u_h\|_{H^1(\Omega)} h \leq C C_a C_* h^{l+1} \|u\|_{H^{l+1}(\Omega)} . \tag{2.79}$$

Let $(r_n, \varphi(r_n)) \to (r, \phi)$ in $L^2(\Omega) \times H^2(\Omega)$ as $n \to \infty$. Since $\varphi(r_n) \in V \cap H^2(\Omega)$, we obtain in particular $\|\varphi(r_n) - \phi\|_{H^1(\Omega)} \to 0$, implying for each $v \in V$:

$$a(v, \phi) = \lim_{n\to\infty} a(v, \varphi(r_n)) = \lim_{n\to\infty} (r_n, v) = (r, v), \tag{2.80}$$

which just states that $\phi = \varphi(r)$. The graph of φ in $L^2(\Omega) \times H^2(\Omega)$ is thus closed and φ is bounded by the closed graph theorem. □

Remark 2.26. The constant C_* appearing in Theorem 2.25 is typically derived from Céa's lemma in conjunction with general interpolation space results. Therefore, by (2.74), it will behave unfavourably (like ϵ^{-1}) for small diffusion $\epsilon > 0$. △

Proposition 2.27 ([QV94, Proposition 11.2.2], modified)**.** *Consider $\Omega \subset \mathbb{R}^d$ for $d \leq 3$, triangulated by the shape-regular family of triangulations $(\mathcal{T}_h)_h$. Assume that problem (2.26) is adjoint regular and let u be its solution. Moreover, assume $u_0 \in H^{k+1}(\Omega)$ and $u' \in L^1(0,T;H^{k+1}(\Omega))$ for $k \in \mathbb{N}_{\geq 1}$. Let u_h be the solution to (2.56) for the space $V_h = V_{h,0}^k$. Then*

$$\begin{aligned}\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq{}& \|u_0 - u_{0,h}\|_{L^2(\Omega)} \\ &+ Ch^{k+1}\left(\|u_0\|_{H^{k+1}(\Omega)} + \left(\int_0^t \left\|u'(\tau)\right\|_{H^{k+1}(\Omega)}^2 \, d\tau\right)^{1/2}\right)\end{aligned} \tag{2.81}$$

holds for all $t \in [0,T]$, where C is dependent in particular on ϵ and T and independent of h. △

Proof. We modify slightly Quarteroni and Valli's proof. For $v \in V = H_0^1(\Omega)$ define by $\pi(v)$ the "elliptic projection operator" defined as the unique solution to

$$\pi \in V_h : a(\pi, v_h) = a(v, v_h) \quad \text{for all } v_h \in V_h. \tag{2.82}$$

Then Céa's lemma for the functional $a(v,\cdot)$ and the Aubin-Nitsche trick yield

$$\|v - \pi(v)\|_{L^2(\Omega)} + h\,\|v - \pi(v)\|_{H^1(\Omega)} \leq Ch^{k+1}\|v\|_{H^{k+1}(\Omega)} \quad \text{for all } v \in H^{k+1}(\Omega) \tag{2.83}$$

with a constant C independent of v (but scaling like ϵ^{-1}). For all fixed $t \in [0,T]$ we decompose

$$u_h(t) - u(t) = w_1(t) + w_2(t) \tag{2.84}$$

with $w_1(t) := u_h(t) - \pi(u(t))$ and $w_2(t) := \pi(u(t)) - u(t)$. The error term $w_2(t)$ is then easily estimated using the results of Céa and Aubin-Nitsche:

$$\begin{aligned}\|w_2(t)\|_{L^2(\Omega)} &\leq Ch^{k+1}\,\|u(t)\|_{H^{k+1}(\Omega)} \\ &\leq Ch^{k+1}\left(\|u_0\|_{H^{k+1}(\Omega)} + \left(\int_0^t \left\|u'(\tau)\right\|_{H^{k+1}(\Omega)}^2 \, d\tau\right)^{1/2}\right).\end{aligned} \tag{2.85}$$

For $w_1(t)$, we have for each $v_h \in V_h$ that

$$\begin{aligned}\big(w_1'(t), v_h\big) + a(w_1(t), v_h) &= (u_h'(t), v_h) + a(u_h(t), v_h) - ((\pi \circ u)'(t), v_h) - a(\pi(u(t)), v_h) \\ &= (f(t), v_h) - a(u(t), v_h) - ((\pi \circ u)'(t), v_h) \\ &= (u'(t), v_h) - ((\pi \circ u)'(t), v_h) = (-w_2'(t), v_h).\end{aligned} \tag{2.86}$$

Setting $v_h := w_1(t)$ and employing coercivity of a gives

$$\frac{1}{2}\frac{d}{dt}\,\|w_1(t)\|_{L^2(\Omega)}^2 + c_a\,\|w_1(t)\|_{H^1(\Omega)}^2 \leq -(w_2'(t), w_1(t)). \tag{2.87}$$

Multiplying by 2, integration and Young's inequality give

$$\|w_1(t)\|_{L^2(\Omega)}^2 + 2c_a \int_0^t \|w_1(s)\|_{H^1(\Omega)}^2 \, ds \leq \|w_1(0)\|_{L^2(\Omega)}^2 + \int_0^t \frac{1}{8c_a}\left\|w_2'(s)\right\|_{L^2(\Omega)}^2 + 2c_a\,\|w_1(s)\|_{L^2(\Omega)}^2 \, ds \tag{2.88}$$

and thus

$$\|w_1(t)\|^2_{L^2(\Omega)} \leq \|w_1(0)\|^2_{L^2(\Omega)} + \frac{1}{8c_a}\int_0^t \left\|w_2'(s)\right\|^2_{L^2(\Omega)}\, ds \tag{2.89}$$

or

$$\|w_1(t)\|_{L^2(\Omega)} \leq \|w_1(0)\|_{L^2(\Omega)} + C\left(\int_0^t \left\|w_2'(s)\right\|^2_{L^2(\Omega)}\, ds\right)^{1/2}. \tag{2.90}$$

The time derivative commutes with π, hence

$$\left\|w_2'(s)\right\|_{L^2(\Omega)} = \left\|\pi(u'(s)) - u'(s)\right\|_{L^2(\Omega)} \leq Ch^{k+1}\left\|u'(s)\right\|_{H^{k+1}(\Omega)}. \tag{2.91}$$

Furthermore,

$$\|w_1(0)\|_{L^2(\Omega)} \leq \|u_{h,0} - u_0\|_{L^2(\Omega)} + \underbrace{\|u_0 - \pi(u(0))\|_{L^2(\Omega)}}_{\leq Ch^{k+1}\|u_0\|_{H^{k+1}(\Omega)}}. \tag{2.92}$$

Adding (2.85) and (2.90) and employing the last two relations yields the assertion. □

Remark 2.28. The previous two a priori convergence result and the fully discrete result from the next section are of limited practical use in our case of small $\epsilon > 0$ because (2.65) and the comment after (2.83) reveal that the term ϵ^{-1} is hidden in the constants. This can be interpreted in two ways: the a priori estimate could be far from being sharp or, if it is rather sharp, the errors due to oscillations arising in the standard Galerkin method manifest themselves in this estimate. △

2.4 Time-Discretisation by θ-Stepping

This section is based on Chapter 11.3 of [QV94]. Let $\theta \in [0,1]$. Then to obtain a computable problem, we need a method to discretise and numerically solve the ordinary differential equation posed by the semi-discrete problem (2.59). The simplest class of methods to accomplish this is the class of θ-stepping methods, in which the time-derivative is approximated by a simple forward-difference and the right-hand side is evaluated at the last known and the current time-step. These evaluations are weighted by $1-\theta$ and θ, respectively. Hence, the parameter θ is aptly called the *implicitness parameter* of the method. For $\theta = 0$, $\theta = 0.5$ and $\theta = 1$ the methods are called *forward* (or *explicit*) *Euler method*, *Crank-Nicolson method* and *backward* (or *implicit*) *Euler method*, respectively.

Let $0 = t_0 < t_1 < \cdots < t_{N_T} \leq T$ be points in time and $\tau_n := t_{n+1} - t_n$, $n = 1, \ldots, N_T - 1$. For simplicity we only consider constant time-step lengths, i.e. $\tau_n = \tau = \text{const.}$ and $t_n = n\tau$. The fully discretised problem is thus to find $u_h^n \in V_h$, $n = 1, \ldots, N_T$, satisfying

$$\begin{cases} \left(\dfrac{u_h^{n+1} - u_h^n}{\tau}, v_h\right) + a(\theta u_h^{n+1} + \overline{\theta} u_h^n, v_h) = (\theta f(t_{n+1}) + \overline{\theta} f(t_n), v_h) \quad \text{for all } v_h \in V_h \\ u_h^0 = u_{0,h} \end{cases} \tag{2.93}$$

for $n = 0, \ldots, N_T - 1$, where $\overline{\theta} := 1 - \theta$. Writing again

$$u_h^n = \sum_{j=1}^N c_j^n \varphi_j, \qquad\qquad u_{0,h} = \sum_{j=1}^N c_{0,j}\varphi_j \tag{2.94}$$

and defining $M_C, A \in \mathbb{R}^{N\times N}$ and $F^n \in \mathbb{R}^N$ by

$$m_{ij} := (\varphi_j, \varphi_i), \qquad a_{ij} := a(\varphi_j, \varphi_i), \qquad F_i^n := (f(t_n), \varphi_i) \tag{2.95}$$

for $i, j = 1, \dots, N$ and $n = 0, \dots, N_T - 1$, we have to solve

$$(M_C + \theta\tau A)c^{n+1} = (M_C - (1-\theta)\tau A)c^n + \theta F^{n+1} + (1-\theta)F^n \tag{2.96}$$

in order to obtain $c^{n+1} \in \mathbb{R}^N$ for $n = 0, \dots, N_T - 1$. The system matrix is positive definite because so are M_C and A.

Lemma 2.29 (Inverse inequality for piecewise polynomials, [QV94, Proposition 6.3.2]). *Let $(\mathcal{T}_h)_h$ be a shape-regular, quasi-uniform family of triangulations of the domain $\Omega \subset \mathbb{R}^d$ and $V_h := V_h^k \subset V$. Then there exists a constant $C_{inv} \in \mathbb{R}_{>0}$ such that*

$$\|\nabla v_h\|^2_{L^2(\Omega)} \le C_{inv} h^{-2} \|v_h\|^2_{L^2(\Omega)} \tag{2.97}$$

for all $v_h \in V_h$. △

Proposition 2.30 (Stability, [QV94, Theorem 11.3.1]). *Assume that the map $t \mapsto \|f\|_{L^2(\Omega)}$ is bounded on $[0,T]$ and that $(\mathcal{T}_h)_h$ is a shape-regular family of triangulations of Ω. For $\theta \in [0, 1/2)$ assume, in addition, that $(\mathcal{T}_h)_h$ is quasi-uniform and that the time-step restriction*

$$\frac{1 + C_{inv}}{h^2}\tau < \frac{2c_a}{(1-2\theta)C_a^2}, \tag{2.98}$$

holds, where C_{inv} is the constant from the inverse inequality in Lemma 2.29. Then u_h^n from (2.93) satisfies the stability relation

$$\|u_h^n\|_{L^2(\Omega)} \le C_\theta \left(\|u_{0,h}\|_{L^2(\Omega)} + \sup_{t\in[0,T]} \|f(t)\|_{L^2(\Omega)} \right) \quad \text{for } n = 0, 1, \dots, N_T, \tag{2.99}$$

where the constant $C_\theta > 0$ is a non-decreasing function of c_a^{-1}, C_a and T and is independent of N_T, τ and h. △

Theorem 2.31 (Convergence, [QV94, Theorem 11.3.2]). *Assume that $u_h'(0) \in L^2(\Omega)$, $f_t \in L^2(\Omega_T)$ and that $(\mathcal{T}_h)_h$ is shape-regular. For $\theta \in [0, 1/2)$, assume that $(\mathcal{T}_h)_h$ is quasi-uniform and that the time-step restriction (2.99) holds. Then the following error estimate holds for the semi-discrete solution $u_h : [0,T] \to V_h$ and the fully discrete solution $u_h^n \in V_h$, $n = 0, \dots, N_T$:*

$$\|u_h^n - u_h(t_n)\|_{L^2(\Omega)} \le C_\theta \tau \left(\left\|u_h'(0)\right\|^2_{L^2(\Omega)} + \int_0^T \|f_t(s)\|^2_{L^2(\Omega)}\, ds \right)^{1/2} \tag{2.100}$$

for $n = 0, \dots, N_T$. If $\theta = 1/2$ and $f_{tt} \in L^2(\Omega_T)$, $u_h''(0) \in L^2(\Omega)$, then

$$\|u_h^n - u_h(t_n)\|_{L^2(\Omega)} \le C(\tau)^2 \left(\left\|u_h''(0)\right\|^2_{L^2(\Omega)} + \int_0^T \|f_{tt}(s)\|^2_{L^2(\Omega)}\, ds \right)^{1/2} \tag{2.101}$$

for $n = 0, \dots, N_T$. The contants C_θ and C are non-decreasing functions of c_a^{-1}, C_a, T and are independent of N_T, τ and h. △

3. The Reduced Problem

Since we are dealing with the class of convection-diffusion-reaction equations where the diffusive part of the operator is dominated by its counterparts, it is interesting to study the relationship between our problem of interest

$$\begin{cases} (u_\epsilon)_t - \epsilon\Delta u_\epsilon + b\cdot\nabla u_\epsilon + cu_\epsilon = f & \text{in } \Omega_T \\ u_\epsilon = 0 & \text{on } \partial\Omega\times[0,T] \\ u_\epsilon = u_0 & \text{on } \Omega\times\{t=0\} \end{cases} \tag{3.1}$$

and the so-called *reduced problem*

$$\begin{cases} u_t + b\cdot\nabla u + cu = f & \text{in } \Omega_T \\ u = 0 & \text{on } \mathcal{G}_-\subset\partial\Omega\times[0,T] \\ u = u_0 & \text{on } \Omega\times\{t=0\} \end{cases} \tag{3.2}$$

that is obtained by dropping the diffusive part of the operator, i.e. by formally setting $\epsilon = 0$, and restricting the Dirichlet portion of the boundary to $\mathcal{G}_-$, cf. the following definition. Cancelling all second-order derivatives changes the nature of the problem from parabolic to hyperbolic and necessitates the above restriction of the Dirichlet boundary, as will become apparent now.

In this chapter we will make generous regularity assumptions (to assure the existence of smooth enough classical solutions and auxiliary functions) and prove the convergence result Theorem 3.15 on special subdomains of Ω_T using a parabolic maximum principle.

Definition 3.1 (Characteristics, exit times, exit locus)**.** Set $\tilde{b} := (b,1) \in W^{1,\infty}(\Omega_T,\mathbb{R}^{d+1})$. For $\overline{x} = (x,t) \in \Omega_T$ define the associated *characteristic* $\gamma_{\overline{x}}$ as the solution to the ordinary differential equation

$$\begin{cases} \gamma_{\overline{x}}(t) = \overline{x} \\ \gamma'_{\overline{x}}(s) = \tilde{b}(\gamma_{\overline{x}}(s)) \quad \text{for } s\in[\tau_-(\overline{x}),\tau_+(\overline{x})], \end{cases} \tag{3.3}$$

where the *forward* and *backward exit times* $\tau_\pm(\overline{x})$ are defined by

$$\tau_+(\overline{x}) := \sup\{\tau \geq t : \gamma_{\overline{x}}(s)\in\Omega\times(0,T) \text{ for } s\in(t,\tau)\} \tag{3.4}$$

$$\tau_-(\overline{x}) := \inf\{\tau \leq t : \gamma_{\overline{x}}(s)\in\Omega\times(0,T) \text{ for } s\in(\tau,t)\}. \tag{3.5}$$

(By "the characteristic of $\overline{x}$" we will also mean the image $\gamma_{\overline{x}}([\tau_-(\overline{x}), \tau_+(\overline{x})]$.)

Then the *forward* and *backward exit points* $\overline{x}_\pm$ of $\gamma_{\overline{x}}$ are defined as

$$\overline{x}_- := \gamma_{\overline{x}}(\tau_-(\overline{x})) \in \partial\Omega_T \tag{3.6}$$

$$\overline{x}_+ := \gamma_{\overline{x}}(\tau_+(\overline{x})) \in \partial\Omega_T. \tag{3.7}$$

$\mathcal{G}_-$, the *backward exit locus in* $\partial\Omega\times[0,T)$, is the subset of $\partial\Omega\times[0,T)$ that consists of all characteristic backward exit points:

$$\mathcal{G}_- := \{\overline{y} \in \partial\Omega \times [0,T) : \overline{y} = \overline{x}_- \text{ for some } \overline{x} \in \Omega_T\} \tag{3.8}$$

and the *backward exit locus* $\mathcal{G}$ is

$$\mathcal{G} := \{\overline{y} \in \Gamma_T : y = \overline{x}_- \text{ for some } \overline{x} \in \Omega_T\} = \mathcal{G}_- \cup \Omega \times \{0\}. \tag{3.9}$$

Then the definition of the functions $\overline{x} \mapsto \tau_+(\overline{x})$ and $\overline{x} \mapsto \overline{x}_+$, which are constant along characteristics, can be naturally extended to $\mathcal{G}$. △

Remark 3.2. Note that in our notation "exit" refers to the interior Ω_T° of the cylinder, not its closure. △

Proposition 3.3. *Suppose that, on top of Assumption 2.1, b, c, f and u_0 are regular enough to admit a classical solution to* (3.2) *and give sense to the following terms. Then for any $\overline{x} := (x,t) \in \Omega_T$, $u(\overline{x})$ can be obtained through the following steps:*

(i) Compute the associated characteristic $\gamma_{\overline{x}}$ and backwards exit time $\tau_-(\overline{x})$.

(ii) Solve the ordinary differential equation

$$\tilde{u}'(s) = \tilde{f}(s) - \tilde{c}(s)\tilde{u}(s) \tag{3.10}$$

on $[\tau_-(\overline{x}), t]$ with initial value $\tilde{u}_0 := u(\overline{x}_-)$, where $\tilde{f}(s) := f(\gamma_{\overline{x}}(s))$ and $\tilde{c}(s) := c(\gamma_{\overline{x}}(s))$.

(iii) Set $u(\gamma_{\overline{x}}(s)) = \tilde{u}(s)$ for $s \in [\tau_-(\overline{x}), t]$. △

Proof. It holds by construction that

$$\begin{aligned} u_t(x,t) + b(x,t)\cdot\nabla u(x,t) &= u_t(\gamma_{\overline{x}}(t)) + b(\gamma_{\overline{x}}(t))\cdot\nabla u(\gamma_{\overline{x}}(t)) = \frac{d}{ds}\bigg|_{s=t} u(\gamma_{\overline{x}}(s)) \\ &= \frac{d}{ds}\bigg|_{s=t} \tilde{u}(s) = f(x,t) - c(x,t)u(x,t). \end{aligned} \tag{3.11}$$

□

From the way solutions are constructed in Proposition 3.3 as solutions of ordinary differential equations along characteristics it follows that the reduced problem in the case of classical solutions is well-posed if values are only prescribed on $\mathcal{G}$ and are left unprescribed on the remaining part of Γ_T. It makes no physical sense to prescribe data on forward exit points.

Obviously, the method of characteristics also yields solutions for discontinuous initial boundary data and discontinuous f, as long as $b \in W^{1,\infty}(\Omega, \mathbb{R}^d)$ or – equivalently – b Lipschitz continuous on Ω is assumed. We shall ignore the seemingly delicate subject of

Definition 3.4 (Boundary decomposition)**.** The boundary $\partial\Omega$ of a C^1 domain $\Omega \subset \mathbb{R}^d$ can be decomposed into the three subsets

$$\Gamma_- := \{x \in \partial\Omega : b(x) \cdot n < 0\} \qquad \text{(the } \textit{inflow boundary of } \Omega\text{)} \tag{3.12}$$

$$\Gamma_0 := \{x \in \partial\Omega : b(x) \cdot n = 0\} \qquad \text{(the } \textit{parabolic boundary of } \Omega\text{)} \tag{3.13}$$

$$\Gamma_+ := \{x \in \partial\Omega : b(x) \cdot n > 0\} \qquad \text{(the } \textit{outflow boundary of } \Omega\text{)}, \tag{3.14}$$

where n is the outside unit normal of $\partial\Omega$. △

Remark 3.5. Obviously, $\Gamma_- \times [0,T) \subset \mathcal{G}$. △

Remark 3.6 (Layers)**.** For small $\epsilon > 0$, it is expected that the solution to (2.1) is in some sense close to the solution of the reduced problem, at least at some distance to $(\Gamma_0 \cup \Gamma_+) \times [0,T]$. Near $(\Gamma_0 \cup \Gamma_+) \times [0,T]$, however, steep gradients in the solution of the parabolic problem are to be expected due to the discrepancy between the values of the reduced problem obtained by integrating (3.10) along characteristics and the homogeneous Dirichlet boundary conditions. These regions of rapid change in the solution are called *boundary layers*. One distinguishes between *exponential* and *parabolic* boundary layers, which occur along $\Gamma_+ \times [0,T]$ and $\Gamma_0 \times [0,T]$, respectively.
The fact that the sequence of solutions of convection-diffusion-reaction equations as $\epsilon \downarrow 0$ does not converge in the L^∞ norm to the solution of the limiting operator ($\epsilon = 0$, no diffusion) makes (2.1) a so-called *singulary-perturbed* problem.
Interior layers are common to both the parabolic and the reduced problem and stem from discontinuities in the initial-boundary data. In the reduced case, these are propagated into the interior of Ω_T along the characteristics, while in the case of small non-vanishing diffusion these interior layers remain steep, but not discontinuous, being also subject to the smoothing effect of diffusion. △

3.1 Convergence to the Reduced Problem as $\epsilon \to 0$

This section is concerned with the adaptation of the proof of [GFLRT83, Theorem 4.3] from the stationary case to our time-dependent problem.

Let us for this section consider the situation that the data is smooth and compatible enough such that the solutions to the parabolic problem (3.1) and to the reduced problem (3.2) are classical solutions in $C^2(\overline{\mathcal{S}_\Sigma})$, where $\mathcal{S}_\Sigma$ is a subdomain of the cylinder to be defined shortly. In particular, no internal boundary layers occur. The previous remark tells us that the parabolic problem exhibits, however, boundary layers in the vicinity of $(\Gamma_0 \cup \Gamma_-) \times [0,T]$; as a consequence, a convergence

$$u_\epsilon \to u \quad \text{in } L^\infty(\mathcal{S}_\Sigma) \text{ as } \epsilon \to 0 \tag{3.15}$$

cannot be expected if $\mathcal{S}_\Sigma$ has accumulation points in $(\Gamma_0 \cup \Gamma_+) \times [0,T]$. Such a result can only be proved by further restricting to a suitable subset of $\mathcal{S}_\Sigma$.

Definition 3.7. For a subset $\Sigma \subset \mathcal{G}$ define its induced *characteristic tube* $\mathcal{S}_\Sigma$ by

$$\mathcal{S}_\Sigma := \{\gamma_{\overline{x}}(t) : \overline{x} = (x,t) \in \Sigma \text{ and } t < \tau < \tau_+(\overline{x})\} \subset \Omega \times (0,T) \tag{3.16}$$

and decompose its boundary into four disjoint parts:

$$\partial\mathcal{S}_\Sigma = \overline{\Sigma} \uplus (\underbrace{\partial\mathcal{S}_\Sigma \cap \Omega \times (0,T)}_{=:\mathcal{B}_1(\Sigma)} \uplus (\underbrace{\partial\mathcal{S}_\Sigma \cap (\partial\Omega \times (0,T]) \setminus \mathcal{G}}_{=:\mathcal{B}_2(\Sigma)} \uplus (\underbrace{\partial\mathcal{S}_\Sigma \cap \Omega \times \{t=T\}}_{=:\mathcal{B}_3(\Sigma)}. \tag{3.17}$$

△

The following two lemmata ensure that $\mathcal{S}_\Sigma$ is a domain under reasonable conditions on $\Sigma \subset \mathcal{G}$.

Lemma 3.8. *(i) The function $\overline{x} \mapsto \tau_-(\overline{x})$ is in $C(\Omega \times (0,T), \mathbb{R})$.*

(ii) $\overline{x} \mapsto \tau_+(\overline{x})$ is in $C(\Omega \times (0,T) \cup \mathcal{G}, \mathbb{R})$.

(iii) $\overline{x} \mapsto \overline{x}_\pm$ are in $C(\Omega \times (0,T), \mathbb{R}^{d+1})$. △

Proof. It suffices to show the assertions for the subscript "+". To see that τ_+ is continuous on $\Omega \times (0,T) \cup \mathcal{G}$, let $\overline{x} = (x,t)$, $\overline{y} = (y,r) \in \Omega \times (0,T) \cup \mathcal{G}$, $\varepsilon > 0$ and $\tau > t$ such that $0 < \tau_+(\overline{x}) - \tau < \varepsilon$. For fixed small enough $\alpha > 0$ the trajectory $\Theta := \gamma_{\overline{x}}([t+\alpha, \tau])$ is compact and contained in $\Omega \times (0,T)$, hence a neighbourhood $B_r(\Theta)$ of this trajectory is also contained in $\Omega \times (0,T)$ for small $r > 0$. From the theory of ordinary differential equations we know that for $|(x,t) - (y,r)| < \delta$ it holds

$$\gamma_{\overline{y}}(\tilde{\tau}) \to \gamma_{\overline{x}}(\tilde{\tau}) \tag{3.18}$$

uniformly in $\tilde{\tau} \in [t+\alpha, \tau]$ as $\delta \to 0$. In particular, for small enough δ we see that $\gamma_{\overline{y}}(\tilde{\tau}) \in \Omega \times (0,T)$ for $\tilde{\tau} \in [t+\alpha, \tau]$ and thus $\tau_+(\overline{y}) \geq \tau > \tau_+(\overline{x}) - \varepsilon$. Interchanging the roles of $\overline{x}$ and $\overline{y}$, we obtain

$$\tau_+(\overline{y}) > \tau_+(\overline{x}) - \epsilon \qquad \text{and} \qquad \tau_+(\overline{x}) > \tau_+(\overline{y}) - \varepsilon \tag{3.19}$$

or, equivalently, $|\tau_+(\overline{y}) - \tau_+(\overline{x})| < \varepsilon$ for $|(x,t) - (y,r)| < \delta$ with $\delta > 0$ small enough.
Now that we have proven that τ_+ is continuous, it follows again from the theory of ordinary differential equations (specifically, from the continuity of an ODE solution in the initial data and in its argument) that $\overline{x} \mapsto \overline{x}_+ = \gamma_{\overline{x}}(\tau_+(\overline{x}))$ is continuous. □

Lemma 3.9. *Let $\Sigma \subset \mathcal{G}$.*

(i) It Σ is open in $\partial\Omega \times [0,T) \cup \Omega \times \{0\}$, then $\mathcal{S}_\Sigma \subset \Omega \times (0,T)$ is open in $\mathbb{R}^{d+1}$.

(ii) If Σ is path-connected, so is $\mathcal{S}_\Sigma$.

In particular, if the premises of (i) and (ii) hold, then $\mathcal{S}_\Sigma$ is a domain. △

Proof. We use Lemma 3.8. To show (i), let $\overline{x} \in \mathcal{S}_\Sigma$. Then the continuity of $\overline{x} \mapsto \overline{x}_-$ on $\Omega \times (0,T)$ implies that small enough open r-balls around $\overline{x}$ have their image in a small neighbourhood $V \subset \Sigma$ containing $\overline{x}_-$. The continuity of τ_+ on $\mathcal{G}$ then yields that $B_r(\overline{x}) \subset \mathcal{S}_\Sigma$ for small enough $r > 0$.
For the proof of (ii), let $\overline{x}, \overline{y} \in \mathcal{S}_\Sigma$. If these points lie on the same characteristic, the latter gives a path in $\mathcal{S}_\Sigma$ between $\overline{x}$ and $\overline{y}$. Otherwise, $\overline{x}$ and $\overline{y}$ can be connected to $\overline{x}_- \in \Sigma$ and $\overline{y}_- \in \Sigma$, respectively, by the trajectories of their respective characteristics. Let $\zeta : [0,1] \to \Sigma$ be continuous with $\zeta(0) = \overline{x}_-$ and $\zeta(1) = \overline{y}_-$. Then $\zeta([0,1]) \subset \mathcal{G}$ is compact and thus there exists $\mu > 0$ such that $\gamma_{\overline{z}}(\tau) \in \mathcal{S}_\Sigma$ for each $\overline{z} = (z,t) \in \zeta([0,1])$ and $\tau \in (t, t+\mu)$. Then the path $\tilde{\gamma} : [0,1] \to \mathcal{S}_\Sigma$ defined by

$$\tilde{\gamma}(\tau) = \gamma_{\zeta(\tau)}\left(t(\zeta(\tau)) + \frac{\mu}{3} + \tau\frac{\mu}{3}\right) \tag{3.20}$$

is continuous and connects the characteristic trajectories of $\overline{x}$ and $\overline{y}$ within $\mathcal{S}_\Sigma$. By concatenating the three paths collected so far, we find a continuous path in $\mathcal{S}_\Sigma$ from $\overline{x}$ to $\overline{y}$. □

Corollary 3.10. *If $\mathcal{S}_\Sigma \subset \Omega \times (0,T)$ is a domain, then so is $\pi_x(\mathcal{S}_\Sigma) = \{x \in \Omega : (x,t) \in \mathcal{S}_\Sigma\}$.* △

Proof. This follows from the fact that $\pi_x : \mathbb{R}^{d+1} \to \mathbb{R}^d$, $\pi_x(x,t) = x$, is an open continuous map. □

Let us fix the following assumptions:

Assumption 3.11. (i) The reaction term is strictly positive: There exists a constant $c_0 > 0$ such that $c \geq c_0$ on Ω_T.

(ii) $\Omega \subset \mathbb{R}^d$ is a domain with C^2 boundary.

(iii) b, c, u_0 and f are sufficiently smooth and compatible to allow for classical $C^2(\overline{\mathcal{S}_\Sigma})$ solutions u_ϵ and u.

(iv) $\Sigma \subset \mathcal{G}$ is compact and such that an open neighbourhood V of Σ in $\partial\Omega \times [0,T) \cup \Omega \times \{0\}$ is contained in $\Gamma_- \times [0,T) \cup \Omega \times \{0\}$ and such the boundary portion $\mathcal{B}_2(V)$ of $\partial\mathcal{S}_V$ is a compact subset of $\Gamma_+ \times [0,T)$.

(v) For V as in (iv), there are constants $\gamma_0, c > 0$ and a function $\Psi \in C^2(\overline{\pi_x(\mathcal{S}_V)})$ such that for $d := \operatorname{dist}(\cdot, \partial\Omega)$

$$\begin{cases} \Psi(x) = -d(x) & \text{on } U_\gamma := \{x \in \overline{\pi_x(\mathcal{S}_V)} : \operatorname{dist}(x, \pi_x(\mathcal{B}_2(V))) < \gamma\} \\ \Psi(x) \leq -c\gamma & \text{on } \overline{\pi_x(\mathcal{S}_V)} \setminus U_\gamma \end{cases} \tag{3.21}$$

holds for all $0 < \gamma \leq \gamma_0$. Then we define $F_+ := \Psi \circ \pi_x \in C^2(\overline{\mathcal{S}_V})$.

△

Remark 3.12. Part (i) of the assumption does not pose a loss of generality, since by a simple transformation, problems (3.1) and (3.2) can be altered such that the reaction term is strictly bounded away from zero. Indeed, for some $\lambda > 0$ define

$$u_\epsilon^\lambda(x,t) := e^{-\lambda t} u_\epsilon(x,t), \qquad u^\lambda(x,t) := e^{-\lambda t} u(x,t) \qquad \text{and} \qquad f^\lambda(x,t) := e^{-\lambda t} f(x,t) \tag{3.22}$$

for $(x,t) \in \Omega_T$. Then u_ϵ and u are solutions of the problems (3.1) and (3.2), respectively, if and only if u_ϵ^λ and u^λ are solutions to the same problems with c replaced by $c + \lambda \geq \lambda > 0$. Thus, if for a subset $S \subset \Omega_T$ we have shown that

$$\left\| u_\epsilon^\lambda - u^\lambda \right\|_{L^\infty(S)} \leq K\epsilon \tag{3.23}$$

for a constant K independent of ϵ, we obtain qualitatively the same result for the original problem:

$$\|u_\epsilon - u\|_{L^\infty(S)} \leq e^{\lambda T} K\epsilon. \tag{3.24}$$

In certain cases, the cylinder Ω_T may be approximated well by $\mathcal{S}_\Sigma$ satisfying assumption (iv), namely if the bundle of characteristics touching the parabolic boundary does not occupy a significant fraction of the volume of Ω_T. △

For proving our desired convergence result, we need a maximum principle on non-cylindrical domains (i.e. not of the form $\Omega \times I$ for a real interval I). To this end, we cite a strong maximum principle from Friedman:

Proposition 3.13 (Strong maximum principle on general domains, [Fri64, Chapter 2, Theorem 1]). *Let $D \subset \mathbb{R}^{d+1}$ be a domain, a_{ij}, b_i and $c \geq 0$ for $i, j = 1, \ldots, d$ continuous functions on D. Let $\mathcal{L}$ be a parabolic operator defined by*

$$\mathcal{L}u := \frac{\partial u}{\partial t} - \sum_{i,j=1}^{d} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} + cu, \tag{3.25}$$

for $u \in C_1^2(D)$ (continuous x-derivatives of degree 2, once continuously differentiable in t) where parabolicity means that

$$\sum_{i,j=1}^{d} a_{ij}(x,t)\xi_i \xi_j > 0 \tag{3.26}$$

for any $(x,t) \in D$ and $\xi \in \mathbb{R}^d \setminus \{0\}$. Then if $\mathcal{L}u \geq 0$ and u has a negative minimum in D attained at a point $\overline{x} = (x,t) \in D$, it follows that $u(\overline{y}) = u(\overline{x})$ for all $\overline{y} \in D$ that can be connected to $\overline{x}$ by a simple connected curve in D along which the t-coordinate is non-increasing going from $\overline{y}$ to $\overline{x}$. △

Corollary 3.14 (Weak maximum principle for characteristic tubes of balls). *Let $\overline{x} \in \mathcal{G}$ and $r > 0$ such that $B := B_r(\overline{x}) \cap (\partial\Omega \times [0,T) \cup \Omega \times \{0\})$ is contained in $\mathcal{G}$. Let $\mathcal{L}$ be as in Proposition 3.13 with $c > 0$ bounded away from zero and let $u, v \in C^2(\overline{\mathcal{S}_B})$ with*

$$|u| \leq v \qquad \text{on } \overline{B} \cup \mathcal{B}_1(B) \cup \mathcal{B}_2(B) \tag{3.27}$$

$$|\mathcal{L}u| \leq \mathcal{L}v \qquad \text{on } \overline{\mathcal{S}_B}. \tag{3.28}$$

Then $|u| \leq v$ on $\overline{\mathcal{S}_B}$. △

Proof. We first show that $w \geq 0$ on $\overline{B} \cup \mathcal{B}_1(B) \cup \mathcal{B}_2(B)$ and $\mathcal{L}w \geq 0$ on $\overline{\mathcal{S}_B}$ imply $w \geq 0$ on $\overline{\mathcal{S}_B}$. The set $\overline{\mathcal{S}_B}$ is compact. If $\mu := \min_{\overline{x} \in \overline{\mathcal{S}_B}} w < 0$ were true, then this minimum could not be attained on $\overline{B} \cup \mathcal{B}_1 \cup \mathcal{B}_2$ by the premise and not on the interior $\mathcal{S}_B$ owing to Proposition 3.13, since every point of $\mathcal{S}_B$ can, by definition of $\mathcal{S}_B$, be connected to a point arbitrarily close to B along a characteristic. By continuity, it follows that $w(\overline{x}) < 0$ is impossible for $\overline{x} \in \mathcal{S}_B$. Hence $\mu < 0$ must be attained at some $\overline{x} \in \mathcal{B}_3(B) \subset \Omega \times \{t = T\}$. Therefore $\nabla w(\overline{x}) = 0$, $w_t(\overline{x}) \leq 0$ and the second order term in $\mathcal{L}w(\overline{x})$ is non-negative, giving

$$\mathcal{L}w(\overline{x}) = \left(\frac{\partial w}{\partial t} - \sum_{i,j=1}^{d} a_{ij} \frac{\partial^2 w}{\partial x_i \partial x_j} + b \cdot \nabla w + \underbrace{cw}_{<0} \right)(\overline{x}) < 0, \tag{3.29}$$

a contradiction. Now the assertion follows from linearity of $\mathcal{L}$ and by considering $w := v \pm u$. □

Theorem 3.15. *Assume the statements and notation from Assumption 3.11. Define $\gamma(\epsilon) := -\epsilon^{1/2} \ln(\epsilon)$ and $U_\epsilon := U_{\gamma(\epsilon)} \times [0,T]$. Then $\gamma(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$ and there exists an $\epsilon_0 > 0$ and a constant K such that*

$$\|u_\epsilon - u\|_{L^\infty(\mathcal{S}_\Sigma \setminus U_\epsilon)} \leq K\epsilon \quad \text{for } \epsilon \leq \epsilon_0 \tag{3.30}$$

and K does not depend on ϵ. △

Proof. All constants $K_i, i = 1, 2, \dots$ in this proof will be independent of ϵ as long as $\epsilon \leq \epsilon_0$ for some $\epsilon_0 > 0$. Set

$$B_s := B_s(\overline{x}_0) \cap (\partial\Omega \times [0, T) \cup \Omega \times \{0\}), \tag{3.31}$$

where $B_s(\overline{x}_0)$ is the open ball in $\mathbb{R}^{d+1}$ of radius $s > 0$ around $\overline{x}_0$. We shall prove that, for any $\overline{x}_0 \in \Sigma$ and $r > 0$ such that $B_r \subset V$, the desired result holds on $\mathcal{S}_{B_{r/2}}$. Then by compactness of Σ, we can cover $\mathcal{S}_\Sigma$ by the characteristic tubes of finitely many balls such that the corresponding balls of halved radius cover Σ and the assertion follows.

Define the differential operators

$$\mathcal{L}_0(v) := v_t + b \cdot \nabla v + cv, \qquad L_1(v) := \Delta v, \qquad \mathcal{L}_\epsilon := \mathcal{L}_0 - \epsilon L_1 \tag{3.32}$$

and the error function $z_\epsilon := u_\epsilon - u$. With the help of the parabolic maximum principle from the last chapter, it can be seen that $\|u_\epsilon\|_{L^\infty(\Omega_T)}$ remains bounded uniformly in ϵ. Thus, because $\mathcal{L}_\epsilon(z_\epsilon) = \mathcal{L}_\epsilon u_\epsilon - \mathcal{L}_0 u + \epsilon L_1 u = \epsilon \Delta u$ on Ω_T and u_ϵ and u agree on $\mathcal{G}$, we have the following estimates:

$$|z_\epsilon| \leq K_1 \qquad \text{on } \overline{\mathcal{S}_{B_r}} \tag{3.33}$$

$$z_\epsilon = 0 \qquad \text{on } \Sigma \tag{3.34}$$

$$|\mathcal{L}_\epsilon z_\epsilon| \leq K_2 \epsilon \qquad \text{on } \overline{\mathcal{S}_{B_r}}. \tag{3.35}$$

In order to circumvent problems arising from the non-smoothness of the boundary $\partial\Omega_T$, we regard a Lipschitz continuous extension of $\tilde{b}$ to a smooth domain $D \supset \overline{\Omega_T}$ with $\operatorname{dist}(\partial\Omega_T, \partial D)$ small enough. Due to our assumption that $B_r \subset \Gamma_- \times [0, T) \cup \Omega \times \{0\}$, the extension backward in time of the characteristics $\gamma_{\overline{z}}$, $\overline{z} \in B_r$, exit $\partial\Omega_T$ at $\overline{z}$ and cut the smooth surface $\mathcal{M} := \partial D$. For $\overline{x} \in B_r$, let $\overline{x}_\gamma$ be the intersection of $\mathcal{M}$ and the extended characteristic through $\overline{x}$. Then we find a smooth function ψ on $\mathcal{M}$ such that

$$\psi(\overline{z}) \begin{cases} = -\frac{r}{2} & \text{for } \overline{z} = \overline{x}_\gamma, \overline{x} \in B_{r/2} \\ \in [-\frac{r}{2}, 0] & \text{for } \overline{z} = \overline{x}_\gamma, \overline{x} \in B_{3r/4} \setminus B_{r/2} \\ = 0 & \text{else.} \end{cases} \tag{3.36}$$

We extend this function constantly along characteristics to obtain a function $F_0 \in C^2(\overline{\mathcal{S}_{B_r}})$. This is achieved by defining F_0 to be the solution of the problem

$$\begin{cases} \partial_t F_0 + b \cdot \nabla F_0 = 0 & \text{on } \mathcal{S}_{B_r} \\ F_0 = \psi & \text{on } \mathcal{M}. \end{cases} \tag{3.37}$$

Now we define the barrier function $S_\epsilon \in C^2(\overline{\mathcal{S}_{B_r}})$ by

$$S_\epsilon(\overline{x}) := K_3\epsilon + \mu(\epsilon) + K_4 \exp\left(\frac{F_0(\overline{x})}{K_5 \epsilon^{1/2}}\right) + K_6 \exp\left(\frac{F_+(\overline{x})}{K_7 \epsilon}\right) \tag{3.38}$$

with $K_3, \dots, K_7, \mu(\epsilon) > 0$ to be determined momentarily. At first we note that

$$S_\epsilon(\overline{x}) \geq K_3\epsilon \geq 0 = |z_\epsilon(\overline{x})| \qquad \text{on } \overline{B_r} \tag{3.39}$$

$$S_\epsilon(\overline{x}) \geq \min(K_4, K_6) \geq K_1 \geq |z_\epsilon(\overline{x})| \qquad \text{on } \mathcal{B}_1(B_r) \cup \mathcal{B}_2(B_r) \tag{3.40}$$

for $K_4, K_6 \geq K_1$ by construction of the functions F_0 and F_+ which vanish on $\mathcal{B}_1(B_r)$ and $\mathcal{B}_2(B_r)$, respectively. In order to be able to apply Corollary 3.14, we compute for $\overline{x} \in \overline{\mathcal{S}_{B_r}}$:

$$\begin{aligned}\mathcal{L}_\epsilon S_\epsilon(\overline{x}) &= c(\overline{x})(K_3\epsilon + \mu(\epsilon)) \\ &\quad + K_4 \exp\left(\frac{F_0(\overline{x})}{K_5\epsilon^{1/2}}\right)\left\{\frac{\partial_t F_0(\overline{x}) + b\cdot\nabla F_0(\overline{x})}{K_5\epsilon^{1/2}} - \epsilon\left(\frac{\Delta F_0(\overline{x})}{K_5\epsilon^{1/2}} + \frac{|\nabla F_0(\overline{x})|^2}{K_5^2\epsilon}\right) + c(\overline{x})\right\} \\ &\quad + K_6 \exp\left(\frac{F_+(\overline{x})}{K_7\epsilon}\right)\left\{c(\overline{x}) - \frac{\Delta F_+(\overline{x})}{K_7} + \frac{1}{K_7\epsilon}\left(b\cdot\nabla F_+(\overline{x}) - \frac{1}{K_7}|\nabla F_+(\overline{x})|^2\right)\right\}.\end{aligned} \tag{3.41}$$

The first braced term is non-negative for $\epsilon \leq \epsilon_0$ for large enough $K_5 > 0$ due to our assumption that $c \geq c_0 > 0$ and (3.37).
Asserting non-negativity of the second brace needs more careful attention. The term $c - \Delta F_+/K_7$ is non-negative for large enough K_7. If $\overline{x} \in \overline{\mathcal{S}_{B_r}} \cap U_{\gamma_0} \times [0,T]$ for $\gamma_0 > 0$ chosen small enough, then Assumption 3.11 (iv) and (v) give for some constants $\mu_1, M_1, M_2 > 0$ that

$$M_1 \geq b\cdot\nabla F_+ \geq \mu_1 > 0 \qquad \text{and} \qquad M_2 \geq |\nabla F_+|^2 \tag{3.42}$$

on $U_{\gamma_0} \times [0,T]$, because $-\nabla d$ is the outward unit normal on $\partial\Omega$ for $d = \operatorname{dist}(\cdot, \partial\Omega)$. From this we infer that

$$E(\overline{x}) := b\cdot\nabla F_+(\overline{x}) - \frac{1}{K_7}|\nabla F_+(\overline{x})|^2 \geq 0 \tag{3.43}$$

for $\overline{x} \in U_\gamma \times [0,T]$ for all $0 < \gamma \leq \gamma_0$ for sufficiently large K_7.
If, on the other hand, $\overline{x} \in \overline{\mathcal{S}_{B_r}} \setminus U_\gamma \times [0,T]$, Assumption 3.11 (v) gives that $F_+(\overline{x}) \leq -c\gamma$ and thus

$$K_6 \exp\left(\frac{F_+(\overline{x})}{K_7\epsilon}\right)\left(\frac{E(\overline{x})}{K_7\epsilon}\right) \leq K_6 \exp\left(\frac{-c\gamma}{K_7\epsilon}\right)\left(\frac{E(\overline{x})/K_7}{\epsilon}\right). \tag{3.44}$$

This – possibly not everywhere non-negative – term can be counterbalanced by setting

$$\mu(\epsilon) := \frac{K_6}{c_0}\exp\left(\frac{-c\gamma}{K_7\epsilon}\right)\frac{K_8}{\epsilon} \tag{3.45}$$

with $K_8 \geq \max_{\overline{x}\in\overline{\mathcal{S}_{B_r}}} E(\overline{x})/K_7$. Overall, we have shown that for suitably chosen $K_3, \ldots, K_8$, we have

$$\mathcal{L}_\epsilon S_\epsilon \geq c_0 K_3 \epsilon \geq K_2\epsilon \geq |\mathcal{L}_\epsilon z_\epsilon| \quad \text{on } \overline{\mathcal{S}_{B_r}}. \tag{3.46}$$

Corollary 3.14 then gives us the desired estimate:

$$|z_\epsilon(\overline{x})| \leq K_3\epsilon + K_4\exp\left(\frac{F_0(\overline{x})}{K_5\epsilon^{1/2}}\right) + K_6\exp\left(\frac{F_+(\overline{x})}{K_7\epsilon}\right) + \frac{K_6}{c_0}\exp\left(\frac{-c\gamma}{K_7\epsilon}\right)\frac{K_8}{\epsilon} \quad \text{for } \overline{x} \in \overline{\mathcal{S}_{B_r}}. \tag{3.47}$$

If we set $\gamma := \gamma(\epsilon) := -\epsilon^{1/2}\ln(\epsilon)$, then $\gamma(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$ and we can show that the right hand side of (3.47) is $\mathcal{O}(\epsilon)$ on $\mathcal{S}_{B_{r/2}} \setminus U_{\gamma(\epsilon)} \times [0,T]$; indeed, on $\mathcal{S}_{B_{r/2}}$ we have $F_0 = -r/2$ by construction, so that

$$\exp\left(\frac{F_0(\overline{x})}{K_5\epsilon^{1/2}}\right) = \exp\left(\frac{-r/2}{K_5\epsilon^{1/2}}\right) = \mathcal{O}(\epsilon^n) \quad \text{as } \epsilon \downarrow 0 \tag{3.48}$$

for any $n \in \mathbb{N}$. On $\mathcal{S}_{B_{r/2}} \setminus U_{\gamma(\epsilon)} \times [0,T]$, by construction, $F_+ \leq -c\gamma(\epsilon) = c\epsilon^{1/2}\ln(\epsilon)$, so that

$$\exp\left(\frac{F_+(\overline{x})}{K_7\epsilon}\right) \leq \exp\left(\frac{c\ln(\epsilon)}{K_7\epsilon^{1/2}}\right) = \epsilon^{\frac{c}{K_7\epsilon^{1/2}}} = \mathcal{O}(\epsilon^n) \tag{3.49}$$

for any $n \in \mathbb{N}$, which also shows that the last summand in (3.47) is $\mathcal{O}(\epsilon^n)$ for any $n \in \mathbb{N}$. This shows that

$$|z_\epsilon(\overline{x})| = \mathcal{O}(\epsilon) \quad \text{on } \mathcal{S}_{B_{r/2}} \setminus U_{\gamma(\epsilon)} \times [0,T] \quad \text{as } \epsilon \downarrow 0, \tag{3.50}$$

which – combined with the remark made in the beginning of this proof – concludes the proof. □

4. First Order Upwinding of the Convective Part and the LED Principle

4.1 Upwinding in One Dimension

We have seen in the introduction that employing an upwind-facing difference instead of a central difference in the convective part of the transport operator in the mass-lumped finite element scheme seems to resolve the oscillation issues in the considered one-dimensional homogeneous convection-diffusion equation with forward Euler stepping.

In order to understand why this is the case, in subsection 4.1.1 we put forth an interesting property of one-dimensional convection-diffusion equations that prohibits oscillations in classical solutions. In the subsequent section, we show that in the considered 1D case, finite element (FE), finite volume (FV) and finite difference (FD) schemes with explicit Euler time-stepping are essentially equivalent and introduce the term *upwinding* and the rationale for its use.

4.1.1 Total Variation

Definition 4.1 (Continuous total variation)**.** Let $\alpha < \beta \in \mathbb{R}$ and $t \in \mathbb{R}_{>0}$. Define the *total variation* of a function $f : [\alpha, \beta] \to \mathbb{R}$ as

$$\mathrm{Var}(f; [\alpha, \beta]) := \sup \left\{ \sum_{i=0}^{n-1} |f(x_{i+1}) - f(x_i)| : n \in \mathbb{N}, \alpha \leq x_0 < \cdots < x_n \leq \beta \right\}.$$

For a function Ψ defined on the U-shaped set $\Gamma_t := [\alpha, \beta] \times \{0\} \cup \{\alpha, \beta\} \times [0, t] \subset \mathbb{R}^2$ define its total variation $\mathrm{Var}(\Psi; \Gamma_t)$ on this set by

$$\mathrm{Var}(\Psi; \Gamma_t) := \mathrm{Var}(\tilde{\Psi}; [\alpha - t, \beta + t] \subset \mathbb{R}), \tag{4.1}$$

where $\tilde{\Psi}$ is defined through Ψ in the obvious way by flapping down the "walls" of the U onto $\mathbb{R} \times \{0\}$. △

Proposition 4.2 (Evolution of total variation for 1D convection-diffusion, [Sat69, Theorem 2])**.** *Let $\alpha < \beta \in \mathbb{R}$, $T > 0$, $\Omega := (\alpha, \beta)$, $\Omega_t := (\alpha, \beta) \times (0, t)$ for $t \in (0, T]$ and $\Gamma_T = \partial\Omega_T \backslash (\alpha, \beta) \times \{t = T\}$*

as usual. Consider the problem

$$\begin{cases} u_t - au_{xx} + bu_x = 0 & \text{in } \Omega_T \\ u = \Psi & \text{on } \Gamma_T, \end{cases} \tag{4.2}$$

where a, b, a_x and b_x are Hölder continuous on $\overline{\Omega_T}$ and Ψ is continuous on Γ_T with bounded variation, i.e. $\mathrm{Var}(\Psi; \Gamma_T) < \infty$. *Then the (classical) solution to this problem satisfies*

$$\mathrm{Var}(u(\cdot, t); [\alpha, \beta]) \leq \mathrm{Var}(\Psi; \Gamma_t) \quad \text{for all } t \in [0, T]. \tag{4.3}$$

△

Remark 4.3. In particular, the above theorem states that for continuous initial-boundary conditions Ψ with homogeneous Dirichlet boundary data, the total variation of the homogeneous 1D convection-diffusion (classical) solution at any time $t > 0$ can never exceed the total variation of the initial value u_0. This justifies calling the oscillations observed in the introductory example *spurious*, as they clearly increase the total variation. A numerical scheme diminishing a discrete total variation for this problem would thus be desirable. △

Definition 4.4 (Discrete total variation, TVD)**.** Let $I \subset \mathbb{Z}$ be an interval and $v = (v_i)_{i \in I}$ a discrete function. Then, analogously to the definition of the total variation of functions on the continuum, its *total variation* is defined as

$$\mathrm{TV}(v) := \sum_{i \in I} |v_{i+1} - v_i|, \tag{4.4}$$

where $v_j := 0$ for $j \notin I$. A scheme updating v^j to $v^{j+1} = (v_i^{j+1})_{i \in I}$ is called *total variation diminishing (TVD)* if

$$\mathrm{TV}(v^{j+1}) \leq \mathrm{TV}(v^j). \tag{4.5}$$

△

4.1.2 Equivalence of FD, FE and FV

Let $\Omega = (\alpha, \beta) \subset \mathbb{R}$, $b \in \mathbb{R}_{>0}$ constant. We would like to solve

$$\begin{cases} u_t - \epsilon u_{xx} + bu_x = 0 & \text{in } \Omega_T \\ u = 0 & \text{on } \partial\Omega \times [0, T] \\ u = u_0 & \text{on } \Omega \times \{t = 0\} \end{cases} \tag{4.6}$$

on a triangulation $\mathcal{T}$ of Ω with nodes $\alpha = x_1 < \cdots < x_N = \beta$ dividing Ω into intervals of equal length $h = (\beta - \alpha)/(N-1)$ by employing forward Euler time-stepping with a fixed time-step length τ. Assume $u_0 \in V_0^1(\Omega)$ and let u_i^j be a shorthand for $u_h(ih, j\tau)$, the sought numerical solution at the grid points.

We show that the finite difference scheme

$$\frac{u_i^{j+1} - u_i^j}{\tau} - \epsilon \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} + b \frac{u_{i+1}^j - u_{i-1}^j}{2h} = 0 \quad \text{for } i \in \{2, \ldots, N-1\} \tag{4.7}$$

is rather universal in the sense that it can also be interpreted as a finite element or finite volume scheme. Indeed, the standard Galerkin FE scheme with explicit Euler time-steps reads

$$\begin{cases} \left(\dfrac{u^{n+1}-u^n}{\tau}, \varphi_i\right) + \epsilon(u_x^n, (\varphi_i)_x) + b(u_x^n, \varphi_i) = 0 \quad \text{for all } i \in \{2, \dots, N-1\} \\ u^0 = u_0. \end{cases} \tag{4.8}$$

Writing this using the matrices $M_C^\circ, C^\circ, D^\circ \in \mathbb{R}^{M\times M}$ from Definition 0.4 with $M = N-2$, the Galerkin method can be restated equivalently (now denoting by $u^n \in \mathbb{R}^{N-2}$ the coefficients of the discrete solution at time $n\tau$ with respect to the hat function basis)

$$\begin{cases} M_C^\circ \left(\dfrac{u^{n+1}-u^n}{\tau}\right) + \epsilon D^\circ u^n + C^\circ u^n = 0 \\ u^0 = u_0. \end{cases} \tag{4.9}$$

Simple calculations show that

$$m_{ii} = \frac{2}{3}h \qquad m_{ij} = \frac{1}{6}h \quad \text{for } |i-j| = 1 \qquad m_{ij} = 0 \text{ else} \tag{4.10}$$

$$d_{ii} = \frac{2}{h} \qquad d_{ij} = -\frac{1}{h} \quad \text{for } |i-j| = 1 \qquad d_{ij} = 0 \text{ else} \tag{4.11}$$

$$c_{ii} = 0 \qquad c_{i,i\pm1} = \pm\frac{1}{2} \qquad c_{ij} = 0 \text{ else} \tag{4.12}$$

for $i, j \in \{1, \dots, N\}$. Replacing M_C° by the restricted lumped mass matrix M_L° and dividing the whole equation by h, we immediately obtain the FD scheme (4.7). Note that, for the equivalence to hold, we really need to sum over all $j \in \{1, \dots, N\}$ in (0.10).

In order to interpret (4.7) as a FV method, we introduce the *dual cells*

$$C_i := [x_{i-1/2}, x_{i+1/2}],$$

where $x_{i+1/2} := (x_i + x_{i+1})/2$ for $i \in \{1, \dots, N-1\}$, $x_{1/2} = x_1, x_{N+1/2} = x_N$, and interpret u_i^j as an approximation to the mean value of the true solution u over C_i at time $j\tau$ (as opposed to the interpretation as nodal values in FD). If we integrate the differential equation in (4.6) over an interior cell C_i, $i \in \{2, \dots, N-1\}$, we re-obtain the underlying conservation law

$$\frac{d}{dt}\frac{1}{h}\int_{C_i} u(\cdot, j\tau)\, dx = \overline{F}_{i-1/2}^j - \overline{F}_{i+1/2}^j, \tag{4.13}$$

where

$$\overline{F}_{i+1/2}^j := \frac{b}{h}u(x_{i+1/2}, j\tau) - \frac{\epsilon}{h}u_x(x_{i+1/2}, j\tau) \tag{4.14}$$

are the true fluxes at the cell boundaries. Now we set $u_1^j = u_N^j = 0$ to implement the boundary conditions, use the natural approximations

$$u(x_{i+1/2}, j\tau) \approx \frac{u(x_{i+1}, j\tau) + u(x_i, j\tau)}{2} \tag{4.15}$$

$$u_x(x_{i+1/2}, j\tau) \approx \frac{u(x_{i+1}, j\tau) - u(x_i, j\tau)}{h} \tag{4.16}$$

to define the numerical fluxes as

$$F^j_{i+1/2} := b\frac{u^j_{i+1} + u^j_i}{2h} - \epsilon\frac{u^j_{i+1} - u^j_i}{h^2}, \tag{4.17}$$

approximate the time derivative by a forward difference and remember our interpretation of u^j_i to obtain the FV scheme

$$\frac{u^{j+1}_i - u^j_i}{\tau} = \epsilon\frac{u^j_{i+1} - 2u^j_i + u^j_{i-1}}{h^2} - b\frac{u^j_{i+1} - u^j_{i-1}}{2h}, \tag{4.18}$$

which is, again, exactly (4.7).

We call the first summand in the numerical flux (4.17) the *central flux* approximation to the convective flux; in this approximation, both adjoining cells of an interface $x_{i+1/2}$ enter with the weighting factor $1/2$. It is well-known that a von Neumann stability analysis for the reduced hyperbolic problem ($\epsilon = 0$ and no boundary condition on $\{\beta\} \times [0, T]$) reveals that the central flux scheme is unconditionally unstable, i.e. unstable no matter how small $\tau > 0$ is chosen. This is commonly explained by the reasoning that the stencil of the numerical convective flux resulting from the central approximation includes a point *downwind* from the interface $x_{i+1/2}$ (i.e. to the right (left) of $x_{i+1/2}$ for $b > 0$ (for $b < 0$), which is unnatural since information is advected along characteristics from *upwind* from $x_{i+1/2}$.

This gives rise to the following alteration in the numerical flux called *upwinding of the convective part*:

$$F^{j,\mathrm{up}}_{i+1/2} := b\frac{u^j_i}{h} - \epsilon\frac{u^j_{i+1} - u^j_i}{h^2}, \tag{4.19}$$

yielding the upwinded scheme

$$\frac{u^{j+1}_i - u^j_i}{\tau} - \epsilon\frac{u^j_{i+1} - 2u^j_i + u^j_{i-1}}{h^2} + b\frac{u^j_i - u^j_{i-1}}{h} = 0 \quad \text{for } i \in \{2, \dots, N-1\}. \tag{4.20}$$

4.1.3 Upwinding in 1D implies TVD

Proposition 4.5 ([Har83, Lemma 2.2])**.** *A difference scheme of the form*

$$u^{j+1}_i = u^j_i + C_{+,i+1/2}(u^j_{i+1} - u^j_i) - C_{-,i-1/2}(u^j_i - u^j_{i-1}) \quad \textit{for } j \in \mathbb{N}_0, i \in \mathbb{Z} \tag{4.21}$$

is TVD provided that

$$C_{-,i+1/2}, C_{+,i+1/2} \geq 0 \textit{ and } C_{-,i+1/2} + C_{+,i+1/2} \leq 1 \tag{4.22}$$

holds for all $i \in \mathbb{Z}$. △

Note that we can extend the scheme (4.20) to have the form needed to apply the previous theorem by requiring $u^j_i = 0$ for any $j \in \mathbb{N}_0, i \in \mathbb{Z} \setminus \{2, \dots, N-1\}$. This extension does not change the total variation.

Corollary 4.6. *The upwinded scheme* (4.20) *is TVD under the CFL-like condition*

$$\tau\left(\frac{b}{h}+\frac{2\epsilon}{h^2}\right)\leq 1. \tag{4.23}$$

△

Proof. Simple shifting of terms shows

$$u_i^{j+1}=u_i^j+\frac{\tau\epsilon}{h^2}(u_{i+1}^j-u_i^j)-\left(\frac{\tau\epsilon}{h^2}+\frac{\tau b}{h}\right)(u_i^j-u_{i-1}^j)\quad\text{for } i\in\{2,\dots,N-1\} \tag{4.24}$$

and of course

$$u_i^{j+1}=u_i^j=0\quad\text{for } i\in\mathbb{Z}\setminus\{2,\dots,N-1\}. \tag{4.25}$$

In view of Proposition 4.5 we define

$$C_{+,i+1/2}:=\begin{cases}\frac{\tau\epsilon}{h^2} & \text{for } i\in\{2,\dots,N-1\}\\ 0 & \text{else}\end{cases},\quad C_{-,i-1/2}:=\begin{cases}\frac{\tau\epsilon}{h^2}+\frac{\tau b}{h} & \text{for } i\in\{2,\dots,N-1\}\\ 0 & \text{else}\end{cases} \tag{4.26}$$

and see that the only condition to be fulfilled is (4.23). □

Note that without upwinding, i.e. if central differences are applied in the convective part, the scheme is

$$\frac{u_i^{j+1}-u_i^j}{\tau}-\epsilon\frac{u_{i+1}^j-2u_i^j+u_{i-1}^j}{h^2}+b\frac{u_{i+1}^j-u_{i-1}^j}{2h}=0\quad\text{for } i\in\{1,\dots,N-1\}, \tag{4.27}$$

and the constants for Harten's lemma are

$$C_{+,i+1/2}:=\begin{cases}\frac{\tau\epsilon}{h^2}-\frac{\tau b}{2h} & \text{for } i\in\{2,\dots,N-1\}\\ 0 & \text{else}\end{cases},\quad C_{-,i-1/2}:=\begin{cases}\frac{\tau\epsilon}{h^2}+\frac{\tau b}{2h} & \text{for } i\in\{2,\dots,N-1\}\\ 0 & \text{else,}\end{cases} \tag{4.28}$$

so that the conditions

$$\tau\leq\frac{h^2}{2\epsilon}\qquad\text{and}\qquad Pe:=\frac{bh}{\epsilon}\leq 2 \tag{4.29}$$

on the time-step and the cell Péclet number Pe have to imposed in order for Harten's criterion to guarantee the TVD property. The second condition can be formulated equivalently as

$$\frac{bh}{2\epsilon}\leq 1\iff h^2b^2\leq 4\epsilon^2\iff\frac{h^2}{2\epsilon}\leq\frac{2\epsilon}{b^2}, \tag{4.30}$$

because all terms involved are positive. This shows that the severe time-step restriction $\tau\leq 2\epsilon/b^2$ applies. We have seen in the introduction that conditions (4.23) and (4.29) seem to not only sufficient but *also necessary* for the upwinded (the non-upwinded) mass-lumped scheme to be TVD.

4.2 Upwinding in Multiple Dimensions

The combination of mass lumping and upwinding the convective part of the differential operator carried out in the previous section can also be applied in multidimensional problems. For this purpose, we want to show an equivalence of the FE convective part to a central numerical FV flux and then propose a manipulation of the stiffness matrix that represents upwinding. This is the content of the following subsection.

4.2.1 Manipulation of the Stiffness Matrix Resulting in Upwinding

We restrict ourselves to the case $d = 2$, but a generalization to conforming simplicial meshes in any dimension should be possible.

Definition 4.7 (Barycentric dual mesh)**.** Let $\mathcal{T}$ be a triangulation of $\Omega \subset \mathbb{R}^2$. Then its associated *barycentric dual mesh* is constructed in the following way:

- Connect the barycenter of each $T \in \mathcal{T}$ with the midpoint of its sides. This partitions each triangle into three quadrangles.
- Out of these quadrangles, define for each node $p_i \in \mathcal{N}$ the cell C_i around this node as the union of quadrangles Q with $p_i \in Q$ (see Figure 4.1). △

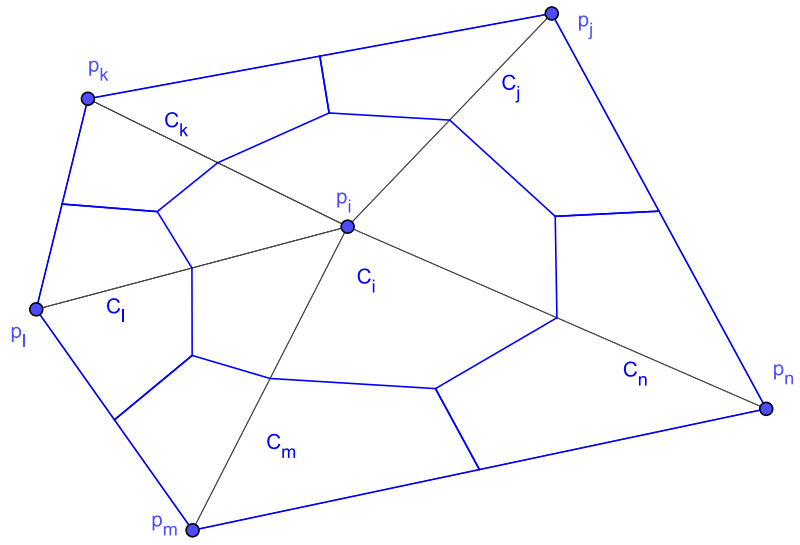

Figure 4.1: A triangulation (black) and its barycentric dual mesh (blue)

Lemma 4.8 (Connection between barycentric dual cells and mass lumping)**.** *For all dual cells C_i, $i = 1, \dots, N$, it holds $m_i = |C_i|$, where the m_i are the diagonal entries of the lumped mass matrix M_L.* △

Proof. The quadrangles of the first step of Definition 4.7 each have one third of the area of the

triangle they lie in, so that

$$|C_i| = \sum_{T\in\mathcal{T}:p_i\in T} \frac{1}{3}|T| = \int_\Omega \varphi_i \, dx = \sum_{j=1}^{N} (\varphi_j, \varphi_i) = m_i, \tag{4.31}$$

because the hat functions form a partition of unity on Ω. □

Motivated by the work in [Sel93], we show that the approximation of convection in the standard finite element approximation method with $V^1_{h,0}$ elements can be interpreted as a central flux finite volume approximation. Deviating from Selmin's paper, we consider only linear convection, but with a flux non-constant in space. However, for our argument to work, we need an additional assumption on the space-dependence of the vector field b, namely that b is elementwise constant and $b \in H(\mathrm{div};\Omega)$.

Assumption 4.9. $b : \Omega \times [0,T] \to \mathbb{R}^2$ is time-independent and piecewise constant on each $T \in \mathcal{T}$ and for each shared edge $e = \partial T \cap \partial T'$, its normal component $b \cdot n_e$ is continuous on e. △

Remark 4.10. This is equivalent to requiring that $b(\cdot,t) \in \mathcal{RT}_0(\Omega)$, the lowest-order Raviart-Thomas space, with $\operatorname{div} b = 0$. If the considered field b is not constant in space, a vector field as required by the assumption can be, for instance, obtained by projecting a time-independent, divergence-free field $b \in W^{1,\infty}(\Omega_T)$ into the lowest order Raviart-Thomas space $\mathcal{RT}_0(\Omega)$ via the standard projection operator π that sets the normal components of $\pi(b(\cdot,t))$ to the value $\int_e b(\cdot,t) \cdot n \, ds$ on each inter-element edge e. △

Lemma 4.11. *Let $T \subset \mathbb{R}^2$ be a triangle. Then, with the notation from Figure 4.2, it holds that*

$$l_1 n_1 + l_2 n_2 = |T| \nabla \varphi_i \tag{4.32}$$

$$l_1 n_1 = l_2 n_2 + l_3 n_3, \tag{4.33}$$

where n_r, $r = 1,2,3$, are supposed to be unit vectors. △

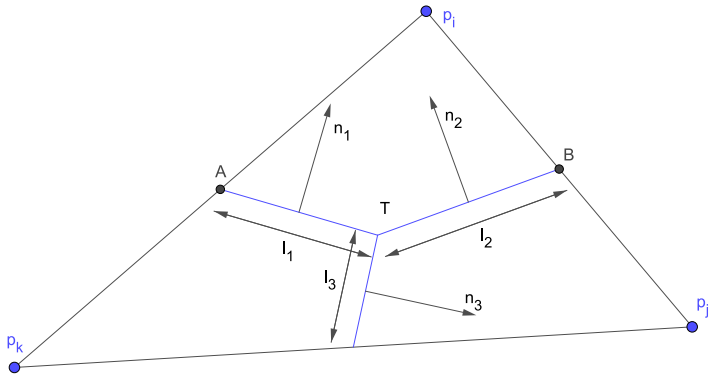

Figure 4.2: Notation for Lemma 4.11

Proof. Simple geometrical vector calculus shows that $\overrightarrow{AB} = \frac{\overrightarrow{p_k p_j}}{2}$. Denoting the triangle side opposite p_i by S_i, the outer normal over S_i by n^i, the perpendicular height over S_i by h_i and recalling that

$$|T|\nabla\varphi_i = \frac{1}{2}|S_i|h_i\nabla\varphi_i = -\frac{1}{2}|S_i|n^i,$$

the first equation follows from the divergence theorem applied to the triangle ATB. Another quick vector calculation shows that the lines labelled by their lengths l_1, l_2, l_3 form a triangle with outer unit normals $-n_1, n_2, n_3$, and thus the second equation follows. □

Theorem 4.12. *Let Assumption 4.9 hold and let p_i be a node of $\mathcal{T}$ such that*

$$\int_{\partial\Omega} \varphi_i\varphi_j b\cdot n \; ds = 0 \tag{4.34}$$

holds for $j = 1, \dots, N$. Then for $u \in \mathbb{R}^N$

$$\sum_{j\in K(i)} (b\cdot\nabla\varphi_j, \varphi_i)u_j = \sum_{j\in K(i)} \eta_{ij}\frac{u_i + u_j}{2}, \tag{4.35}$$

where

$$\eta_{ij} := l_{ij,1}\; b|_{T_1}\cdot n_{ij,1} + l_{ij,2}\; b|_{T_2}\cdot n_{ij,2} = \int_{\Gamma_{ij}} b\cdot n \; ds \qquad \text{for } p_i \text{ or } p_j \text{ interior} \tag{4.36}$$

$$\eta_{ij} := l_{ij,1}\; b|_{T_1}\cdot n_{ij,1} = \int_{\Gamma_{ij}} b\cdot n \; ds \qquad \text{for } p_i \text{ and } p_j \text{ boundary nodes} \tag{4.37}$$

and $n_{ij,k}$ is the outward unit normal of C_i on $\Gamma_{ij,k} := \partial C_i \cap \partial C_j \cap T_k$ and $l_{ij,k}$ the length of $\Gamma_{ij,k}$ (see Figure 4.3). △

Proof. Since, by Assumption 4.9, $b \in \mathcal{RT}_0(\Omega)$ with $b|_T$ constant for all $T \in \mathcal{T}$, we see that

$$\begin{aligned}(b\cdot\nabla\varphi_i, \varphi_j) &= \sum_{T\in\mathcal{T}}\int_T b\cdot\nabla\varphi_i\varphi_j \; dx = \sum_{T\in\mathcal{T}}\int_T \operatorname{div}(b\varphi_i)\varphi_j \; dx \\ &= -\sum_{T\in\mathcal{T}}\int_T \varphi_i b\cdot\nabla\varphi_j \; dx + \sum_{T\in\mathcal{T}}\int_{\partial T}\varphi_i\varphi_j b\cdot n \; ds \\ &= -\int_\Omega \varphi_i b\cdot\nabla\varphi_j \; dx = -(b\cdot\nabla\varphi_j, \varphi_i)\end{aligned} \tag{4.38}$$

holds for $j = 1, \dots, N$. Therefore we get

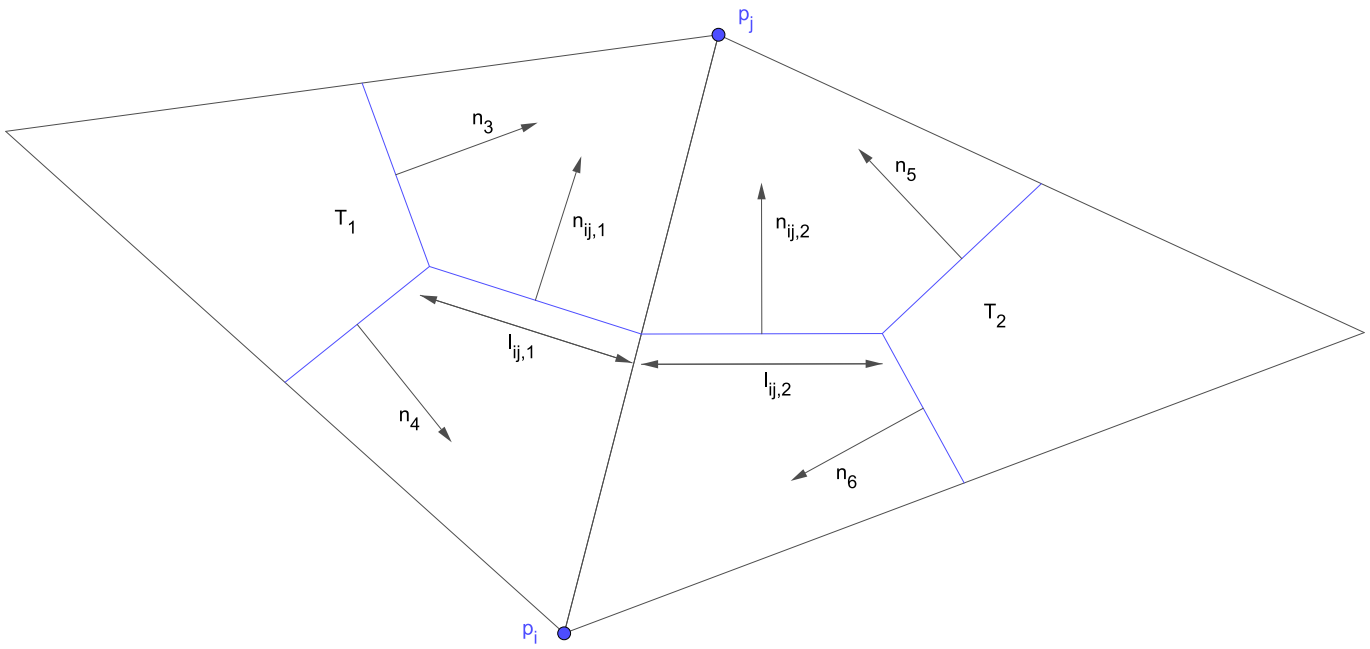

Figure 4.3: Notation for Theorem 4.12

$$
\begin{aligned}
2(b\cdot\nabla\varphi_i,\varphi_j) &= (b\cdot\nabla\varphi_i,\varphi_j)-(b\cdot\nabla\varphi_j,\varphi_i)\\
&= \sum_{k=1}^{2}\left(b|_{T_k}\cdot\nabla\varphi_i|_{T_k}\frac{|T_k|}{3}-b|_{T_k}\cdot\nabla\varphi_j|_{T_k}\frac{|T_k|}{3}\right)\\
&\overset{(4.32)}{=} \frac{b|_{T_1}}{3}\cdot(-l_{ij,1}n_{ij,1}+l_4n_4-(l_{ij,1}n_{ij,1}+l_3n_3))\\
&\quad+\frac{b|_{T_2}}{3}\cdot(-l_{ij,2}n_{ij,2}+l_6n_6-(l_{ij,2}n_{ij,2}+l_5n_5))\\
&\overset{(4.33)}{=} \frac{b|_{T_1}}{3}\cdot(-2l_{ij,1}n_{ij,1}+l_4n_4-l_3n_3)\\
&\quad+\frac{b|_{T_2}}{3}\cdot(-2l_{ij,2}n_{ij,2}+l_6n_6-l_5n_5)\\
&\overset{(4.33)}{=} -\eta_{ij},
\end{aligned}
\tag{4.39}
$$

where the terms involving T_2 are simply dropped if p_i and p_j are adjacent boundary nodes. Furthermore, the η_{ij} sum up to zero:

$$
\sum_{j\in K(i)}\eta_{ij}=\sum_{T:p_i\in T}\int_{\partial(T\cap C_i)}b\cdot n\;ds=\sum_{T:p_i\in T}\int_{T\cap C_i}\operatorname{div}b\;dx=0, \tag{4.40}
$$

which holds true for boundary nodes p_i only because of the assumption made in (4.34). We thus conclude that

$$
\sum_{j\in K(i)}(b\cdot\nabla\varphi_j,\varphi_i)u_j=\sum_{j\in K(i)}\eta_{ij}\frac{u_i+u_j}{2}. \tag{4.41}
$$

□

Remark 4.13. The technical assumption made in equation (4.34) is not restrictive when treating the homogeneous Dirichlet problem, because then only values u_i for interior nodes p_i are non-trivial and therefore the boundary integral vanishes for all nodes of interest. However, for different boundary conditions like no-flux or homogeneous Neumann boundary conditions, this condition amounts to assuming that $b \cdot n = 0$ on $\partial\Omega$ and that therefore the convective part is skew-symmetric. We shall see that we need to make this assumption if we want to upwind our finite element scheme by algebraic manipulations using only the information given by the convective part of the stiffness matrix. △

If – similar to the one-dimensional case – we reinterpret the vector $u \in \mathbb{R}^N$ as the values of a function piecewise constant over the dual cells, we see from Theorem 4.12 that the convective part C of the stiffness matrix can be interpreted as being discretised by a central flux. Then we can mimick the strategy that seemed to work in 1D and use an upwind flux instead, hoping that the resulting modified Galerkin scheme will be a non-oscillatory one. Hence we replace the terms $\eta_{ij}(u_i + u_j)/2$ in (4.35) by

$$\begin{cases} \eta_{ij} u_i & \text{for } \eta_{ij} \geq 0 \\ \eta_{ij} u_j & \text{for } \eta_{ij} \leq 0. \end{cases} \tag{4.42}$$

Proposition 4.14. *Let Assumption 4.9 hold and $C \in \mathbb{R}^{N\times N}$ be the full convective matrix, $c_{ij} = (b \cdot \nabla\varphi_j, \varphi_i)$. Let $\mathcal{I} \subset \{1, \dots, N\}$ be the set of indices of non-Dirichlet nodes, i.e. $u_j = 0$ holds a priori for all $j \notin \mathcal{I}$. Assume that* (4.34) *holds for all $i \in \mathcal{I}$. Let $Y \in \mathbb{R}^{N\times N}$ be given by*

$$y_{ij} = \begin{cases} -2 \sum_{\substack{j\in K(i) \\ c_{ij}\leq 0}} c_{ij} & \text{for } i = j \\ -|c_{ij}| & \text{for } i \neq j. \end{cases} \tag{4.43}$$

Then the upwinded version of the convective part $C^\circ = C_{\mathcal{I}\mathcal{I}} \in \mathbb{R}^{M\times M}$ of the Galerkin stiffness matrix should be defined as $(C + Y)^\circ = (C + Y)_{\mathcal{I}\mathcal{I}}$. △

Proof. Due to homogeneous boundary conditions, we have some $u \in \mathbb{R}^N$ with $u_j = 0$ for $j \notin \mathcal{I}$ and seek a matrix $\tilde{C} \in \mathbb{R}^{M\times M}$ such that

$$(\tilde{C} u_{\mathcal{I}})_i = \sum_{\substack{j\in K(i) \\ \eta_{ij}\geq 0}} \eta_{ij} u_i + \sum_{\substack{j\in K(i) \\ \eta_{ij}\leq 0}} \eta_{ij} u_j.$$

With the choice $\tilde{C} = (C + Y)_{\mathcal{I}\times\mathcal{I}}$, we obtain exactly that:

$$\begin{aligned} (\tilde{C} u_{\mathcal{I}})_i &= \sum_{j\in\mathcal{I}\setminus\{i\}} (c_{ij} - |c_{ij}|) u_j - 2 \sum_{\substack{j\in K(i) \\ c_{ij}\leq 0}} c_{ij} u_i = \sum_{j\in K(i)} (c_{ij} - |c_{ij}|) u_j - 2 \sum_{\substack{j\in K(i) \\ c_{ij}\leq 0}} c_{ij} u_i \\ &= \sum_{\substack{j\in K(i) \\ c_{ij}\leq 0}} 2 c_{ij} (u_j - u_i) \overset{(4.39)}{=} \sum_{\substack{j\in K(i) \\ \eta_{ij}\leq 0}} \eta_{ij} (u_j - u_i) \overset{(4.40)}{=} \sum_{\substack{j\in K(i) \\ \eta_{ij}\leq 0}} \eta_{ij} u_j + \sum_{\substack{j\in K(i) \\ \eta_{ij}\geq 0}} \eta_{ij} u_i. \end{aligned} \tag{4.44}$$

□

Remark 4.15. The matrix Y from (4.43) satisfies $y_{ij} = y_{ji}$ for all i, j with (4.34) and has vanishing row sums:

$$\sum_{j=1}^{N} y_{ij} = -2 \sum_{\substack{j=1 \\ c_{ij} \leq 0}}^{N} c_{ij} - \sum_{\substack{j=1 \\ j \neq i}}^{N} |c_{ij}| = \sum_{\substack{j=1 \\ c_{ij} \leq 0}}^{N} |c_{ij}| - \sum_{\substack{j=1 \\ c_{ij} \geq 0}}^{N} |c_{ij}| = -\sum_{j=1}^{N} c_{ij} = 0. \tag{4.45}$$

Such matrices are called "discrete diffusion operators" in [Kuz10] and we have just argued that they may be called *upwinding matrices*. △

4.2.2 The Upwind Finite Element Method of Baba and Tabata

In [BT81], Baba and Tabata have developed and analysed an upwinded finite element scheme for the transient convection-diffusion equation with zero-flux boundary conditions:

$$\begin{cases} u_t - \epsilon \Delta u + \operatorname{div}(ub) = f & \text{in } \Omega_T \\ (\epsilon \nabla u - ub) \cdot n = 0 & \text{on } \partial\Omega \times (0,T) \\ u = u_0 & \text{on } \partial\Omega \times \{t = 0\}. \end{cases} \tag{4.46}$$

for a time-independent vector field $b \in C^{0,1}(\Omega_T)$, $f \in C(0,T;L^2(\Omega))$ and $u_0 \in C(\Omega)$. It is in particular defined for arbitray dimension $d \in \mathbb{N}$ on regularly simplicially partitioned domains. The proposed scheme – recast from their original variational formulation into an algebraic one – reads:

$$\begin{cases} m_i \dfrac{u_i^{k+1} - u_i^k}{\tau} + \displaystyle\sum_{j=1}^{N} \epsilon(\nabla\varphi_i, \nabla\varphi_j) u_j^k + \sum_{j=1}^{N} \max\{0, \beta_{ij}\} u_i^k + \min\{0, \beta_{ij}\} u_j^k = (f(\cdot, k\tau), \varphi_i) \\ u^0 = I_h u_0 \end{cases} \tag{4.47}$$

for $i = 1, \ldots, N$, where I_h is the nodal interpolator onto V_h^1 and the β_{ij} are defined for any two adjacent nodes $p_i, p_j \in \mathcal{N}$ and satisfy

$$\beta_{ij} = -\beta_{ji} \tag{4.48}$$

$$|\beta_{ij}| \leq \|b\|_{L^\infty(\Omega)} |\Gamma_{ij}| \tag{4.49}$$

$$\left| \beta_{ij} - \int_{\Gamma_{ij}} b \cdot n \, ds \right| \leq C \|b\|_{W^{1,\infty}(\Omega)} h_T^d, \tag{4.50}$$

where n is the unit outward normal to C_i on Γ_{ij} and $T \in \mathcal{T}$ is a d-simplex containing the edge between p_i and p_j.

They then prove in their Theorem 1.1 a discrete mass conservation of this scheme and that, given $f, u_0 \geq 0$, the solution remains non-negative under the assumptions that $\mathcal{T}_h$ is of acute type and that the CFL-like time-step condition

$$\tau \leq \frac{\kappa^2}{(d+1)\epsilon + c_d \kappa \|b\|_{L^\infty(\Omega)}}, \tag{4.51}$$

holds, where κ is the minimal perpendicular length of all simplices $T \in \mathcal{T}_h$ and c_d is a dimension-dependent constant with, e.g., $c_2 = 4$ and $c_3 = 6$. Their Theorem 1.2 contains the following:

Theorem 4.16 (Baba and Tabata, 1981)**.** *Let $\mathcal{T}_h$ be a family of shape-regular triangulations and let the time-step condition*

$$\tau \leq \frac{2\kappa^2}{(d+1)^2\epsilon}(1-\delta) \tag{4.52}$$

hold, where $\delta \in (0,1)$ is some number independent of h. Assume that the solution u to (4.46) satisfies the regularity condition $u \in Z_1 := C^{1,0.5}(0,T;L^2(\Omega)) \cap C^1(0,T;H^1(\Omega)) \cap C(0,T;H^m(\Omega))$ for $m > d/2$. Then for the error $e_k := u^k - I_h(u(\cdot, k\tau))$ it holds that

$$\max_{k=0,\dots,N_T} \|e^k\|_{L^2(\Omega)} \leq C \|u\|_{Z_1} h \tag{4.53}$$

$$\left(\tau \sum_{k=0}^{N_T-1} \left\| \frac{e^{k+1}+e^k}{2} \right\|^2_{H^1(\Omega)}\right)^{1/2} \leq C \|u\|_{Z_1} h \tag{4.54}$$

with $C = C(\sigma_{\mathcal{T}_h}, \epsilon, \delta, \Omega, d, m, \|b\|_{W^{1,\infty}(\Omega)})$. △

Remark 4.17 (Algebraic upwinding vs the upwinding of Baba and Tabata)**.** Under the additional assumption that $\operatorname{div} b = 0$ and $b \cdot n = 0$ on $\partial\Omega$, the convective matrix $C = (b \cdot \nabla\varphi_j, \varphi_i)_{i,j=1,\dots,N}$ is skew-symmetric. The choice $\beta_{ij} := 2c_{ij}$ therefore satisfies (4.48). The following lemma asserts that also the slight variation (4.49′) of (4.49) and (4.50) hold with this choice. The additional constant $\sigma_{\mathcal{T}}/2$ in (4.49′) has to be accounted for in the time-step condition (4.51) for positivity, changing it in the following way:

$$\tau \leq \frac{\kappa^2}{(d+1)\epsilon + c_d\kappa \|b\|_{L^\infty(\Omega)} \sigma_{\mathcal{T}}/2}, \tag{4.51′}$$

whereas the the difference between (4.49) and (4.49′) can be absorbed in the constant C in the last theorem, and thus Therorem 4.16 still holds true.

The proof of Proposition 4.14 with $\mathcal{I}$ replaced by $\{1,\dots,N\}$ and η_{ij} substituted by β_{ij} then shows that adding the upwinding matrix Y defined there amounts to what can be interpreted as an upwinding method in the sense of the paper of Baba and Tabata. Note that we have now gotten rid of Assumption 4.9 but have required $b \cdot n = 0$ on $\partial\Omega$, an assumption we need to make in order for the boundary terms in (4.34) to vanish.

Indeed, this assumption seems hard to avoid when one tries to define and derive information about the signs of the β_{ij} solely from the convective part of the stiffness matrix (i.e. through the terms c_{ij}), because of the skew-symmetry property required in (4.48) and because one has to infer the signs of an approximation to $\int_{\Gamma_{ij}} b \cdot n \, ds$ exclusively from the c_{ij}, which is then hard to do because in order to obtain something resembling the normal n over Γ_{ij}, one certainy needs both the directions $\nabla\varphi_i$ and $\nabla\varphi_j$, hence both c_{ij} and c_{ji}. But without $b \cdot n = 0$ on $\partial\Omega$, their relation is (for adjacent boundary nodes p_i and p_j) polluted by the boundary term $\int_{\partial\Omega} \varphi_i\varphi_j b \cdot n \, ds$ basically unrelated to what happens on Γ_{ij}.

This indicates that, if the boundary conditions are not purely homogeneous Dirichlet and if $b \cdot n \neq 0$ on $\partial\Omega$, then the manipulation suggested in Proposition 4.14 can no longer be interpreted as upwinding. △

Lemma 4.18. *Let $\Omega \subset \mathbb{R}^d$, $d = 2$, be partitioned by a regular, shape-regular triangulation, $b \in W^{1,\infty}(\Omega)$ with $\operatorname{div} b = 0$ and let $\pi : W^{1,\infty}(\Omega) \to \mathcal{RT}_0$ be the interpolator onto the lowest order Raviart-Thomas space characterized by*

$$(\pi(b) \cdot n)|_S = \int_S b \cdot n \; ds \tag{4.55}$$

for each simplex side S with outer unit normal n. Let Γ_{ij} and n_{ij} be defined as in Theorem 4.12 and set $\beta_{ij} := 2(b \cdot \nabla\varphi_j, \varphi_i)$. Then it holds that

$$\left| \beta_{ij} - \int_{\Gamma_{ij}} b \cdot n_{ij} \; ds \right| \leq C(d, \sigma_T) h_T^d \, \|b\|_{W^{1,\infty}(\Omega)} \tag{4.56}$$

$$|\beta_{ij}| \leq \frac{\sigma_{\mathcal{T}}}{2} \, \|b\|_{L^\infty(\Omega)} \, |\Gamma_{ij}| \tag{4.49$'$}$$

for all nodes p_i with (4.34) and neighbours p_j, where T is a triangle containing p_i and p_j. △

Proof. We decompose the error into three parts, one of which we already know to be vanishing:

$$\begin{aligned} \left| 2(b \cdot \nabla\varphi_j, \varphi_i) - \int_{\Gamma_{ij}} b \cdot n \; ds \right| \leq & |2((b - \pi b) \cdot \nabla\varphi_j, \varphi_i)| + \underbrace{\left| 2(\pi b \cdot \nabla\varphi_j, \varphi_i) - \int_{\Gamma_{ij}} \pi b \cdot n \; ds \right|}_{=0 \text{ by (4.38) and (4.39)}} \\ & + \left| \int_{\Gamma_{ij}} (\pi b - b) \cdot n \; ds \right|. \end{aligned} \tag{4.57}$$

An L^∞ estimate of $\varphi_i \nabla\varphi_j$ on $\Omega_i \cap \Omega_j$ gives

$$|((b - \pi b) \cdot \nabla\varphi_j, \varphi_i)| \leq C(\sigma_{\mathcal{T}}) h_{T_k}^{-1} \, \|b - \pi b\|_{L^1(\Omega_i \cap \Omega_j)} \tag{4.58}$$

and Corollary A.2 allows us to estimate

$$\|b - \pi b\|_{L^1(\Gamma_{ij,k})} \leq C(d, \sigma_{\mathcal{T}}) \left(h_{T_k}^{-1} \, \|b - \pi b\|_{L^1(T_k)} + \|\nabla(b - \pi b)\|_{L^1(T_k)} \right), \tag{4.59}$$

where $T_k, k = 1, 2$ are the triangles constituting $\Omega_i \cap \Omega_j$. In order to employ the Bramble-Hilbert type Lemma A.3, we assert that for the standard d-simplex $\hat{T}$, the embedding $W^{1,\infty}(\hat{T}) \hookrightarrow W^{1,1}(\hat{T})$ holds and that the Raviart-Thomas interpolator on $\hat{T}$ is bounded:

$$\left\| \pi_{\hat{T}}(v) \right\|_{W^{1,1}(\hat{T})} \leq C(d) \max_{i=1,\dots,d+1} \left| \int_{S_i} v \cdot n \; ds \right| \leq C(d) \, \|v\|_{W^{1,1}(\hat{T})} \leq C(d) \, \|v\|_{W^{1,\infty}(\hat{T})} \tag{4.60}$$

by norm-equivalence in finite-dimensional spaces, the trace inequality (Theorem A.1) and the embedding. Also, $\pi_{\hat{T}}$ is the identity on $\mathbb{P}^0(\hat{T})^d$. Now Lemma A.3 yields for any $T \in \mathcal{T}$

$$\|b - \pi b\|_{L^1(T)} \leq C(d, \sigma_{\mathcal{T}}) h_T^{d+1} |b|_{W^{1,\infty}(T)} \tag{4.61}$$

$$\|\nabla(b - \pi b)\|_{L^1(T)} \leq C(d, \sigma_{\mathcal{T}}) h_T^d |b|_{W^{1,\infty}(T)}. \tag{4.62}$$

Collecting all the estimates, the first assertion follows.

For the second estimate, note that

$$\left| \int_{\Omega_i \cap \Omega_j} b \cdot \nabla \varphi_j \varphi_i \, dx \right| \leq \frac{\|b\|_{L^\infty(\Omega)}}{3} \left(|T_1| |\nabla \varphi_j|_{T_1}| + |T_2| |\nabla \varphi_j|_{T_2}| \right). \tag{4.63}$$

For a node p_r of T_k let $S_{r,k}$ be the side of T_k opposite this node and $h_{r,k}$ the height of T_k perpendicular to $S_{r,k}$, for $k = 1, 2$. Then it holds that $|\nabla \varphi_r|_{T_k}| = h_{r,k}^{-1}$ and therefore

$$|T_k| |\nabla \varphi_j|_{T_k}| = \frac{|T_k|}{h_{j,k}} = \frac{1}{2} |S_{j,k}| \leq \frac{1}{2} h_{T_k} = \frac{1}{2} \sigma_{T_k} \rho_{T_k} = \frac{1}{2} \sigma_{T_k} \frac{\rho_{T_k}}{|\Gamma_{ij,k}|} |\Gamma_{ij,k}| \tag{4.64}$$

and we can estimate the last quotient by

$$\frac{\rho_{T_k}}{|\Gamma_{ij,k}|} = 3 \frac{\rho_{T_k}}{m_{r,k}} \leq 3 \frac{\rho_{T_k}}{h_{r,k}} \leq \frac{3}{2}, \tag{4.65}$$

where $m_{r,k}$ is the length of the median of the third node p_r in T_k and the last inequality holds because the incircle of T_k is completely contained in T_k and tangential to all sides. It follows that

$$|\beta_{ij}| \leq \frac{2}{3} \|b\|_{L^\infty(\Omega)} \frac{3}{4} \sigma_{\mathcal{T}} |\Gamma_{ij}| = \frac{\sigma_{\mathcal{T}}}{2} \|b\|_{L^\infty(\Omega)} |\Gamma_{ij}|, \tag{4.66}$$

which proves the second inequality. □

Remark 4.19 (Total variation on triangular meshes and the LED property)**.** In [Jam95], Jameson makes the important observation that for $r = 1$ the total variation

$$\mathrm{Var}_p^r(v) = \left(\int_\Omega \|\nabla v\|_p^r \, dx \right)^{1/r} \tag{4.67}$$

loses its qualification as a measure of oscillation in the two-dimensional case on triangular meshes, which he shows by comparing this term for the two continuous piecewise linear functions displayed in Figure 4.4 (triangles have side lengths 1) and $p = 1, 2, \infty$. If we denote the left function by v_1 and the right one by v_2, we obtain the results

r	p	$\mathrm{Var}_p^r(v_1)$	$\mathrm{Var}_p^r(v_2)$
1	1	$4 + 2\sqrt{3}$	$6 + \sqrt{3}$
	2	6	7
	∞	$2 + 2\sqrt{3}$	$5 + \sqrt{3}$
2	2	$(4\sqrt{3})^{1/2}$	$\left(\frac{14}{3}\sqrt{3}\right)^{1/2}$.

Our result here differs from Jameson's for $p = \infty$ and v_2, where he states $5 + 3\sqrt{3}$, so we carry out in detail the computation for this case.

Let the 14 triangles be indexed like the entries of a (2×7)-matrix. Then it suffices to compute the area of these triangles and $\|\nabla v_2|_{T_{11}}\|_\infty$ and $\|\nabla v_2|_{T_{12}}\|_\infty$, since the gradient vectors on all other triangles are mirror images of the aforementioned gradients across the coordinate axes and the ∞-norm is invariant under these particular reflections.

All triangles have height $h = \frac{\sqrt{3}}{2}$ and therefore area $|T| = \frac{\sqrt{3}}{4}$. The vector $\nabla v_2|_{T_{12}}$ has no x-component, so that

$$\|\nabla v_2|_{T_{12}}\|_\infty = \|\nabla v_2|_{T_{12}}\|_2 = \frac{1}{h} = \frac{2}{\sqrt{3}}. \tag{4.68}$$

The ∞-norm of $\nabla v_2|_{T_{11}}$ clearly is the modulus of its x-component, so that

$$\|\nabla v_2|_{T_{11}}\|_\infty = \cos(30°)\ \|\nabla v_2|_{T_{12}}\|_\infty = \frac{\sqrt{3}}{2} \cdot \frac{2}{\sqrt{3}} = 1. \tag{4.69}$$

Summing up, we obtain

$$\mathrm{Var}^1_\infty(v_2) = \int_\Omega \|v_2\|_\infty \, dx = \frac{\sqrt{3}}{4}\left(10 \cdot \frac{2}{\sqrt{3}} + 4 \cdot 1\right) = 5 + \sqrt{3}. \tag{4.70}$$

Be that as it may, the total variation of v_2 is greater than that of v_1 in all considered cases (we have added $r = p = 2$), although v_2 certainly oscillates less than v_1.

Jameson then proposes the *local extremum diminishing (LED)* property (requiring that local extrema not be accentuated and no local extrema be created) as a suitable property that is readily applicable to scalar functions on domains of any dimension. The following observation shows that this property implies the TVD property in 1D:

Let $I = (\alpha, \beta) \subset \mathbb{R}$ be an interval, $N \in \mathbb{N}$, $\mathcal{T}$ a triangulation of I with vertices $\alpha = x_1 < \cdots < x_N = \beta$ and $u \in V^1(\mathcal{T})$. It is $\mathcal{I} = \{2, \ldots, N-1\}$ and we set

$$\mathcal{I}_{\max} := \{i \in \mathcal{I} : u(x_i) \text{ is a local maximum}\}, \quad \mathcal{I}_{\min} := \{i \in \mathcal{I} : u(x_i) \text{ is a local minimum}\}. \tag{4.71}$$

Then for the total variation it holds

$$\mathrm{TV}(u) = \sum_{i=2}^{N} |u(x_i) - u(x_{i-1})| = 2\left(\sum_{i \in \mathcal{I}_{\max}} u(x_i) - \sum_{i \in \mathcal{I}_{\min}} u(x_i)\right) + \sigma_\alpha u(\alpha) + \sigma_\beta u(\beta), \tag{4.72}$$

where

$$\sigma_\alpha = \begin{cases} -1 & \text{if } u(\alpha) \le u(x_2) \\ 1 & \text{else} \end{cases} \qquad \sigma_\beta = \begin{cases} -1 & \text{if } u(\beta) \le u(x_{N-1}) \\ 1 & \text{else.} \end{cases} \tag{4.73}$$

This shows that a scheme which diminishes all existing extrema and does not create new ones also diminishes the total variation. △

4.3 LED conditions for semi-discrete problems

The content and proofs of the following are (with some slight variations and unless other citations are given) based on Chapter 3 of [Kuz10].

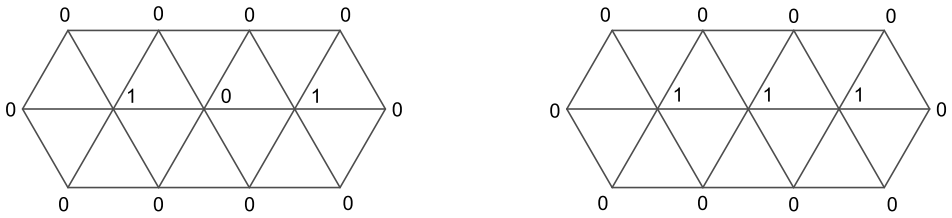

Figure 4.4: Two piecewise linear functions; the left one has lower total variation

Definition 4.20 (Matrices of non-negative type)**.** Let $K \in \mathbb{R}^{N\times N}$ be a matrix and $\mathcal{J} \subset \{1,\dots,N\}$ a set of indices. Then we call K *of $\mathcal{J}$-non-negative type*, if

$$\sum_{j=1}^{N} k_{ij} = 0 \quad \text{for all } i \in \mathcal{J} \text{ and} \tag{4.74}$$

$$k_{ij} \geq 0 \quad \text{for all } i \in \mathcal{I}, j \in \{1,\dots,N\} \setminus \{i\} \tag{4.75}$$

holds and simply *of non-negative type* if $\mathcal{J} = \{1,\dots,N\}$. △

Definition 4.21 (Semi-discrete LED scheme)**.** Consider the semi-discrete scheme in algebraic form

$$M\frac{du}{dt} = Ku + r. \tag{4.76}$$

Then – motivated by the properties of classical solutions to the transient convection-diffusion equation shown in Proposition 2.18 – we call the scheme *local extremum diminishing (LED)* if

$$\frac{du_i}{dt} \leq 0 \quad \text{whenever } i \in \mathcal{I}, u_i \geq \max_{j\in K(i)} u_j \text{ and } r_i \leq 0, \tag{4.77}$$

$$\frac{du_i}{dt} \geq 0 \quad \text{whenever } i \in \mathcal{I}, u_i \leq \min_{j\in K(i)} u_j \text{ and } r_i \geq 0. \tag{4.78}$$

△

Theorem 4.22. *Let u, M, K and r define semi-discrete scheme as in* (4.76)*. Then if $M = \operatorname{diag}(m_1,\dots,m_N)$ is a diagonal matrix with positive diagonal entries and $\sum_{j=1}^{N} k_{ij} = 0$ for $i \in \mathcal{I}$, the scheme* (4.76) *is LED if and only if K is of $\mathcal{I}$-non-negative type.* △

Proof. Assume first $k_{ij} \geq 0$ for all $i \in \mathcal{I}$ and $j \in \{1,\dots,N\} \setminus \{i\}$. Then if $u_i \geq \max_{j\in K(i)} u_j$ and $r_i \leq 0$ for some $i \in \mathcal{I}$, we obtain because of the zero row sum property

$$\frac{du_i}{dt} = \frac{1}{m_i}\left(\sum_{j=1}^{N} k_{ij}u_j + r_i\right) = \underbrace{\frac{1}{m_i}}_{\geq 0}\left(\sum_{j\in K(i)} \underbrace{k_{ij}}_{\geq 0}\underbrace{(u_j - u_i)}_{\leq 0} + \underbrace{r_i}_{\leq 0}\right) \leq 0 \tag{4.79}$$

and analogously $\frac{du_i}{dt} \geq 0$ for local minima and $r_i \geq 0$. To see that condition (4.75) is also necessary, assume that $k_{ij_0} < 0$ for some $j_0 \neq i$. Then the following situation is possible: u has a local maximum at p_i, $i \in \mathcal{I}$, for some $t > 0$ and $u_{j_0} - u_i < 0$ is arbitrarily large in modulus while u_j remains bounded for $j \in K(i) \setminus \{j_0\}$. This destroys (4.79). □

Definition 4.23 (Delaunay triangulation)**.** Let $\Omega \subset \mathbb{R}^d$ with $d \in \{2,3\}$ be a polygonal domain triangulated by $\mathcal{T}$. Then $\mathcal{T}$ is called a *Delaunay triangulation* if $(C_T)^\circ \cap \mathcal{N} = \emptyset$ for all $T \in \mathcal{T}$, where C_T is the (filled, i.e. d-dimensional) circumsphere of T. △

Lemma 4.24 (Characterisation and Properties of the 2D Delaunay triangulation, [Bar92, Section 3.2])**.** *For a planar domain $\Omega \subset \mathbb{R}^2$ triangulated by $\mathcal{T}$ the following are equivalent:*

(i) $\mathcal{T}$ is a Delaunay triangulation.

(ii) For any two adjacent triangles $T_1, T_2 \in \mathcal{T}$ with $T_1 = \operatorname{conv}\{p_k, p_i, p_j\}$ and $T_2 = \operatorname{conv}\{p_i, p_j, p_l\}$ it holds that

$$\angle p_j p_k p_i + \angle p_j p_l p_i \leq 180^\circ. \tag{4.80}$$

Furthermore, out of all triangulations of a given point set, a Delaunay triangulation maximises $\min_{T\in\mathcal{T}} \angle_{\min}(T)$ and minimises $\max_{T\in\mathcal{T}} \angle_{\max}(T)$. △

The following assertion from [Bar92, page 48] showcases the significance of the Delaunay triangulation for the discrete Laplacian to be of non-negative type.

Proposition 4.25. *Let $\Omega \subset \mathbb{R}^2$ be triangulated by $\mathcal{T}$ and $\mathcal{N} = \{p_i : i = 1, \dots, N\}$. Then the discrete Laplace operator $\Delta := -D = -(\nabla\varphi_j, \nabla\varphi_i)_{i,j=1,\dots,N}$ satisfies*

$$\delta_{ij} \geq 0 \quad \text{for } i \neq j \tag{4.81}$$

if and only if $\mathcal{T}$ is a Delaunay triangulation. △

Proof. Let $p_i, p_j \in \mathcal{N}$ be adjacent nodes and $T = \operatorname{conv}\{p_i, p_j, p_k\}$ and S_r the side opposite p_r, h_r the associated perpendicular height, n_r the associated unit outward normal and $\alpha := \angle p_i p_k p_j$. Then

$$\cos\alpha = -\cos\angle n_i n_j = -n_i \cdot n_j = -\nabla\varphi_i|_T \cdot \nabla\varphi_j|_T \; h_i h_j \tag{4.82}$$

and using

$$\sin\alpha = \frac{h_j}{|S_i|} \quad \Rightarrow \quad h_i h_j = \sin\alpha \; h_i |S_i| = 2\sin\alpha \; |T| \tag{4.83}$$

we obtain

$$\nabla\varphi_i|_T \cdot \nabla\varphi_j|_T = -\frac{\cos\alpha}{2\sin\alpha \; |T|} = -\frac{\cot\alpha}{2|T|} \tag{4.84}$$

and thus, if $T' = \operatorname{conv}\{p_i, p_j, p_l\}$ is the triangle sharing the side $S_k = \overline{p_i p_j}$ with T and $\beta = \angle p_i p_l p_j$

$$\delta_{ij} = -(\nabla\varphi_j, \nabla\varphi_i) = \frac{1}{2}\left(\cot\alpha + \cot\beta\right). \tag{4.85}$$

Then arguing as in [Bar92, page 48], we see

$$\delta_{ij} = \frac{1}{2}\frac{\sin(\alpha+\beta)}{\sin\alpha\sin\beta}, \tag{4.86}$$

which is non-negative if and only if $\alpha + \beta \leq 180^\circ$, which is the case for any two triangles sharing a side precisely if and only if $\mathcal{T}$ is Delaunay, as stated in Lemma 4.24. □

Remark 4.26. Unfortunately, for $d = 3$, $\mathcal{T}$ being a Delaunay triangulation no longer guarantees that property (4.81) holds, which Barth shows by means of a counterexample in [Bar92, page 48 ff.]. △

A simple but more restrictive sufficient condition can be given for the discrete Laplacian to satisfy (4.81) in any dimension d:

Definition 4.27 (Non-obtuseness and acuteness)**.** Let $d \in \mathbb{N}$, $\mathcal{T}$ a collection of d-simplices, $T = \operatorname{conv}\{q_1, \dots, q_{d+1}\} \in \mathbb{R}^d$ a particular d-simplex, $f_i = \operatorname{conv}\{q_1, \dots, q_{i-1}, q_{i+1}, \dots, q_{d+1}\}$ the facet opposite q_i and n_i its outer unit normal. Then T is called *non-obtuse* if

$$n_i \cdot n_j \leq 0 \quad \text{for all } i, j \in \{1, \dots, d+1\}, i \neq j \tag{4.87}$$

or equivalently

$$\alpha_{ij} := \arccos(n_i \cdot n_j) \geq \frac{\pi}{2} \quad \text{for all } i, j \in \{1, \dots, d+1\}, i \neq j \tag{4.88}$$

holds, where the α_{ij} are the *(exterior) dihedral angles.*

T is called *acute* if the inequality is strict for all pairs of facets. The mentioned properties can be assigned to $\mathcal{T}$ by requiring them to hold for all $T \in \mathcal{T}$. △

Apparently, Ciarlet and Raviart [CR73, page 23 f.] were the first to propose the following result (although they do not speak about angles and use a strenghtened condition to deal with an additional reaction term):

Proposition 4.28. *Let $d \geq 2$, $\Omega \subset \mathbb{R}^d$ be triangulated by $\mathcal{T}$ and assume all $T \in \mathcal{T}$ are non-obtuse. Then* (4.81) *holds.* △

Proof. Let $i, j \in \{1, \dots, N\}$, $i \neq j$ and $T \subset \operatorname{supp}(\nabla\varphi_i \cdot \nabla\varphi_j)$. Then p_i and p_j are adjacent nodes. In local coordinates $T = \operatorname{conv}\{q_1, \dots, q_{d+1}\}$, w.l.o.g. with $q_1 = p_i$ and $q_2 = p_j$. Let f_k, n_k, h_k for $k = 1, \dots, d+1$ be the facets, outer unit normals and perpendicular heights associated to vertex q_k. Recall that $\nabla\varphi_i|_T = -h_1^{-1} n_1$ and $\nabla\varphi_j|_T = -h_2^{-1} n_2$ and that this implies that $\operatorname{sign}(\nabla\varphi_i|_T \cdot \nabla\varphi_j|_T) = \operatorname{sign}(n_1 \cdot n_2)$. Therefore

$$\delta_{ij} = -(\nabla\varphi_i, \nabla\varphi_j) = -\sum_{T \in \mathcal{T}} \int_T \underbrace{\nabla\varphi_i|_T \cdot \nabla\varphi_j|_T}_{\leq 0} \, dx \geq 0. \tag{4.89}$$

□

Remark 4.29 (Restrictiveness of non-obtuse and acute triangulations)**.** Especially acute triangulations are so restrictive that they need not even exist in higher dimensions. Also, given an initital triangulation, local or even global mesh refinement maintaining non-obtuseness can require special conditions on the initial mesh. Here we give some results about acute triangulations mentioned or proved in [KPP12]:

- Every n-gon in $\mathbb{R}^2$ has a triangulation into $\mathcal{O}(n)$ acute triangles.
- The 3-cube has an acute triangulation.

- The d-cube cannot be acutely triangulated for $d \geq 4$ and $\mathbb{R}^d$ cannot be triangulated for $d \geq 5$.

But even in two dimensions, while a global red refinement of all triangles in $\mathcal{T}$ into four similar triangles does not change the occurring angles, a simple red-green local refinement step can introduce obtuse angles.

In [KK05] the authors use so-called *path tetrahedra* to take on the task of global and local refinement of non-obtuse triangulations in three dimensions. A path tetrahedron is a tetrahedron in which there exist three edges that form a non-closed path and are mutually orthogonal. They show that in 3D a red refinement of a single tetrahedron usually introduces obtuse interior dihedral angles and propose a global refinement into a new regular triangulation of so-called *path tetrahedra* under the condition that all $T \in \mathcal{T}$ contain their circumcenter and have non-obtuse triangles as their faces. For their proposed non-obtuse local refinement around a vertex of a cluster of tetrahedra containing that vertex they need to assume that one of the tetrahedra of the unrefined cluster is already a path tetrahedron and that all the tetrahedra of that cluster are mirror images of another tetrahedron of the cluster.

In summary, without going into much detail, we see that even the concept of non-obtuse triangulations seems rather restrictive when the task is to triangulate complex domains in two or three dimensions and to refine these meshes globally or locally such that non-obtuseness is preserved. $\triangle$

Now we recall that we intend to solve the convection-diffusion equation with homogeneous Dirichlet boundary conditions. Since in this case the boundary values are vanishing a priori, one needs only to compute the values of the discrete solution at interior nodes. A clean way of doing this in the standard Galerkin approach is to restrict the mass and stiffness matrix so that only entries corresponding to pairs of inner nodes remain. More specifically, let $\mathcal{N} = \{p_1, \ldots, p_N\}$ and let $\mathcal{I}$ be the index set of interior nodes, $\#\mathcal{I} = M$. Then the semi-discrete standard Galerkin problem is to find $u \in V_0^1(\mathcal{T})$ such that

$$(u_t, v) + a(u, v) = (f, v) \quad \text{for all } v \in V_0^1(\mathcal{T}), \tag{4.90}$$

where $a(u, v) = \epsilon(\nabla u, \nabla v) + (b \cdot \nabla u, v)$.

Remark 4.30. We remember that, since $\{\varphi_i : i = 1, \ldots, N\}$ is a partition of unity on Ω, D and C have zero row sums. $\triangle$

Then problem (4.90) reads in algebraic formulation, where $u_{\mathcal{I}} \in \mathbb{R}^M$ are the coordinates of the discrete solution with respect to the basis $\{\varphi_i : i \in \mathcal{I}\}$ of $V_0^1(\mathcal{T})$:

$$(M_C)_{\mathcal{I}\mathcal{I}} \frac{du_{\mathcal{I}}}{dt} + (D + C)_{\mathcal{I}\mathcal{I}} u_{\mathcal{I}} = (f, \varphi_i)_{i \in \mathcal{I}}. \tag{4.91}$$

Now on the way towards an LED version of this problem, there are two questions:

1.) How to obtain the lumped mass matrix? Two obvious possible ways to do this would be:

   a) to lump the row entries of the full matrix M_C into the diagonal to obtain M_L and then restrict to $(M_L)_{\mathcal{I}\mathcal{I}}$

b) or to lump the row entries of $(M_C)_{\mathcal{I}\mathcal{I}}$ into the diagonal, i.e. to first restrict and then lump.

2.) If a diffusion or upwinding matrix Y has to be applied to the stiffness matrix for the semi-discrete scheme to become LED, the same question arises. Should the order be

a) to first add such a matrix to $K = -(D + C)$ and then restrict to $K_{\mathcal{I}\mathcal{I}}$

b) or to add such a matrix to $K_{\mathcal{I}\mathcal{I}}$?

As for the first question, answer a) seems to be the right one, because if we choose b), we lose the property that the diagonal entries of the lumped mass matrix equal the area or volume of the associated barycentric dual cell (see Lemma 4.8) for nodes adjacent to boundary nodes. This was, however, part of the interpretation of the upwind finite element method as a finite volume scheme. As for the second question, let us recall Proposition 4.14 and Remark 4.15, where we made the case for the order given in a) when upwinding by adding the matrix Y. In this case of algebraic upwinding, the order given in a) ensures that the interior barycentric dual cells receive their convective contribution from the surrounding cells which are upwind, even if these cells are boundary dual cells. In the context of adding a diffusion matrix to the stiffness matrix to guarantee the LED criterion from Theorem 4.22, proceeding in this order guarantees that the zero boundary values are used in determining whether u has a local extremum at a node next to the boundary; see the following proof for this.

We can now restate Theorem 4.22 for the homogeneous Dirichlet problem:

Theorem 4.31. *Let the domain $\Omega \subset \mathbb{R}^d$ be triangulated by $\mathcal{T}$, the node set be numbered $\mathcal{N} = \{p_i : i = 1, \dots, N\}$, $\mathcal{I}$ the index set of $\mathcal{N}^\circ$, $\#\mathcal{I} = M$, and denote for $u \in V_0^1(\mathcal{T})$ by $u \in \mathbb{R}^M$ also its coordinates with respect to the standard nodal basis. Let $M_L = \operatorname{diag}(m_1, \dots, m_M)$ with $m_i > 0$, $K \in \mathbb{R}^{N\times N}$ with $k_{ij} = 0$ for $j \notin K(i)$, $\sum_{j=1}^N k_{ij} = 0$ for $i \in \mathcal{I} \subset \{1, \dots, N\}$ and $r \in \mathbb{R}^M$. Then the semi-discrete scheme*

$$M_L \frac{du}{dt} = K_{\mathcal{I}\mathcal{I}} u + r \tag{4.92}$$

is LED if and only if K is of $\mathcal{I}$-non-negative type. △

Proof. We repeat the proof of Theorem 4.22 setting $u_j = 0$ for $j \notin \mathcal{I}$. Assume first $k_{ij} \geq 0$ for all $i \in \mathcal{I}$ and $j = 1, \dots, N$, $j \neq i$. Then if $u_i \geq \max_{j\in K(i)} u_j$ and $r_i \leq 0$ for some $i \in \mathcal{I}$, we obtain because of the zero row sum property

$$\frac{du_i}{dt} = \frac{1}{m_i}\left(\sum_{j\in\mathcal{I}} k_{ij}u_j + r_i\right) = \frac{1}{m_i}\left(\sum_{j=1}^N k_{ij}u_j + r_i\right) = \underbrace{\frac{1}{m_i}}_{\geq 0}\left(\sum_{j\in K(i)} \underbrace{k_{ij}}_{\geq 0}\underbrace{(u_j - u_i)}_{\leq 0} + \underbrace{r_i}_{\leq 0}\right) \leq 0 \tag{4.93}$$

and analogously $\frac{du_i}{dt} \geq 0$ for local minima and $r_i \geq 0$. The necessity proof can be copied verbatim from the proof of Theorem 4.22. □

If the triangulation $\mathcal{T}$ of the domain $\Omega \subset \mathbb{R}^d$ is such that for the full diffusive matrix D, its negative $-D$ is of positive type, then we see from Proposition 4.14 that upwinding the convective part C algebraically in the way suggested there is sufficient to ensure for the semi-discrete scheme (4.91)

to be LED. We have seen in Proposition 4.25 a sufficient and necessary condition, namely that $\mathcal{T}$ is a Delaunay trianguation, for this to be the case when $d = 2$ and a sufficient but restrictive criterion for any $d \geq 2$ in Proposition 4.28, namely the non-obtuse angle condition on $\mathcal{T}$. We have also mentioned in Remark 4.26 Barth's negative result that a 3D Delaunay triangulation does not guarantee the non-negative off-diagonal entries of $-D$.

Let us now state two dissatisfactory properties of the LED schemes developed thus far.

Remark 4.32 (Oscillations in LED schemes on irregular meshes)**.** LED schemes in $d \geq 2$ are not necessarily free of oscillations. Consider the simple case of a triangulated domain $\Omega \subset \mathbb{R}^2$ containing the triangulated portion displayed in Figure 4.5 which in turn contains a portion of $\{(x, y) \in \mathbb{R}^2 : y = 0\}$ and an initial state $u_0 \in C(\Omega)$ such that $u_0 \equiv 0$ on $\Omega \cap \{y = 0\}$ and that the values -1 and 1 are attained at the nodes above (below) $\{y = 0\}$. Let $b \equiv (1, 0)$.

Then for very small $\epsilon > 0$ and away from $\partial\Omega$, u solving $u_t - \epsilon\Delta u + b \cdot \nabla u = 0$ on Ω_T should display almost pure advection along lines $\{y = \text{const.}\}$ without ripples. The triangulation shown in Figure 4.5 is non-obtuse, so if it is extended non-obtusely to all of Ω, the semi-discrete finite element method with upwinded convective part will be an LED scheme. However, as can be seen from the inclinations of the effective dual cell boundaries (dashed lines) of C_1 and C_2, the upwind advective flux balance into C_1 will be positive and the flux balance into C_2 will be negative, thus allowing a ripple structure along $\Omega \cap \{y = 0\}$ to develop, even though these ripples will not constitute new local extrema.

The reason such ripples can occur in the considered situation of three not quite parallel lines $\{u = -1\}, \{u = 0\}$ and $\{u = 1\}$ is the irregularity of the shown mesh with respect to the direction of b. It allows for fluxes into C_1 and C_2 from nodes that should not have an advective influence. The effective dual cell tops and bottoms would be horizontal for a regular Friedrichs-Keller triangulation on the other hand, thus not allowing polluting fluxes into these cells from the line of constantly valued nodes above and below $\{y = 0\}$. For an advection field b variable in space and general u_0, even a regular mesh could not preclude such an effect.

We conclude that suppressing all conceivable notions of oscillation is somewhere between hard and impossible to reconcile with the desire for developing methods involving spatially variable fields b on geometrically irregular simplicial meshes, at least if these meshes are not specially adapted to b. We therefore carry on with the LED principle. △

Remark 4.33 (Godunov type order barrier)**.** Similar to [HV03, page 59] for 1D constant advection, we can show that in general we run into a *first consistency order barrier* with our linear compact stencil LED schemes.

To show this, we consider the 1D problem $u_t = Lu := \epsilon u_{xx} - bu_x$ in $\Omega = (\alpha, \beta) \subset \mathbb{R}$, where $b > 0$ is a constant. Let $\alpha = x_1 < \cdots < x_N = \beta$ be equidistant grid points defining the triangulation $\mathcal{T}$ with $|T| = h = (c - a)/(N - 1)$ for all $T \in \mathcal{T}$. Then for $i \in \{2, \ldots, N - 2\}$, row i of a semi-discrete

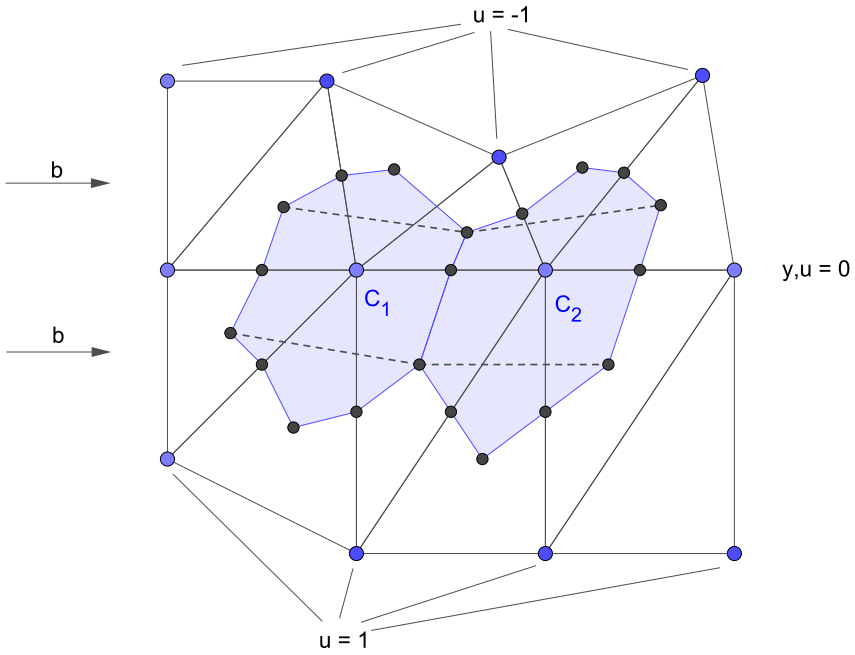

Figure 4.5: A situation where the upwinded finite element method produces oscillations

LED scheme with compact stencil has the form

$$\frac{du_i}{dt} = (L_h u)_i := \sum_{k=-1}^{1} a_k u_{i+k} \tag{4.94}$$

with $\sum_{k=-1}^{1} a_k = 0$ and $a_{-1}, a_1 \geq 0$, where the coefficients α_k do not depend on i because $\mathcal{T}$ is equidistant and ϵ and b are constants. Denote by Pe the cell Péclet number, $Pe = hb/\epsilon$, and let u, u_x, u_{xx} be shorthands for $u(x_i), u_x(x_i), u_{xx}(x_i)$. Assuming the exact solution u is smooth enough, Taylor series expansion gives

$$\begin{aligned}(Lu)(x_i) - \sum_{k=-1}^{1} a_k u(x_{i+k}) &= \epsilon u_{xx} - b u_x - \sum_{k=-1}^{1} a_k \left(u + khu_x + \frac{1}{2}k^2h^2u_{xx} \right) + \mathcal{O}(h^3) \\ &= -\left(\sum_{k=-1}^{1} a_k \right) u - \left(b + h \sum_{k=-1}^{1} k a_k \right) u_x + \left(\epsilon - \frac{h^2}{2} \sum_{k=-1}^{1} k^2 a_k \right) u_{xx} + \mathcal{O}(h^3).\end{aligned} \tag{4.95}$$

The first bracket vanishes as long as $a_0 = -(a_{-1} + a_1)$. For the other two brackets to vanish, we need to solve

$$\begin{pmatrix} h & -h \\ h^2/2 & h^2/2 \end{pmatrix} \begin{pmatrix} a_{-1} \\ a_1 \end{pmatrix} = \begin{pmatrix} b \\ \epsilon \end{pmatrix} \iff \begin{pmatrix} a_{-1} \\ a_1 \end{pmatrix} = \frac{1}{h^3} \begin{pmatrix} h^2/2 & h \\ -h^2/2 & h \end{pmatrix} \begin{pmatrix} b \\ \epsilon \end{pmatrix}. \tag{4.96}$$

Since a_1 must be non-negative, the second equation cannot be satisfied in our case of interest $Pe > 2$ and thus second order consistent approximation of L by L_h is impossible. First order consistency

implies $(a_1 - a_{-1})h = -b$.

The first order limitation gives numerical solutions of LED schemes an overly diffused quality. $\triangle$

The LED schemes we have looked at so far are linear, which makes them unable to adapt their behaviour to the solution they are producing. On the other hand, the necessity part of the proof of Theorem 4.22 and our observations in the introduction suggest that spurious local extrema develop because of large jumps in the solution at neighbouring points, e.g. at layers or spikes. On smoother parts of the solution, the second-order standard Galerkin approximation will usually be well-behaved and therefore superior to a linear LED scheme because of its second order convergence rate. A hybridisation of the two approaches is what we need in order to avoid the excessive diffusion of LED or upwind schemes (manifested in the Godunov order barrier) in solution regions where no spurious effects are created by the standard scheme while fully engaging the LED schemes at places where oscillations would otherwise develop. Since such a hybrid scheme would necessarily have to act upon information about the shape of the solution at the current time-step, it has to be solution-dependent. Hence, such a scheme cannot not be linear, although we are dealing with a linear problem.

5. Flux Corrected Transport

5.1 Zalesak's Original FCT

In [Zal79], Zalesak presents an extension of the *flux corrected transport (FCT)* method invented by Boris and Book [BB73; BBH75; BB76] to multi-dimensional problems in a clear and adaptable fashion. We thus take Zalesak's paper as our entry point to FCT and also take the liberty of adapting his presentation for two-dimensional tensor product grids to general tessellations $\mathcal{C}$ of domains.

So let $\Omega \subset \mathbb{R}^d$ be a domain, $\mathcal{J} := \{1, \dots, N\}$ and $\mathcal{C} = \{C_j : j \in \mathcal{J}\}$ a collection of closed cells $C_j \subset \overline{\Omega}$ such that $C_i^\circ \cap C_j^\circ = \emptyset$ for $i \neq j$ and $\bigcup_{C \in \mathcal{C}} C = \overline{\Omega}$. Set $\Gamma_j := \partial C_j$ for $j \in \mathcal{J}$ and $\Gamma_{ij} := \Gamma_i \cap \Gamma_j$ and call two distinct cells C_i, C_j adjacent if $\Gamma_{ij} \neq \emptyset$. If $\partial C_j \cap \partial\Omega = \emptyset$, we call C_j an interior cell, otherwise a boundary cell. For each boundary cell C_j we introduce a ghost cell $C_j^g \in \mathbb{R}^d \setminus \Omega$ such that $C_j^g \cap C_j = \partial C_j \cap \partial\Omega$ and set $\mathcal{C}^g := \{C_j^g : C_j \text{ is a boundary cell}\}$ and $N_g := \#\mathcal{C}^g$. We number the cells in $\mathcal{C}^g$ as cells $C_{N+1}, \dots, C_{N+N_g}$. This allows us to describe the boundary of all cells as composed of portions shared with another cell: Define $K(j) = \{k \in \{1, \dots, N + N_g\} : C_k \text{ is adjacent to } C_j\}$ for $j \in \mathcal{J}$.

Then Zalesak's version of multi-dimensional FCT applies to schemes in *flux form* that compute numerical solution values u_i^n associated to cell C_i and time step n and have the following form:

$$w_i^{n+1} = w_i^n + \frac{1}{|C_i|} \sum_{j \in K(i)} F_{ij}^n \quad \text{for } i \in \mathcal{J}. \tag{5.1}$$

The F_{ij} are called the inter-cell fluxes and are defined for any pair of adjacent cells/points. They must satisfy $F_{ij} = -F_{ji}$ in order for F_{ij} to signify a transfer of substance from C_j into C_i. The upper index n in F_{ij}^n will be henceforth omitted.

FCT offers a way to combine two separate methods of the form (5.1) that should be designed to solve the same problem but have different properties. These methods in flux form are then determined by two sets of fluxes $\{F_{ij}^L\}$ and $\{F_{ij}^H\}$, the *low order* and the *high order* fluxes. Typically, the methods given by the low order fluxes will be highly diffusive but non-oscillatory (or in our case: LED), while the higher order flux method will – as indicated by the name – be of higher convergence order. FCT proceeds with the following steps, where always $i \in \mathcal{J}, j \in K(i)$:

(1) Compute the fluxes F_{ij}^L and F_{ij}^H.

(2) Compute the *antidiffusive fluxes* $A_{ij} := F_{ij}^H - F_{ij}^L$.

(3) Compute the low order or *transported and diffused* solution

$$u_i^{td} := u_i^L := u_i^n + \frac{1}{|C_i|} \sum_{j \in K(i)} F_{ij}^L. \tag{5.2}$$

(4) Compute the *corrected* or *limited antidiffusive fluxes* $A_{ij}^C := \alpha_{ij} A_{ij}$, where $\alpha_{ij} = \alpha_{ji} \in [0,1]$.

(5) Apply limited antidiffusion:

$$u_i^{n+1} := u_i^{td} + \frac{1}{|C_i|} \sum_{j \in K(i)} A_{ij}^C. \tag{5.3}$$

The flux limiting step (4) is of course the crucial one. The constants $\alpha_{ij} = \alpha_{ji} \in [0,1]$ should be chosen as large as possible but as small as is necessary so that the limited antidiffusive fluxes in step (5) do not cause any unwanted effects in u^{n+1}. The objective is to ensure that $u_i^{n+1} \in [u_i^{\min}, u_i^{\max}]$ for each $i \in \mathcal{J}$ and user-defined bounds $u_i^{\min} \le u_i^{\max}$ whose suitable definition we will come back to in a moment. For each $i \in \mathcal{J}$ define

$$P_i^+ := \sum_{j \in K(i)} A_{ij}^+ \tag{5.4}$$

$$Q_i^+ := |C_i|(u_i^{\max} - u_i^{td})^+ \tag{5.5}$$

$$R_i^+ := \begin{cases} \min\left(1, \frac{Q_i^+}{P_i^+}\right) & \text{if } P_i^+ > 0 \\ 1 & \text{if } P_i^+ = 0 \end{cases} \tag{5.6}$$

$$P_i^- := \sum_{j \in K(i)} A_{ij}^- \tag{5.7}$$

$$Q_i^- := |C_i|(u_i^{\min} - u_i^{td})^- \tag{5.8}$$

$$R_i^- := \begin{cases} \min\left(1, \frac{Q_i^-}{P_i^-}\right) & \text{if } P_i^- < 0 \\ 1 & \text{if } P_i^- = 0, \end{cases} \tag{5.9}$$

where we use the definition $x^+ := \max(0,x) \ge 0$ and $x^- := \min(0,x) \le 0$ and have some changes in comparison to the definitions in [Zal79]: we define P_i^- and Q_i^- with the opposite sign, which has no influence on the quotient Q_i^-/P_i^- and in the second case in R_i^+ and R_i^- we assign the value 1 instead of 0 and instead of requiring a priori that $u_i^{\min} \le u_i^{td} \le u_i^{\max}$ we have added cut-offs in the definitions of Q_i^+ and Q_i^-.

Zalesak's proposed limiting procedure is given by setting for any $i \in \mathcal{J}$ and $j \in K(i)$

$$\alpha_{ij} := \begin{cases} \min(R_i^+, R_j^-) & \text{if } A_{ij} \ge 0 \\ \min(R_j^+, R_i^-) & \text{if } A_{ij} < 0. \end{cases} \tag{5.10}$$

We note that the choice (5.10) ensures the necessary symmetry condition $\alpha_{ij} = \alpha_{ji}$ for $A_{ij} \ne 0$ because $A_{ji} = -A_{ij}$ and that, given $u_i^{\min} \le u_i^{td} \le u_i^{\max}$ for all $i \in \mathcal{J}$, the choices in (5.7) – (5.10)

ensure $u_i^{\min} \le u_i^{n+1} \le u_i^{\max}$ for all $i \in \mathcal{J}$:

$$\begin{aligned}
u_i^{n+1} &= u_i^{td} + \frac{1}{|C_i|} \sum_{j \in K(i)} \alpha_{ij} A_{ij} \\
&= u_i^{td} + \frac{1}{|C_i|} \left(\sum_{j \in K(i): A_{ij} < 0} \alpha_{ij} A_{ij} + \sum_{j \in K(i): A_{ij} > 0} \alpha_{ij} A_{ij} \right) \\
&\in u_i^{td} + \frac{1}{|C_i|} \left[\sum_{j \in K(i): A_{ij} < 0} \alpha_{ij} A_{ij}, \sum_{j \in K(i): A_{ij} > 0} \alpha_{ij} A_{ij} \right] \\
&\subset u_i^{td} + \frac{1}{|C_i|} \left[R_i^- \sum_{j \in K(i): A_{ij} < 0} A_{ij}, R_i^+ \sum_{j \in K(i): A_{ij} > 0} A_{ij} \right] \\
&= u_i^{td} + \frac{1}{|C_i|} \left[R_i^- P_i^-, R_i^+ P_i^+ \right] \subset u_i^{td} + \frac{1}{|C_i|} \left[Q_i^-, Q_i^+ \right] \\
&= [u_i^{\min}, u_i^{\max}].
\end{aligned} \tag{5.11}$$

Remark 5.1 (Preprocessing of antidiffusive fluxes before correction step (4))**.** Zalesak mentions two ways of altering the fluxes A_{ij} before limiting them in step (4). He uses two-dimensional tensor product grids with discrete coordinates (i, j) and denotes the flux from cell (i, j) into cell $(i + 1, j)$ by $A_{i+1/2,j}$, the flux from (i, j) into $(i, j + 1)$ by $A_{i,j+1/2}$ (opposite sign convention!).

- *Cancellation of diffusive "antidiffusive" fluxes:*
  Set $A_{i+1/2,j} = 0$ if

$$\left.\begin{aligned}
& A_{i+1/2,j}(u_{i+1,j}^{td} - u_{i,j}^{td}) < 0 \\
\text{and} \quad & A_{i+1/2,j}(u_{i+2,j}^{td} - u_{i+1,j}^{td}) < 0 \\
\text{or} \quad & A_{i+1/2,j}(u_{i,j}^{td} - u_{i-1,j}^{td}) < 0
\end{aligned}\right\} \tag{5.12}$$

  and $A_{i,j+1/2} = 0$ if

$$\left.\begin{aligned}
& A_{i,j+1/2}(u_{i,j+1}^{td} - u_{i,j}^{td}) < 0 \\
\text{and} \quad & A_{i,j+1/2}(u_{i,j+2}^{td} - u_{i,j+1}^{td}) < 0 \\
\text{or} \quad & A_{i,j+1/2}(u_{i,j}^{td} - u_{i,j-1}^{td}) < 0
\end{aligned}\right\}. \tag{5.13}$$

  This is done to suppress unexpected diffusive behaviour of the antidiffusive fluxes when their direction of mass transport is down gradients. Conditions (5.12) and (5.13) seem to be designed to detect situations where u^{td} is monotone rather than zig-zagging in x- or y-direction, respectively, and where additional diffusion in this respective direction is thus uncalled for. However, Zalesak claims this adjustment to be of cosmetic nature in most cases.

  On irregular meshes, the grid points are no longer aligned along two orthogonal directions and there is no obvious way of generalising these three-piece conditions. One way (called *prelimiting* in [Kuz10]) is to set $A_{ij} = 0$ (we are back to our own notation and sign convection!) whenever $A_{ij}(u_i^{td} - u_j^{td}) > 0$, which amounts to reducing conditions (5.12) and (5.13) to their first inequality.

- *Limiting along coordinate directions first:* If the low order scheme produces a solution u^{td} that is monotonic along an axis parallel grid line and the antidiffusive step destroys that property (this usually happens if there is a large gradient in the solution transverse to this line), then one-dimensional limiting of the fluxes $A_{i+1/2,j}$ and $A_{i,j+1/2}$ along the x- and y-direction can be performed prior to the two-dimensional limiting step in order to keep new extrema in the axis direction restrictions from being accentuated or created.

△

Zalesak's proposed choices for the $u_i^{\min}$ and $u_i^{\max}$ are either

$$u_i^{\min} := \min_{j \in K(i) \cup \{i\}} u_j^n \qquad\qquad u_i^{\max} := \max_{j \in K(i) \cup \{i\}} u_j^n \tag{5.14}$$

or

$$u_i^{\min} := \min_{j \in K(i) \cup \{i\}} u_j^a \qquad\qquad u_i^{\max} := \max_{j \in K(i) \cup \{i\}} u_j^b, \tag{5.15}$$

where $u_j^a := \min(u_j^{td}, u_j^n)$ and $u_i^b := \max(u_j^{td}, u_j^n)$ and the right-hand choice is expected to perform better because it can undo to some extent excessive diffusion developed in u^{td}.

5.2 Intermezzo: M-Matrices

Definition 5.2 (Z-, monotone and M-matrix)**.** Let $A \in \mathbb{R}^{n \times n}$ be a matrix. A is called a *Z-matrix* if $a_{ij} \leq 0$ for all pairs $i \neq j$. A non-singular matrix is called *monotone* if its inverse has non-negative entries, i.e. if $A^{-1} \geq 0$. A monotone Z-matrix is called an *M-matrix*. △

Definition 5.3 (Irreducibility)**.** A matrix $A \in \mathbb{C}^{n \times n}$ is called *irreducible* if there exists *no* permutation matrix $P \in \{0,1\}^{n \times n}$ such that

$$PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \tag{5.16}$$

with square blocks A_{11} and A_{22}. A more usable equivalent definition is given in terms of *directed paths*: For $1 \leq i, j \leq n$, a sequence

$$a_{ik_1}, a_{k_1k_2}, \ldots, a_{k_{r-1}k_r}, a_{k_rj} \tag{5.17}$$

of *non-zero* entries of A is called a directed path from i to j. Now A is irreducible if and only if there exists a directed path connecting i and j for any pair $(i,j) \in \{1, \ldots, n\}^2$. △

Definition 5.4 (Diagonal dominance)**.** A matrix $A \in \mathbb{C}^{n \times n}$ is called *(weakly) diagonally dominant* if

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \tag{5.18}$$

holds for $i = 1, \ldots, n$ and *strictly* so if the inequality is strict for all $i = 1, \ldots, n$. A is called *irreducibly diagonally dominant* if it is an irreducible matrix and (5.18) holds for at least one $i \in \{1, \ldots, n\}$. △

Lemma 5.5. *A strictly or irreducibly diagonally dominant matrix A is non-singular.* △

Proof. This is well-known and a simple consequence of the Gershgorin circle theorem in the strictly diagonally dominant case, see [Var00, Theorem 1.21] for the proof in this case. If A is irreducibly diagonally dominant, then a sharpened version of Gershgorin's theorem, see [Var00, Theorem 1.18], has to be used. □

Theorem 5.6 (Perron-Frobenius, [Var00, Theorem 2.7])**.** *Let $0 \leq A \in \mathbb{R}^{n\times n}$ be an irreducible matrix. Then*

(i) $\rho(A) > 0$ and $\rho(A)$ is an eigenvalue of A.

(ii) There exists $x > 0$ such that $Ax = \rho(A)x$.

(iii) $\rho(A)$ increases strictly when any entry of A is increased.

(iv) $\rho(A)$ is a simple eigenvalue of A. △

Lemma 5.7. *If $0 \leq A \in \mathbb{R}^{n\times n}$ is irreducible, then either*

$$\sum_{j=1}^{n} a_{ij} = \rho(A) \qquad \text{for } i = 1, \ldots, n \tag{5.19}$$

or

$$\min_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} < \rho(A) < \max_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} \tag{5.20}$$

△

Proof. This proof is some simplification of the one of [Var00, Lemma 2.8]. Let $\mathbb{1}$ denote the vector with 1 as each component. Then if all row sums of A are equal to some $\sigma \geq 0$, $A\mathbb{1} = \sigma\mathbb{1}$, $\mathbb{1}$ is an eigenvector with eigenvalue σ and hence $\sigma \leq \rho(A)$. If $\lambda \in \mathbb{C}$ is an eigenvalue of A, then by Gershgorin's theorem

$$|\lambda| - a_{ii} \leq |\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} = \sigma - a_{ii} \qquad \text{for some } i \in \{1, \ldots, n\}, \tag{5.21}$$

which shows $\rho(A) \leq \sigma$ and concludes the proof of (5.19) if all row sums are equal. For the case where the row sums are not all equal to some common value, we confine ourselves to proving the second inequality in (5.20). Since A is irreducible, each row must contain some positive entry. We define a new irreducible $B \geq 0$ by increasing such an entry in all rows k with

$$\sum_{j=1}^{n} a_{kj} < \max_{i=1,\ldots,N} \sum_{j=1}^{n} a_{ij} =: \sigma \tag{5.22}$$

to the extent that all row sums of B are exactly equal to σ. Now we conclude from the case just treated and part (iii) of Theorem 5.6 that $\rho(A) < \rho(B) = \sigma$. □

Theorem 5.8. *Let $A \in \mathbb{R}^{n\times n}$ be a strictly diagonally dominant or irreducibly diagonally dominant Z-matrix with $a_{ii} > 0$ for $i = 1, \dots, n$. Then A is an M-matrix.* △

Proof. From Lemma 5.5 it follows that A is invertible. Define $D := \operatorname{diag}(a_{11}^{-1}, \dots, a_{NN}^{-1})$ and $B := I - DA$. Then

$$b_{ij} = \begin{cases} 0 & \text{if } i = j \\ -\frac{a_{ij}}{a_{ii}} & \text{if } i \neq j, \end{cases} \tag{5.23}$$

which shows $B \geq 0$. If A is strictly diagonally dominant, then

$$\sum_{j=1}^{n} |b_{ij}| < 1 \quad \text{for } i = 1, \dots, n \tag{5.24}$$

shows that $\rho(B) \leq \|B\|_\infty < 1$.

If A is irreducibly diagonally dominant and not strictly diagonally dominant, then B is irreducible, too. To prove this claim, we note that irreducibility of a matrix A depends only on its pattern P, where

$$p_{ij} = \begin{cases} 1 & \text{if } a_{ij} \neq 0 \\ 0 & \text{else.} \end{cases} \tag{5.25}$$

Therefore $X := -DA$ is irreducible. Let $1 \leq i \neq j \leq n$ and let

$$x_{ik_1}, \dots, x_{k_r j} \tag{5.26}$$

be a directed path of non-zero elements from i to j. Then we can eliminate any members of the form x_{kk} from this chain and see that there exists a connected path

$$b_{ik_1}, \dots, b_{k_r j}, \tag{5.27}$$

since $b_{kl} = x_{kl}$ for $k \neq l$. For $j = i$ we take a path without members of the form x_{kk} from i to some auxiliary $l \neq i$ and concatenate it with its reverse path to obtain a directed path of non-zero elements of B from i to i.

Now we infer from

$$\min_{i=1,\dots,n} \sum_{j=1}^{n} b_{ij} < 1 = \max_{i=1,\dots,n} \sum_{j=1}^{n} b_{ij}, \tag{5.28}$$

and Lemma 5.7 that $\rho(B) < 1$ and argue as in the proof of [Var00, Theorem 3.18] to finish the proof: The matrix $DA = I - B$ is invertible because the Neumann series $(I - B)^{-1} = \sum_{k=0}^{\infty} B^k$ converges. All powers of B are non-negative, hence $(I - B)^{-1} \geq 0$. Therefore $A^{-1}D^{-1} \geq 0$ and $A^{-1} \geq 0$. □

5.3 Proposition of a Two-Step FCT Method for the Finite Element Dirichlet Problem

Let $\Omega \subset \mathbb{R}^d$ be triangulated by $\mathcal{T}$, $\#\mathcal{N} = N$ and let $\mathcal{I}$ with $\#\mathcal{I} = M$ be the index set of interior nodes. Take $M_C, M_L, C, D, K \in \mathbb{R}^{N\times N}$ to be as in Definition 0.4 and $Y \in \mathbb{R}^{N\times N}$ to be a discrete

diffusion matrix with zero row sums such that

$$L := K + Y \tag{5.29}$$

has non-negative off-diagonal matrices. Then we have the standard Galerkin and the LED method for the convection-diffusion equation with $f \equiv 0$:

$$M_C^\circ \frac{du^H}{dt} = K^\circ u^H \tag{5.30}$$

$$M_L^\circ \frac{du^L}{dt} = L^\circ u^L. \tag{5.31}$$

Let us propose a two-step FCT method even though the present high and low order schemes are not in flux form, but in which there are antidiffusive fluxes that satisfy the anti-symmetry property. Assume we have computed u^k and want to combine the high order and low order fully discrete solution by a θ-stepping,

$$(M_C - \theta\tau K)^\circ u^{H,k+1} = (M_C + (1-\theta)\tau K)^\circ u^k \tag{5.32}$$

$$(M_L - \theta\tau L)^\circ u^{L,k+1} = (M_L + (1-\theta)\tau L)^\circ u^k \tag{5.33}$$

in order to compute an "in-between" new time-step solution u^{k+1}. Defining the matrices

$$\hat{A} := M_C - \theta\tau K \qquad \hat{B} := M_C + (1-\theta)\tau K \tag{5.34}$$

$$A := M_L - \theta\tau L \qquad B := M_L + (1-\theta)\tau L \tag{5.35}$$

and $f^\circ := (A - \hat{A})^\circ u^{H,k+1} + (\hat{B} - B)^\circ u^k$, the high order step can be equivalently written as

$$A^\circ u^{H,k+1} = A^\circ u^{L,k+1} + f^\circ = B^\circ u^k + f^\circ \tag{5.36}$$

and dropping f° in this equation would yield the low order solution. Now we set

$$f := (A - \hat{A})u^{H,k+1} + (\hat{B} - B)u^k \tag{5.37}$$

and decompose each component of f_i, $i \in \mathcal{I}$, into a sum of internodal fluxes f_{ij}. We have

$$\begin{aligned} f &= (M_L - M_C - \theta\tau Y)u^{H,k+1} + (M_C - M_L - (1-\theta)\tau Y)u^k \\ &= (M_L - M_C)(u^{H,k+1} - u^k) - \tau Y(\theta u^{H,k+1} + (1-\theta)u^k) \end{aligned} \tag{5.38}$$

and thus for $i = 1, \dots, N$

$$\begin{aligned} f_i &= m_i(u^{H,k+1} - u^k)_i - \sum_{j=1}^N m_{ij}(u^{H,k+1} - u^k)_j - \tau \sum_{j=1}^N y_{ij}(\theta u^{H,k+1} + (1-\theta)u^k)_j \\ &= \sum_{\substack{j=1 \\ j\neq i}}^N m_{ij}\left[(u_i^{H,k+1} - u_j^{H,k+1}) - (u_i^k - u_j^k)\right] + \tau y_{ij}\left[\theta(u_i^{H,k+1} - u_j^{H,k+1}) + (1-\theta)(u_i^k - u_j^k)\right] \\ &= \sum_{\substack{j=1 \\ j\neq i}}^N \underbrace{(m_{ij} + \theta\tau y_{ij})(u_i^{H,k+1} - u_j^{H,k+1}) - (m_{ij} - (1-\theta)\tau y_{ij})(u_i^k - u_j^k)}_{=:f_{ij}}, \end{aligned} \tag{5.39}$$

where we set $u_j = 0$ for $j \notin \mathcal{I}$. The symmetry of M_C and $y_{ij} = y_{ji}$ for $i \in \mathcal{I}$, $j \in \{1, \dots, N\}$, $i \neq j$ then immediately gives $f_{ji} = -f_{ij}$. Note that for $\theta = 0$ (and only in that case) this decomposition of f renders (5.36) to have the form (5.3) with $\alpha_{ij} = 1$ for all i, j. For $\theta \neq 0$ we therefore work with solutions multiplied from the left by A° so we do not lose the local nature of antidiffusive fluxes.

If now $\alpha \in [0,1]^{N \times N}$ is symmetric and $F := (f_{ij})_{i,j=1,\dots,N}$ we obtain a flux corrected solution u^{k+1} by solving

$$A^\circ u^{k+1} = A^\circ u^{L,k+1} + \operatorname{diag}(F\alpha)_{\mathcal{I}} = B^\circ u^k + \operatorname{diag}(F\alpha)_{\mathcal{I}}. \tag{5.40}$$

Given values $u_i^{\min}, u_i^{\max}$ for $i \in \mathcal{I}$ we set

$$u^{\min} := (u_i^{\min})_{i \in \mathcal{I}} \qquad u^{\max} := (u_i^{\max})_{i \in \mathcal{I}} \tag{5.41}$$

$$u_A^{\min} := A^\circ u^{\min} \qquad u_A^{\max} := A^\circ u^{\max} \tag{5.42}$$

and then apply Zalesak's limiting strategy to obain suitable α_{ij} to ensure that

$$A^\circ u^{k+1} \in [u_A^{\min}, u_A^{\max}]. \tag{5.43}$$

The matrix A° is an M-matrix (see Lemma 5.9), hence $(A^\circ)^{-1} \geq 0$ (non-negative entries). This implies that

$$u^{k+1} \in [u^{\min}, u^{\max}], \tag{5.44}$$

which was the declared goal. If $\theta = 0$ and thus $A^\circ = M_L^\circ$, these identities are even equivalent.

Lemma 5.9. *The matrix A° (for A defined in (5.35)) is an M-matrix.* △

Proof. We have by definition $A = M_L - \theta\tau L$, where M_L is a positive diagonal matrix, $\theta \geq 0$, $\tau > 0$, L has zero row sums and $l_{ij} \geq 0$ for $i \neq j$. Therefore A is a Z-matrix with positive diagonal and positive row sums and thus strictly diagonally dominant for $\theta > 0$. Restricting A to $A^\circ = A_{\mathcal{I}\mathcal{I}}$ removes some non-positive column entries from each remaining row, thus only increasing the row sums. It follows from Theorem 5.8 that A° is an M-matrix for $\theta > 0$. In the case $\theta = 0$ the matrix A reduces to M_L and the assertion is trivial. □

5.4 The FCT Approach of Kuzmin

In Kuzmin's method, no independent calculation of the high order solution step is performed; rather, the new time step solution is attempted to be attained in a single non-linear step or by a linearised version of a non-linear step.

From now on, f will no longer be a linear source term as in (2.1), but f will stand for fluxes unless locally defined otherwise.

5.4.1 Formal Semi-Discrete Limited Scheme

Disregarding boundary conditions, in Section 4.1 of [Kuz10], Kuzmin operates with the full consistent and lumped mass matrices and uses the following definition for the discrete diffusion matrix:

Definition 5.10 (Kuzmin's discrete diffusion matrix)**.** A discrete diffusion matrix $Y \in \mathbb{R}^{N \times N}$ in the sense of Kuzmin has to satisfy

(i) Y is symmetric with $\sum_{j=1}^{N} y_{ij} = 0$ for $i = 1, \ldots, N$.

(ii) For $L := K + Y$ it holds $l_{ij} \geq 0$ for all $i, j \in \{1, \ldots, N\}$ with $i \neq j$.

Hence the lower bound for y_{ij}, $i \neq j$ is

$$y_{ij} := \max(0, -k_{ij}, -k_{ji}). \tag{5.45}$$

It is this choice for Y that will henceforth be referred to as *Kuzmin's discrete diffusion matrix.* $\triangle$

Remark 5.11. With Kuzmin's discrete diffusion matrix, the whole operator $K = -(C + \epsilon D)$ is manipulated, not just the convective part C. This can be thought of a "minimally invasive" way to obtain an LED semi-discrete scheme; the natural diffusion introduced by ϵD is used to reduce the amount of artificial diffusion that would be introduced by defining Y as the upwinding matrix from (4.43). $\triangle$

The linear high order and low order semi-discrete schemes

$$M_C \frac{du}{dt} = Ku \qquad \text{and} \qquad M_L \frac{du}{dt} = Lu, \tag{5.46}$$

where $L := K + Y$, are connected formally by noticing that

$$M_L \frac{du}{dt} = Lu + f \tag{5.47}$$

returns the high order method for

$$f = (M_L - M_C)\frac{du}{dt} - Yu \tag{5.48}$$

and the low order method for $f = 0$. Therefore, f can be regarded as a sum of unlimited or *raw* antidiffusive fluxes. Kuzmin then suggests the following decomposition of f into *raw antidiffusive internodal fluxes* f_{ij}:

$$f_{ij} := \left(m_{ij} \frac{d}{dt} + y_{ij} \right) (u_i - u_j) \qquad \text{for } i \neq j. \tag{5.49}$$

This decomposition is motivated by the following calculation which makes use of the zero row sum property of $M_L - M_C$ and Y:

$$\begin{aligned} f_i &= \left((M_L - M_C)\frac{du}{dt} \right)_i - (Tu)_i = m_i \frac{du_i}{dt} - \sum_{j=1}^{N} m_{ij} \frac{du_j}{dt} - \sum_{j=1}^{N} y_{ij} u_j \\ &= \sum_{\substack{j=1 \\ j \neq i}}^{N} m_{ij} \frac{d}{dt}(u_i - u_j) - y_{ij}(u_j - u_i) = \sum_{\substack{j=1 \\ j \neq i}}^{N} f_{ij}. \end{aligned} \tag{5.50}$$

As should for a flux, it holds that $f_{ij} = -f_{ji}$ for all $i \neq j$ since M_C and Y are symmetric matrices. Now each such flux can be limited by multiplying f_{ij} by some $\alpha_{ij} \in [0, 1]$, where $\alpha_{ij} = \alpha_{ji}$ should hold in order to maintain antisymmetry. Setting

$$\overline{f}_{ij} := \alpha_{ij} f_{ij} \qquad \text{and} \qquad \overline{f}_i := \sum_{\substack{j=1 \\ j \neq i}}^{N} \overline{f}_{ij} \qquad \text{and} \qquad \overline{f} = (\overline{f}_i)_{i=1,\ldots,N}, \tag{5.51}$$

a new formal limited scheme

$$M_L \frac{du}{dt} = Lu + \overline{f} \tag{5.52}$$

is generated. It should be noted that, with the Zalesak limiter, the α_{ij} will depend non-linearly and non-smoothly on u and f and thus (5.52) cannot be cast into the shape of an explicit ODE in an obvious way. Hence (5.52) may not be a well-posed semi-discrete problem. Nevertheless, a fully discrete version may be solvable.

5.4.2 Fully Discrete Limited Scheme

In [Kuz10, Section 4.4], a full discretisation by means of a θ-stepping is suggested:

$$(M_L - \theta\tau L)u^{n+1} = (M_L + (1-\theta)\tau L)u^n + \tau\overline{f}(u^{n+1}, u^n), \tag{5.53}$$

where $\overline{f}$ is defined as in (5.51) and the differential d/dt in (5.49) is replaced by a difference quotient to define

$$f_{ij} := m_{ij}\frac{(u_i^{n+1} - u_j^{n+1}) - (u_i^n - u_j^n)}{\tau} + y_{ij}\left(\theta(u_i^{n+1} - u_j^{n+1}) + (1-\theta)(u_i^n - u_j^n)\right). \tag{5.54}$$

A limiting strategy for the α_{ij} motivated by Zalesak's work is given by setting

$$P_i^+ := \sum_{j\in K(i)} f_{ij}^+ \tag{5.55}$$

$$Q_i^+ := \frac{m_i}{\tau}(\tilde{u}_i^{\max} - \tilde{u}_i) \tag{5.56}$$

$$R_i^+ := \begin{cases} \min\left(1, \frac{Q_i^+}{P_i^+}\right) & \text{if } P_i^+ > 0 \\ 1 & \text{if } P_i^+ = 0 \end{cases} \tag{5.57}$$

$$P_i^- := \sum_{j\in K(i)} f_{ij}^- \tag{5.58}$$

$$Q_i^- := \frac{m_i}{\tau}(\tilde{u}_i^{\min} - \tilde{u}_i) \tag{5.59}$$

$$R_i^- := \begin{cases} \min\left(1, \frac{Q_i^-}{P_i^-}\right) & \text{if } P_i^- < 0 \\ 1 & \text{if } P_i^- = 0 \end{cases} \tag{5.60}$$

and

$$\alpha_{ij} := \begin{cases} \min(R_i^+, R_j^-) & \text{for } f_{ij} \geq 0 \\ \min(R_i^-, R_j^+) & \text{for } f_{ij} < 0. \end{cases} \tag{5.61}$$

Here, the values involving $\tilde{u}$ are gathered from time steps before time mark $n+1$ and thus the Q_i^+ and Q_i^- are constants with respect to the variable u^{n+1} of the nonlinear problem (5.53).

We might be interested in investigating whether problem (5.53) is well-posed for small $\tau > 0$. We therefore restate it in a way that is equivalent for positive τ but also defined in the limiting case $\tau = 0$. Namely, we consider

$$(M_L - \theta\tau L)u^{n+1} = (M_L + (1-\theta)\tau L)u^n + \overline{f}(u^{n+1}, u^n), \tag{5.62}$$

with (5.51) unchanged,

$$f_{ij} := m_{ij}\left((u_i^{n+1} - u_j^{n+1}) - (u_i^n - u_j^n)\right) + y_{ij}\tau\left(\theta(u_i^{n+1} - u_j^{n+1}) + (1-\theta)(u_i^n - u_j^n)\right) \tag{5.63}$$

and

$$P_i^+ := \sum_{j \in K(i)} f_{ij}^+ \tag{5.64}$$

$$Q_i^+ := m_i(\tilde{u}_i^{\max} - \tilde{u}_i) \tag{5.65}$$

$$R_i^+ := \begin{cases} \min\left(1, \frac{Q_i^+}{P_i^+}\right) & \text{if } P_i^+ > 0 \\ 1 & \text{if } P_i^+ = 0 \end{cases} \tag{5.66}$$

$$P_i^- := \sum_{j \in K(i)} f_{ij}^- \tag{5.67}$$

$$Q_i^- := m_i(\tilde{u}_i^{\min} - \tilde{u}_i) \tag{5.68}$$

$$R_i^- := \begin{cases} \min\left(1, \frac{Q_i^-}{P_i^-}\right) & \text{if } P_i^- < 0 \\ 1 & \text{if } P_i^- = 0. \end{cases} \tag{5.69}$$

as well as (5.61) unchanged.

A natural question which is – unfortunately – not adressed in any of the works of Kuzmin et al. is for which choices of time step $\tau > 0$ the problem (5.62) possesses a solution or even a unique solution. With the Zalesak limiter just presented, this is not trivial because of the nonlinear and nonsmooth nature of the problem. Apart from abstract existence (and maybe uniqueness) results, it would be desirable to have a practical algorithm at hand (such as a fixed point iteration or a Newton method) that can be shown to converge to such solutions. While so-called fixed point iterations with acceleration techniques and Newton methods have been implemented by Kuzmin et al. in [Kuz10] and [MKK07], respectively, and seem to work, the questions of existence and whether (5.62) represents in some form a contraction that would justify using a fixed point algorithm or if convergence criteria for non-smooth Newton methods are met has not been addressed.

5.4.3 An Attempt to Establish Unique Solvability

5.4.3.1 Problem Reformulation

Let us introduce the notation

$$h := u^{n+1} - u^n, \qquad c_{ij}(\tau) := m_{ij} + \theta\tau y_{ij}, \qquad g_{ij}(\tau) := \tau y_{ij}(u_i^n - u_j^n) \tag{5.70}$$

to restate problem (5.62) in slightly simplified form. To this end, we note that we may indeed switch to the variable h because R_i^+ and R_i^- depend on u^{n+1} only through the f_{ij} and each f_{ij} can be written as

$$\begin{aligned} f_{ij}(h,\tau) &= (m_{ij} + \theta\tau y_{ij})\left((u_i^{n+1} - u_i^n) - (u_j^{n+1} - u_j^n)\right) + \tau y_{ij}(u_i^n - u_j^n) \\ &= c_{ij}(\tau)(h_i - h_j) + g_{ij}(\tau). \end{aligned} \tag{5.71}$$

After defining

$$F(h,\tau) := (M_L - \theta\tau L)h - \overline{f}(h,\tau) - \tau L u^n \tag{5.72}$$

we notice that

$$F(0,0) = 0, \tag{5.73}$$

which can be interpreted as u^n being the sought solution at $\tau = 0$, and that solving problem (5.62) for some $\tau > 0$ is equivalent to finding some (unique?) $h(\tau)$ such that

$$F(h(\tau),\tau) = 0. \tag{5.74}$$

This renders the problem in a form that might be amenable to some form of implicit function theorem argument.

5.4.3.2 Implicit Function Theorem for Locally Lipschitz Continuous Functions

The singular goal of this subparagraph it to prove Theorem 5.14.

Theorem 5.12 (Rademacher)**.** *Let $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}^m$ be locally Lipschitz continuous. Then f is differentiable almost everywhere in U.* △

Definition 5.13 (Generalised Jacobian)**.** Let $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}^m$ locally Lipschitz continuous at $x \in U$. Let $D_f \subset U$ be the set of points in which f is differentiable.

(a) The set $\partial_B f(x) := \{G \in \mathbb{R}^{m\times n} : \exists\ (x_n)_{n\in\mathbb{N}} \subset D_f \text{ with } x_n \to x \text{ and } \nabla f(x_n) \to G\}$ is called the *B-* or *Bouligand subdifferential*.

(b) The set $\partial f(x) := \operatorname{conv}(\partial_B f(x))$ is called the *generalised Jacobian* in the sense of Clarke.

(c) If $f : \mathbb{R}^n \times \mathbb{R}^m \supset U \to \mathbb{R}^n$, we denote by $\Pi_x \partial f(x, y)$ the set of projections $G \in \mathbb{R}^{n\times n}$ of elements $[G\ H] \in \partial f(x, y)$. △

Theorem 5.14 (Implicit function theorem for Lipschitz functions, [Hin10, Theorem 2.6])**.** *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be Lipschitz continuous in a neighbourhood of a point $(x_0, y_0) \in \mathbb{R}^n \times \mathbb{R}^m$ for which $f(x_0, y_0) = 0$. Assume that all matrices in $\Pi_x \partial f(x_0, y_0)$ are non-singular. Then there exist open neighbourhoods V_{x_0} of x_0 and V_{y_0} of y_0 such that for every $y \in V_{y_0}$ the equation $f(x, y) = 0$ has a unique solution $x = \varphi(y) \in V_{x_0}$ and in particular $\varphi(y_0) = x_0$. Furthermore, the function $\varphi : V_{y_0} \to V_{x_0}$ is Lipschitz continuous.* △

Lemma 5.15. *For $U \subset \mathbb{R}^n$ open and $f : U \to \mathbb{R}^m$ locally Lipschitz continuous and $x \in U$ the generalised Jacobian $\partial f(x)$ is a non-empty, compact, convex set and the set-valued map ∂f is upper semicontinuous.* △

Proposition 5.16 (Specialised chain rule, [Hin10, Corollary 2.1])**.** *Let U, f, x be as before. Then for any $y \in \mathbb{R}^m$ it holds $\partial(y^T f)(x) = y^T \partial f(x)$.* △

Definition 5.17 (Clarke's generalised directional derivative)**.** Let $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}^m$ a locally Lipschitz continuous function, $x \in U$ and $v \in \mathbb{R}^n$. Then define its *generalised directional derivative* $f^\circ(x; \cdot) : \mathbb{R}^n \to \mathbb{R}^m$ in the sense of Clarke as

$$f^\circ(x; v) := \sup \mathcal{D}_f(x; v), \tag{5.75}$$

where

$$\mathcal{D}_f(x; v) := \left\{ \lim_{n\to\infty} \frac{f(x_n + t_n v) - f(x_n)}{t_n} : x_n \to x \text{ and } t_n \downarrow 0 \right\}. \tag{5.76}$$

Note that $(\mathbb{R}^m, \leq)$ with $x \leq y$ understood elementwise is a vector lattice, hence suprema make sense. △

Lemma 5.18. *Let U, f, x, v be as in the previous definition. Then*

(i) $\mathcal{D}_f(x; v) \subset \mathbb{R}^m$ is non-empty and compact.

(ii) $y^T \mathcal{D}_f(x; v) = D_{y^T f}(x; v)$ for all $y \in \mathbb{R}^m$. △

Proof. For $x_n \to x$ and $t_n \downarrow 0$ the sequence $(f(x_n + t_n v) - f(x_n))/t_n$ remains bounded (uniformly with respect to the choice of sequence) and a convergent subsequence can be selected because $\mathbb{R}^m$ is finite-dimensional. Hence the set is non-empty and bounded. Sequential closedness follows from a diagonal sequence argument:

Let $d_1, d_2, \cdots \in \mathcal{D}_f(x; v)$ and $d \in \mathbb{R}^m$ such that $|d_k - d| \leq 2^{-k}$ for $k \in \mathbb{N}$. We want to show that $d \in \mathcal{D}_f(x; v)$ and thus need to find a sequence (x_n, t_n) with $x_n \to x$, $t_n \downarrow 0$ such that

$$\frac{f(x_n + t_n v) - f(x_n)}{t_n} \to d \quad \text{for } n \to \infty. \tag{5.77}$$

For each $k \in \mathbb{N}$, let x_n^k and $t_n^k \geq 0$ be sequences with $|x_n^k - x|, t_n^k, |q_n^k - d_k| \leq 2^{-n}$, where

$$q_n^k := \frac{f(x_n^k + t_n^k v) - f(x_n^k)}{t_n^k}. \tag{5.78}$$

Then the choice $x_n := x_n^n$ and $t_n := t_n^n$ gives us all the desired properties since $t_n \geq 0$, $|x_n - x|, t_n \leq 2^{-n}$ and

$$|q_n^n - d| \leq |q_n^n - d_n| + |d_n - d| \leq 2^{-n+1} \to 0 \quad \text{for } n \to \infty. \tag{5.79}$$

Assertion (ii) is not hard to see. □

Proposition 5.19 (Generalised gradient = subdifferential, [Cla75, Proposition 1.4])**.** *Let $U \subset \mathbb{R}^n$ be open, $x \in U$ and $g : U \to \mathbb{R}$ locally Lipschitz continuous. Then $g^\circ(x; \cdot)$ is the support function of $\partial g(x)$. More explicitly, for any $v \in \mathbb{R}^n$*

$$g^\circ(x; v) = \max\{u \cdot v : u \in \partial g(x)^T\}, \tag{5.80}$$

$$\partial g(x)^T = \{u \in \mathbb{R}^n : u \cdot w \leq g^\circ(x; w) \ \forall \ w \in \mathbb{R}^n\} \tag{5.81}$$

and therefore

$$\partial g(x) v = [-(-g)^\circ(x; v), g^\circ(x; v)]. \tag{5.82}$$

△

In [Thi82, Section 4], Thibault extends the notion of Clarke's generalised derivative and subdifferential to Hausdorff locally convex vector spaces X and Y and functions $f : X \to Y$ which are Lipschitz at a point $\overline{x} \in X$ in a certain sense such that $\partial(y^* \circ f)(\overline{x}) \in X^*$ is declared for all $y^* \in Y^*$. He then asks the question whether a set-valued mapping $\Delta_f(\overline{x}; \cdot) : X \to CC(Y)$ into the set $CC(Y)$ of non-empty closed convex subsets of Y exists such that

$$\partial(y^* \circ f)(\overline{x}) \cdot v = y^*(\Delta_f(x; v)) \quad \text{for all } y^* \in Y^* \text{ and } v \in X, \tag{5.83}$$

where he defines ∂g for real-valued functions g as the subdifferential, such that the definition reduces to (5.81) for $X = \mathbb{R}^n$. He calls such a mapping – when it exists – the generalized Clarke derivative. The following proposition will show that this introduces no ambiguity with our previous definition of the generalised Jacobian.

In order to avoid the long and technical proofs, let us strip down the generality of Thibault's results to our needs by only considering $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. We know from Proposition 5.16 that a

mapping as in (5.83) exists for locally Lipschitz continuous $f : \mathbb{R}^n \to \mathbb{R}^m$: for $x, v \in \mathbb{R}^n$, simply set $\Delta_f(x; v) := \partial f(x)v$. What is new is that such a set-valued mapping, when it exists, must be unique:

Proposition 5.20 ([Thi82, Proposition 4.1])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous and $\Delta_f : \mathbb{R}^n \times \mathbb{R}^n \to CC(\mathbb{R}^m) \subset \mathcal{P}(\mathbb{R}^m)$ be a set-valued mapping into the set of all non-empty closed convex subsets of $\mathbb{R}^m$ satisfying at each fixed $x \in \mathbb{R}^n$*

$$\partial(y^T f)(x) \cdot v = y^T \Delta_f(x; v) \quad \text{for all } y \in \mathbb{R}^m, v \in \mathbb{R}^n, \tag{5.84}$$

where $\partial g(x) = \{u \in \mathbb{R}^n : u \cdot v \leq g^\circ(x; v) \ \forall \ v \in \mathbb{R}^n\}$ for $g : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz continuous. Then this mapping is unique. △

Proof. Fix $(x, v) \in \mathbb{R}^n \times \mathbb{R}^n$ and assume there exist two different such sets $\Delta_f^1(x; v)$ and $\Delta_f^2(x; v)$. W.l.o.g. there exists $z \in \Delta_f^1(x; v) \setminus \Delta_f^2(x; v)$. By the hyperplane separation theorem (also known as the Hahn-Banach Theorem in the context of locally convex topological vector spaces), there exists $y \in \mathbb{R}^m$ such that

$$y^T z < \inf \left(y^T \Delta_f^2(x; v)\right) \tag{5.85}$$

and thus $y^T \Delta_f^1(x; v)$ and $y^T \Delta_f^2(x; v)$ cannot be both equal to $\partial(y^T \circ f)(x) \cdot v$. □

Proposition 5.21 ([Thi82, Proposition 4.5])**.** *Let $U \subset \mathbb{R}^n$ be open, $f : U \to \mathbb{R}^m$ locally Lipschitz continuous and $x \in \mathbb{R}^n$. Then $\partial f(x)v = \operatorname{conv} \mathcal{D}_f(x; v)$ for any $v \in \mathbb{R}^n$.* △

Proof. Let $y \in \mathbb{R}^m$. Then from Definition 5.17, (5.82) and Lemma 5.18 we obtain

$$\begin{aligned} \partial(y^T f)(x)v &= [-(-y^T f)^\circ(x; v), (y^T f)^\circ(x; v)] = \operatorname{conv} \mathcal{D}_{y^T f}(x; v) = \operatorname{conv} y^T \mathcal{D}_f(x; v) \\ &= y^T \operatorname{conv} \mathcal{D}_f(x; v). \end{aligned} \tag{5.86}$$

The claim follows from closedness (even compactness) of $\mathcal{D}_f(x; v)$ and the uniqueness asserted in Proposition 5.20. □

We are now ready to prove an even more general result than Theorem 5.14, the theorem and its proof being taken from [Kum91]. For the context of this theorem we shall make the following definition:

Definition 5.22 (Regularity)**.** Let $U \subset \mathbb{R}^n \times \mathbb{R}^m$ be open, $f : U \to \mathbb{R}^n$ locally Lipschitz continuous and $(x^*, t^*, a^*) \in U \times \mathbb{R}^n$ such that $f(x^*, t^*) = a^*$. We call f *regular* at (x^*, t^*, a^*) if there exist neighbourhoods $N(x^*) \subset \mathbb{R}^n$ of x^* and $N(a^*, t^*) \subset \mathbb{R}^n \times \mathbb{R}^m$ of (a^*, t^*) and a Lipschitz continuous function $g : N(a^*, t^*) \to N(x^*)$ such that $f(g(a, t), t) = a$ for all $(a, t) \in N(a^*, t^*)$. △

Theorem 5.23 ([Kum91, Theorem 1])**.** *The function f is regular at (x^*, t^*, a^*) if and only if*

$$0 \notin \mathcal{D}_f((x^*, t^*); (v, 0)) \quad \text{for each } v \in \mathbb{R}^n \setminus \{0\}. \tag{5.87}$$

If (5.87) *holds, then*

$$v \in \mathcal{D}_g((a^*, t^*); (\alpha, \tau)) \iff \alpha \in \mathcal{D}_f((x^*, t^*); (v, \tau)). \tag{5.88}$$

△

Lemma 5.24. *Condition* (5.87) *is equivalent to the existence of some* $\varepsilon > 0$ *such that*

$$\left\|f(x^1,t^1) - f(x^2,t^2)\right\| \geq \varepsilon(\left\|x^1 - x^2\right\| + \left\|t^1 - t^2\right\|) \tag{5.89a}$$

whenever

$$x^1, x^2 \in B_\varepsilon(x^*),\ t^1, t^2 \in B_\varepsilon(t^*) \text{ and } \left\|t^1 - t^2\right\| \leq \varepsilon \left\|x^1 - x^2\right\|. \tag{5.89b}$$

△

Proof. Let (5.89) be true, $v \in \mathbb{R}^n \setminus \{0\}$, $x_k \to x^*$, $t_k \to t^*$ and $\lambda_k \downarrow 0$ such that

$$u := \lim_{k\to\infty} \frac{f(x_k + \lambda_k v, t_k) - f(x_k, t_k)}{\lambda_k} \in \mathcal{D}_f((x^*,t^*);(v,0)) \tag{5.90}$$

exists. Then $\|u\| \geq \varepsilon \|v\| > 0$ and (5.87) follows. Now assume (5.89) is false for all $\varepsilon > 0$. Then there exist sequences $\varepsilon_k, x_k^1, x_k^2, t_k^1, t_k^2$ with $\varepsilon_k \downarrow 0$ for $k \to \infty$ satisfying (5.89b) for each $k \in \mathbb{N}$ such that

$$\left\|f(x_k^1,t_k^1) - f(x_k^2,t_k^2)\right\| < \varepsilon_k(\left\|x_k^1 - x_k^2\right\| + \left\|t_k^1 - t_k^2\right\|) \tag{5.91}$$

for each $k \in \mathbb{N}$. In particular, $\lambda_k := \left\|x_k^2 - x_k^1\right\| \neq 0$ for each $k \in \mathbb{N}$. Set $v_k := (x_k^2 - x_k^1)/\lambda_k$. W.l.o.g. $v_k \to v \in \mathbb{S}^{n-1} \subset \mathbb{R}^n$ for $k \to \infty$. Keeping in mind that $x_k^1 + \lambda_k v_k = x_k^2$, we obtain

$$\begin{aligned}
\left\|f(x_k^1 + \lambda_k v, t_k^1) - f(x_k^1, t_k^1)\right\| &\leq \left\|f(x_k^2,t_k^2) - f(x_k^1,t_k^1)\right\| + \left\|f(x_k^1 + \lambda_k v, t_k^1) - f(x_k^1 + \lambda_k v_k, t_k^2)\right\| \\
&\leq \left\|f(x_k^2,t_k^2) - f(x_k^1,t_k^1)\right\| + Lip(f)\left(\left\|t_k^1 - t_k^2\right\| + \lambda_k \left\|v_k - v\right\|\right) \\
&< \varepsilon_k((1+\varepsilon_k)\lambda_k) + Lip(f)(\varepsilon_k\lambda_k + \lambda_k \left\|v_k - v\right\|) \\
&\leq \lambda_k\left(2\varepsilon_k + Lip(f)(\varepsilon_k + \left\|v_k - v\right\|)\right)
\end{aligned} \tag{5.92}$$

for each $k \in \mathbb{N}$. Hence $0 \in \mathcal{D}_f((x^*,t^*);(v,0))$ and (5.87) is false. □

Proof of Theorem 5.23. For our purpose it suffices to show that (5.87) (or equivalently (5.89)) implies regularity. We may assume f is Lipschitz continuous on all of U with constant L.
Suppose for some neighbourhoods $N(x^*)$ and $N(a^*,t^*)$ there exists a function $g : N(a^*,t^*) \to N(x^*)$ such that $f(x,t) = a$ for $x = g(a,t)$. Then this function is unique and Lipschitz continuous after possibly reducing the neighbourhoods: Let $(a^1,t^1),(a^2,t^2) \in N(a^*,t^*)$ and $x^k = g(a^k,t^k)$. Then the relation

$$\varepsilon \left\|x^1 - x^2\right\| \leq \left\|t^1 - t^2\right\| + \left\|a^1 - a^2\right\| \tag{5.93}$$

is trivial for $\left\|t^1 - t^2\right\| > \varepsilon \left\|x^1 - x^2\right\|$ and follows from (5.89) for $\left\|t^1 - t^2\right\| \leq \varepsilon \left\|x^1 - x^2\right\|$.

In order to obtain existence of g, set $t^1 = t^2 = t^*$ in (5.89). It follows that $f(\cdot,t^*)$ is a Lipschitz homeomorphism between $B_\varepsilon(x^*)$ and its image $S := f(B_\varepsilon(x^*),t^*) \subset \mathbb{R}^n$. S is open by the invariance of domain theorem, so there exists a $\delta > 0$ such that $\phi := f(\cdot,t^*)^{-1} : B_\delta(a^*) \to B_\varepsilon(x^*)$ is well-defined and Lipschitz continuous. Now let (a,t) be such that $L\left\|t - t^*\right\| + \left\|a - a^*\right\| < \delta$. This choice implies that

$$h_{a,t}(x) := \phi(f(x,t^*) - f(x,t) + a) \tag{5.94}$$

is a well-defined continuous map from $\{x \in \mathbb{R}^n : (x,t) \in U\}$ into $B_\varepsilon(x^*)$. Indeed,

$$\left\|f(x,t^*) - f(x,t) + a - a^*\right\| \leq L\left\|t - t^*\right\| + \left\|a - a^*\right\| < \delta \tag{5.95}$$

by assumption, hence $h_{a,t}$ makes sense and maps into $B_\varepsilon(x^*)$. We can regard it as a self-map on $\overline{B_\varepsilon(x^*)}$. There exists a fixed point $x_0 \in \overline{B_\varepsilon(x^*)}$ by Brouwer's fixed point theorem and

$$x_0 = h_{a,t}(x_0) \iff f(x_0, t^*) = f(x_0, t^*) - f(x_0, t) + a \iff f(x_0, t) = a. \tag{5.96}$$

Set $g(a,t) := x_0$. □

Proof of the implicit function theorem, Theorem 5.14. If all matrices $[G\ H] \in \partial f(x_0, y_0)$ have a non-singular left block $G \in \mathbb{R}^{n\times n}$, then for any $v \in \mathbb{R}^n \setminus \{0\}$ it holds

$$[G\ H](v, 0) = Gv \neq 0, \tag{5.97}$$

or equivalently

$$0 \notin \partial f(x_0, y_0)(v, 0) = \operatorname{conv} \mathcal{D}_f\left((x_0, y_0); (v, 0)\right) \quad \text{for any } v \in \mathbb{R}^n \setminus \{0\}, \tag{5.98}$$

where we have used the characterisation from Proposition 5.21. In particular this implies

$$0 \notin \mathcal{D}_f((x_0, y_0); (v, 0)) \quad \text{for any } v \in \mathbb{R}^n \setminus \{0\}, \tag{5.99}$$

which is just condition (5.87) ensuring regularity of f at $(x_0, y_0, 0)$. This in turn implies the existence of neighbourhoods and an implicit function φ with the stated properties. □

5.4.3.3 Piecewise C^k Functions and Local Lipschitz Continuity of F

Definition 5.25 (Piecewise C^k functions)**.** Let $V \subset \mathbb{R}^n$ be open, $f \in C(V, \mathbb{R}^m)$ and $k \in \mathbb{N} \cup \infty$. Then f is called a *PC^k function* and we write $f \in PC^k(V, \mathbb{R}^m)$ if for every $x_0 \in V$ there exists a neighbourhood $W \subset V$ of x_0 and a finite collection of functions $f^i \in C^k(W, \mathbb{R}^m)$, $i = 1, \dots, r$ such that f is a continuous selection of $f^1, \dots, f^r$ on W, meaning that

$$f(x) \in \{f^1(x), \dots, f^r(x)\} \quad \text{for all } x \in W. \tag{5.100}$$

We call the sets

$$I(x) := \{i : f(x) = f^i(x)\} \quad \text{and} \quad I^e(x) := \left\{i \in I(x) : x \in \overline{\{y \in W : f(y) = f^i(y)\}^\circ}\right\} \tag{5.101}$$

the *active* and the *essentially active index set* at x, respectively. △

We take [Ulb02, Proposition 2.20] and add a calculus rule for quotients and continuous selections:

Proposition 5.26. *Let $k \in \mathbb{N} \cup \{\infty\}$ and $V \subset \mathbb{R}^n$ open.*

(i) The class of PC^k functions is closed under composition, finite summation and multiplication whenever these operations make sense.

(ii) If $f, g \in PC^k(V, \mathbb{R})$ and g is a continuous selection of functions $g^1, \dots, g^r \in C^k(V, \mathbb{R})$ such that $0 \notin g^k(V)$ for all $k \in \{1, \dots, r\}$, then $f/g \in PC^k(V, \mathbb{R})$.

(iii) If $f : V \to \mathbb{R}^m$ is a continuous selection of finitely many functions $f_1, \dots, f_r \in PC^k(V, \mathbb{R}^m)$, then also $f \in PC^k(V, \mathbb{R}^m)$. △

Proof. We prove only the last statement. Each function $f_k \in PC^k(V, \mathbb{R}^m)$ is itself a continuous selection of functions $f_k^1, \dots, f_k^r \in C^k(V, \mathbb{R}^m)$ on a neighbourhood $W_k \subset V$ of x_0, for each $x_0 \in V$. It follows that

$$f(x) \in \{f_1^1(x), \dots, f_1^r(x), \dots, f_r^1(x), \dots, f_r^r(x)\} \quad \text{for all } x \in \bigcap_{k=1}^{r} W_k, \tag{5.102}$$

which implies $f \in PC^k(V, \mathbb{R}^m)$. □

Lemma 5.27. *The function $F(h,\tau) = (M_L - \theta\tau L)h - \overline{f}(h,\tau) - \tau L u^n$ from (5.72) is piecewise smooth in the sense of Definition 5.25, i.e. $F \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R}^N)$.* △

Proof. The first and last summand of F are obviously in $C^\infty(\mathbb{R}^{N+1}, \mathbb{R}^N)$ and piecewise smoothness of $\overline{f}$ can be shown componentwise. Then, because sums of PC^∞ functions are again PC^∞ according to Proposition 5.26, it is sufficient to show that $\overline{f}_{ij} = \alpha_{ij} f_{ij} \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$ for all $i = 1, \dots, N$ and all $j \in K(i)$. We define $\alpha_{ij}^+, \alpha_{ij}^- : \mathbb{R}^{N+1} \to \mathbb{R}$ on the whole space $\mathbb{R}^{N+1}$ by

$$\alpha_{ij}^+ := \min(R_i^+, R_j^-) \qquad \text{and} \qquad \alpha_{ij}^- := \min(R_i^-, R_j^+). \tag{5.103}$$

With this definition, we see that

$$\alpha_{ij} f_{ij} = \begin{cases} \alpha_{ij}^+ f_{ij} & \text{for } f_{ij} \geq 0 \\ \alpha_{ij}^- f_{ij} & \text{for } f_{ij} < 0. \end{cases} \tag{5.104}$$

Set $H^\pm := H_{ij}^\pm := \{(h,\tau) \in \mathbb{R}^{N+1} : f_{ij}(h,\tau) \gtrless 0\}$. If $Q_i^+ = Q_i^- = Q_j^+ = Q_j^- = 0$, then $\alpha_{ij} f_{ij} \equiv 0$ on $\mathbb{R}^{N+1}$ and there is nothing to show. In case $0 \in \{Q_i^+, Q_j^-\}$ and $0 \notin \{Q_i^-, Q_j^+\}$, $\alpha_{ij} f_{ij} \equiv 0$ on $\overline{H^+}$ and it suffices to show that $\alpha_{ij}^- f_{ij} \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$. Then the selection

$$\alpha_{ij} f_{ij} = \begin{cases} 0 & \text{for } f_{ij} \geq 0 \\ \alpha_{ij}^- f_{ij} & \text{for } f_{ij} < 0 \end{cases} \tag{5.105}$$

is continuous because of $|a_{ij}^- f_{ij}| \leq |f_{ij}|$ and the assertion follows from Proposition 5.26 (iii). The case $0 \notin \{Q_i^+, Q_j^-\}$ and $0 \in \{Q_i^-, Q_j^+\}$ is analogous.

Hence let us show that $a_{ij}^+ f_{ij} \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$ for $0 \notin \{Q_i^+, Q_j^-\}$. In this case we can even show that $\alpha_{ij}^+ \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$; then Proposition 5.26 (i) yields $\alpha_{ij}^+ f_{ij} \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$ as a product of a PC^∞ function and a C^∞ function.
Since $\alpha_{ij}^+ = \min(R_i^+, R_j^-)$ and $\min \in PC^\infty(\mathbb{R}^2, \mathbb{R})$, it is sufficient to show $R_i^+, R_j^- \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$, again using Proposition 5.26 (i). W.l.o.g. we restrict ourselves to proving $R_i^+ \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$. We can rewrite

$$R_i^+ = \begin{cases} \min\left(1, \frac{Q_i^+}{P_i^+}\right) & \text{if } P_i^+ > 0 \\ 1 & \text{if } P_i^+ = 0 \end{cases} \tag{5.106}$$

as

$$R_i^+ = \frac{Q_i^+}{\tilde{P}_i^+} \quad \text{with } \tilde{P}_i^+ := \max(P_i^+, Q_i^+) > 0 \tag{5.107}$$

In light of Proposition 5.26 (ii) it suffices to show that $\tilde{P}_i^+$ is a continuous selection of finitely many $PC^\infty(\mathbb{R}^{N+1}, \mathbb{R}_{>0})$ functions. Recall that

$$f_{ij}^+(h,\tau) = [c_{ij}(\tau)(h_i - h_j) + g_{ij}(\tau)]^+ \quad \text{for } j \in K(i). \tag{5.108}$$

Clearly, $f_{ij}^+ \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$ for all $j \in K(i)$ and thus $P_i^+ \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$ as a finite sum of such functions. We therefore have functions $P_i^{+,1}, \dots, P_i^{+,n} \in C^\infty(\mathbb{R}^{N+1}, \mathbb{R})$ such that P_i^+ is a continuous collection of those on $\mathbb{R}^{N+1}$. In order to obtain strictly positive functions to select $\tilde{P}_i^+$ from, we introduce a monotonically non-decreasing function $\eta \in C^\infty(\mathbb{R}, \mathbb{R})$ such that

$$\eta(x) = \begin{cases} x & \text{for } x \geq Q_i^+ \\ \frac{Q_i^+}{2} & \text{for } x \leq \frac{Q_i^+}{2}. \end{cases} \tag{5.109}$$

Then $\eta \circ P_i^{+,k} \in C^\infty(\mathbb{R}^{N+1}, \mathbb{R})$ with $\eta \circ P_i^{+,k}(h,\tau) \geq Q_i^+/2 > 0$ for $k = 1, \dots, n$ and all $(h,\tau) \in \mathbb{R}^{N+1}$. Of course $\eta \circ P_i^+$ is continuous as a concatenation of continuous functions and hence is a continuous selection of the smooth functions $\eta \circ P_i^{+,k}$. Using $\eta \circ P_i^+$ instead of P_i^+ does not change the definition of $\tilde{P}_i^+$:

$$\tilde{P}_i^+ = \max(P_i^+, Q_i^+) = \max(\eta \circ P_i^+, Q_i^+). \tag{5.110}$$

We have thus shown that $\tilde{P}_i^+$ is a continuous selection of the functions $\eta \circ P_i^{+,1}, \dots, \eta \circ P_i^{+,n}, Q_i^+ \in C^\infty(\mathbb{R}^{N+1}, [Q_i^+/2, \infty))$, from which we infer that $\alpha_{ij}^+ \in PC^\infty(\mathbb{R}^{N+1,N})$ for $0 \notin \{Q_i^+, Q_j^-\}$ (and $\alpha_{ij}^- \in PC^\infty(\mathbb{R}^{N+1,N})$ for $0 \notin \{Q_i^-, Q_j^+\}$). Now it remains to show that the selection

$$\alpha_{ij} f_{ij} = \begin{cases} \alpha_{ij}^+ f_{ij} & \text{for } f_{ij} \geq 0 \\ \alpha_{ij}^- f_{ij} & \text{for } f_{ij} < 0 \end{cases} \tag{5.111}$$

is continuous for $0 \notin \{Q_i^+, Q_i^-, Q_j^+, Q_j^-\}$. Let $(h,\tau) \in \mathbb{R}^{N+1}$ such that $f_{ij}(h,\tau) = 0$ and consider a sequence $(h_n, \tau_n) \to (h,\tau)$ for $n \to \infty$. Then $f_{ij}(h_n,\tau_n) \to f_{ij}(h,\tau) = 0$ and $|\alpha_{ij}^\pm f_{ij}(h_n,\tau_n)| \leq |f_{ij}(h_n,\tau_n)| \to 0 = \alpha_{ij}^+ f_{ij}(h,\tau) = \alpha_{ij} f_{ij}(h,\tau)$. Proposition 5.26 (iii) gives $\alpha_{ij} f_{ij} \in PC^\infty(\mathbb{R}^{N+1}, \mathbb{R})$. □

The next proposition allows to show that a PC^1 function is locally Lipschitz (on compact convex neighbourhoods of points). This is the statement of the subsequent Corollary 5.29.

Proposition 5.28 ([Sch12, Proposition 4.1.2])**.** *Let $V \subset \mathbb{R}^n$ be convex, $f^1, \dots, f^l : V \to \mathbb{R}^m$ Lipschitz continuous on V with constants $L^1, \dots, L^l$. Then if $f \in C(V, \mathbb{R}^m)$ is a continuous selection of $f^1, \dots, f^l$, it is also Lipschitz continuous with constant $L = \max(L^1, \dots, L^l)$.* △

Corollary 5.29 ([Sch12, Corollary 4.1.1])**.** *Let $V \subset \mathbb{R}^n$ be open and $f \in PC^1(V, \mathbb{R}^m)$. Then f is locally Lipschitz continuous.* △

Now we can apply the PC^k theory to our function F from (5.72) to show local Lipschitz continuity.

Corollary 5.30. *The function $F(h,\tau) = (M_L - \theta\tau L)h - \overline{f}(h,\tau) - \tau L u^n$ from (5.72) is locally Lipschitz continuous.* △

Corollary 5.31. *Let $i \in \{1,\dots,N\}$ and $j \in K(i)$ and define the open sets $H^{\pm} \subset \mathbb{R}^{N+1}$ by*

$$H^{\pm} := H^{\pm}_{ij} := \{(h,\tau) \in \mathbb{R}^{N+1} : f_{ij}(h,\tau) \gtrless 0\}. \tag{5.112}$$

Then $\alpha_{ij}|_{H^+} \in PC^\infty(H^+,\mathbb{R})$ and $\alpha_{ij}|_{H^-} \in PC^\infty(H^-,\mathbb{R})$. In particular, these two functions are locally Lipschitz continuous. △

Proof. For reasons of analogy, showing $\alpha_{ij}|_{H^+} \in PC^\infty(H^+,\mathbb{R})$ suffices. This is trivial if $0 \in \{Q_i^+, Q_j^-\}$, for then $\alpha_{ij}|_{H^+} \equiv 0$. On the other hand, in the proof of Lemma 5.27 it was shown that $\alpha_{ij}^+ \in PC^\infty(\mathbb{R}^{N+1},\mathbb{R})$ for $0 \notin \{Q_i^+, Q_j^-\}$. Since $H^+ \subset \mathbb{R}^{N+1}$ is open and $\alpha_{ij}|_{H^+} = \alpha_{ij}^+|_{H^+}$, piecewise smoothness follows immediately and Lipschitz continuity follows from Corollary 5.29. □

5.4.3.4 Non-Singularity of the Generalised Jacobian $\Pi_h \partial F(0,0)$

Lemma 5.32. *There exists an open neighbourhood V of $(0,0)$ such that for all $i \in \{1,\dots,N\}$, $j \in K(i)$ and for all $(h,\tau) \in V$ with $P(h,\tau) \neq 0$ either*

$$\min\left(\frac{Q}{P}, 1\right) = 0 \qquad \text{or} \qquad \min\left(\frac{Q}{P}, 1\right) = 1, \tag{5.113}$$

holds on a neighbourhood $U \subset V$ of (h,τ), where $Q \in \{Q_i^+, Q_j^-, Q_i^-, Q_j^+\}$ and $P \in \{P_i^+, P_j^-, P_i^-, P_j^+\}$ with matching super- and subscript. △

Proof. P is continuous, thus $P(h,\tau) \neq 0$ implies $P \neq 0$ on a neighbourhood of (h,τ). The first identity is obviously true for $Q = 0$. Define now $S := \{Q_k^+, |Q_k^-| : k = 1,\dots,N\} \setminus \{0\}$ and set $m := \min S > 0$. Since $f_{ij}^{\pm}$ is continuous for each $i \in \{1,\dots,N\}, j \in K(i)$ and $f_{ij}(0,0)^{\pm} = 0$, we have for an open neighbourhood $V \ni (0,0)$ that

$$\left|\frac{m}{P(h,\tau)}\right| > 1 \tag{5.114}$$

whenever $P(h,\tau) \neq 0$, $P \in \{P_i^+, P_j^-, P_i^-, P_j^+\}$ and thus the second identity in (5.113) holds for $Q \neq 0$. □

Lemma 5.33. *Let V be the neighbourhood of $(0,0)$ from Lemma 5.32, $i \in \{1,\dots,N\}$ and $j \in K(i)$.*

(i) If $0 \in \{Q_i^+, Q_j^-\}$ and $0 \in \{Q_i^-, Q_j^+\}$, then $\alpha_{ij} f_{ij} \equiv 0$ on $\mathbb{R}^{N+1}$.

(ii) If $0 \notin \{Q_i^+, Q_j^+, Q_i^-, Q_j^+\}$, then $\alpha_{ij} \equiv 1$ on V. △

Proof. For assertion (i), we notice that

$$\alpha_{ij}(h,\tau) = \min(R_i^+, R_j^-) = 0 \qquad \text{on } H_{ij}^+ = \{(h,\tau) \in \mathbb{R}^{N+1} : f_{ij}(h,\tau) > 0\} \tag{5.115}$$

$$\alpha_{ij}(h,\tau) = \min(R_i^-, R_j^+) = 0 \qquad \text{on } H_{ij}^- = \{(h,\tau) \in \mathbb{R}^{N+1} : f_{ij}(h,\tau) < 0\} \tag{5.116}$$

$$f_{ij}(h,\tau) = 0 \qquad \text{on } H_{ij}^0 = \{(h,\tau) \in \mathbb{R}^{N+1} : f_{ij}(h,\tau) = 0\}, \tag{5.117}$$

since $P_i^+(h,\tau) > 0, P_j^-(h,\tau) < 0$ on H_{ij}^+ and $P_i^-(h,\tau) < 0, P_j^+(h,\tau) > 0$ on H_{ij}^- and thus $R = Q/P$ in all four index cases. It follows that the product $\alpha_{ij} f_{ij}$ vanishes everywhere. For assertion (ii) we can apply the previous lemma to see that

$$\alpha_{ij}(h,\tau) = \min(R_i^+, R_j^-) = 1 \qquad \text{on } H_{ij}^+ \cap V \tag{5.118}$$

$$\alpha_{ij}(h,\tau) = \min(R_i^-, R_j^+) = 1 \qquad \text{on } H_{ij}^- \cap V. \tag{5.119}$$

For $(h,\tau) \in V$ with $f_{ij}(h,\tau) = 0$ we have, either due to $(h,\tau) \in V$ and $P(h,\tau) \neq 0$ or due to $R = 1$ for $P = 0$, that $\alpha_{ij}(h,\tau) = 1$. □

Lemma 5.34. *Let V be the neighbourhood from Lemma 5.32. If necessary, shrink V such that $c_{kl}(\tau) \geq m_{kl}/2 > 0$ for all $k \in \{1,\dots,N\}$, $l \in K(k)$ and τ with $(h,\tau) \in V$. Fix $i \in \{1,\dots,N\}$. If $\overline{f}_i$ is differentiable at $(h,\tau) \in V$, then for each $j \in K(i)$ one of the following conditions must hold:*

(i) $f_{ij}(h,\tau) \neq 0$,

(ii) $0 \in \{Q_i^+, Q_j^-\}$ *and* $0 \in \{Q_i^-, Q_j^+\}$,

(iii) $0 \notin \{Q_i^+, Q_j^-, Q_i^-, Q_j^+\}$. △

Proof. Let $(h,\tau) \in V$. We show that if there exists $j_0 \in K(i)$ such that neither of the three conditions hold, then $\overline{f}_i$ is not directionally differentiable at (h,τ). Hence, suppose that for some $j_0 \in K(i)$ it holds $f_{ij_0}(h,\tau) = 0$ and one of the following is the case:

- Case 1: $0 \in \{Q_i^-, Q_{j_0}^+\}$ and $0 \notin \{Q_i^+, Q_{j_0}^-\}$ or
- Case 2: $0 \in \{Q_i^+, Q_{j_0}^-\}$ and $0 \notin \{Q_i^-, Q_{j_0}^+\}$.

W.l.o.g. we may assume the first case. Set $v \in \mathbb{R}^N \setminus \{0\}$ such that $v_i - v_{j_0} > 0$ and $v_i - v_k = 0$ for all $k \neq j_0$. Then for any $\delta \in \mathbb{R}$ and for all $k \in K(i)$:

$$f_{ik}(h + \delta v, \tau) = c_{ik}(\tau)(h_i - h_k + \delta(v_i - v_k)) + g_{ik}(\tau) = \begin{cases} f_{ik}(h,\tau) & \text{for } k \neq j_0 \\ \delta c_{ij_0}(\tau)(v_i - v_{j_0}) & \text{for } k = j_0. \end{cases} \tag{5.120}$$

For $\delta \in (0, \delta_0]$ and small enough δ_0, $(h \pm \delta v, \tau) \in V$. Furthermore, from (5.120) we see

$$f_{ij_0}(h + \delta v, \tau) > 0 \qquad \text{and} \qquad f_{ij_0}(h - \delta v, \tau) < 0, \tag{5.121}$$

so from the definition of α_{ij}, our assumption Case 1 and Lemma 5.32 it follows that

$$\alpha_{ij_0}(h + \delta v, \tau) = 1 \qquad \text{and} \qquad \alpha_{ij_0}(h - \delta v, \tau) = 0. \tag{5.122}$$

Now we can show that $\overline{f}_i$ is not directionally differentiable in direction of v at (h,τ) because the right-sided and left-sided limits disagree. To this end, set

$$\mathcal{K} := \{j \in K(i) : f_{ij}(h,\tau) \neq 0\} \qquad \text{and} \qquad \mathcal{L} := \{j \in K(i) : f_{ij}(h,\tau) = 0\} \ni j_0. \tag{5.123}$$

From $f_{ij}(h,\tau) > 0$ it follows that $P_i^+, P_j^- \neq 0$ on a neighbourhood of (h,τ); likewise, for $f_{ij}(h,\tau) < 0$ it follows that $P_i^-, P_j^+ \neq 0$ on a neighbourhood of (h,τ). Employing Lemma 5.32, we see that α_{ij} is locally constant around (h,τ) for $j \in \mathcal{K}$ and therefore the first sum in the decomposition

$$\overline{f}_i(h,\tau) = \sum_{j \in \mathcal{K}} \alpha_{ij} f_{ij} + \sum_{j \in \mathcal{L}} \alpha_{ij} f_{ij} \tag{5.124}$$

is differentiable at (h,τ) for smoothness of the f_{ij}. We can thus focus on the second sum:

$$\lim_{\delta\downarrow 0}\sum_{j\in\mathcal{L}}\frac{\overbrace{\alpha_{ij}(h+\delta v,\tau)}^{=1 \text{ for } j=j_0}\ \overbrace{f_{ij}(h+\delta v,\tau)}^{=f_{ij}(h,\tau)=0 \text{ for } j\neq j_0}-\alpha_{ij}(h,\tau)\overbrace{f_{ij}(h,\tau)}^{=0}}{\delta}=\lim_{\delta\downarrow 0}\frac{\delta c_{ij_0}(\tau)(v_i-v_{j_0})}{\delta}>0, \quad (5.125)$$

but for the limit from the other side we obtain

$$\lim_{\delta\downarrow 0}\sum_{j\in\mathcal{L}}\frac{\overbrace{\alpha_{ij}(h-\delta v,\tau)}^{=0 \text{ for } j=j_0}\ \overbrace{f_{ij}(h-\delta v,\tau)}^{=f_{ij}(h,\tau)=0 \text{ for } j\neq j_0}-\alpha_{ij}(h,\tau)\overbrace{f_{ij}(h,\tau)}^{=0}}{-\delta}=0. \quad (5.126)$$

This finishes the proof, as we have shown that the negation of conditions (i) – (iii) for some $j_0\in K(i)$ precludes differentiability of $\overline{f}_i$ at (h,τ). □

Corollary 5.35. *Let V be the neighbourhood from Lemma 5.32 and $i\in\{1,\dots,N\}$. If $\overline{f}_i$ is differentiable at $(h,\tau)\in V$, then for each $j\in K(i)$ such that $\alpha_{ij}f_{ij}$ does not vanish identically on $\mathbb{R}^{N+1}$, it holds that α_{ij} is constant with $\alpha_{ij}\equiv 0$ or $\alpha_{ij}\equiv 1$ on a neighbourhood of (h,τ).* △

Proof. According to Lemma 5.34, for each $j\in K(i)$ one of three conditions stated there must hold, the second of which implies that $\alpha_{ij}f_{ij}$ vanishes identically on $\mathbb{R}^{N+1}$.
The first sufficient condition is $f_{ij}(h,\tau)\neq 0$. If $f_{ij}(h,\tau)>0$ it follows that $P_i^+,P_j^-\neq 0$ on a neighbourhood of (h,τ); likewise, for $f_{ij}(h,\tau)<0$ it follows that $P_i^-,P_j^+\neq 0$ on a neighbourhood of (h,τ). Employing Lemma 5.32, we see that α_{ij} is locally constant around (h,τ).
In Lemma 5.33 it has been shown that the third sufficient condition implies $\alpha_{ij}\equiv 1$ on V. □

We are now in a position to show existence of a unique solution $h(\tau)$ for small $\tau>0$. Let us note that F can be decomposed into a smooth part $\overline{g}$ and $\overline{f}$:

$$F(h,\tau):=\underbrace{(M_L-\theta\tau L)h-\tau Lu^n}_{=:\overline{g}(h,\tau)}-\overline{f}(h,\tau). \quad (5.127)$$

Proposition 5.36. *$\Pi_h\partial F(0,0)$ is symmetric positive definite, in particular it is non-singular.* △

Proof. Firstly, let us note that $\nabla_h\overline{g}(0,0)=M_L$, so that we need only investigate the structure of the set $\Pi_h\partial\overline{f}(0,0)$. Secondly, recall that $\Pi_h\partial\overline{f}(0,0)$ is the convex hull of $\Pi_h\partial_B\overline{f}(0,0)$, the projection onto the left square part of the Bouligand subdifferential $\partial_B\overline{f}(0,0)$, see Definition 5.13. Since convex combinations of symmetric positive definite matrices are again symmetric positive definite, it is sufficient to show the claim for each $H:=\Pi_h\tilde{H}$, where $\tilde{H}\in\partial_B\overline{f}(0,0)$.

To obtain an element $\tilde{H}\in\partial_B\overline{f}(0,0)$, let $((h_n,\tau_n))_{n\in\mathbb{N}}$ be an arbitrary sequence in $D_F=D_{\overline{f}}$, the set of differentiability points of F or $\overline{f}$ – which are the same set – such that $(h_n,\tau_n)\to(0,0)$ and $\nabla\overline{f}(h_n,\tau_n)\to\tilde{H}$ for some $\tilde{H}\in\mathbb{R}^{N\times(N+1)}$. Since we are only interested in the limit, we may assume that $(h_n,\tau_n)\in V$ for all $n\in\mathbb{N}$, where V is the neighbourhood from Lemma 5.32. Corollary 5.35 assures us that we can regard all α_{ij} for $i\in\{1,\dots,N\}, j\in K(i)$ as locally constant around (h_n,τ_n), with the values attained lying in $\{0,1\}$. It is therefore easy to compute the Jacobian $H^n:=\nabla_h\overline{f}(h_n,\tau_n)$ of the function $\overline{f}$ at (h_n,τ_n) as it is affine on a neighbourhood of this point:

$$h_{ik}^n = \begin{cases} \sum_{j\in K(i)} \alpha_{ij}^n c_{ij}(\tau_n) & \text{for } k = i \\ -\alpha_{ik}^n c_{ik}(\tau_n) & \text{for } k \in K(i) \\ 0 & \text{else,} \end{cases} \tag{5.128}$$

for $\alpha_{ij}^n \in \{0,1\}$ and $\alpha_{ij}^n = \alpha_{ji}^n$ for all $i \in \{1,\dots,N\}, j \in K(i)$. Considering that $\alpha_{ij}^n \in \{0,1\}$ and $c_{ij}(\tau_n) \to c_{ij}(0) = m_{ij} > 0$ for each $i \in \{1,\dots,N\}, j \in K(i)$ as $n \to \infty$, convergence of the sequence $(H^n)_{n\in\mathbb{N}}$ requires that each sequence $(\alpha_{ij}^n)_{n\in\mathbb{N}}$ eventually become constant: $\alpha_{ij}^n = \alpha_{ij} \in \{0,1\}$ for $n \in \mathbb{N}$ large enough. It follows that $H^n \to H$ with

$$h_{ik} = \begin{cases} \sum_{j\in K(i)} \alpha_{ij} m_{ij} & \text{for } k = i \\ -\alpha_{ik} m_{ik} & \text{for } k \in K(i) \\ 0 & \text{else.} \end{cases} \tag{5.129}$$

Hence, each $G \in \Pi_h \partial_B F(0,0)$ is of the form $M_L - H$ for symmetric matrices H of the form (5.129). Define $P := M_L - M_C - H$, where M_C is the consistent (non-lumped) mass matrix, so that $M_L - H = M_C + P$. More explicitly,

$$p_{ik} = \begin{cases} \sum_{j\in K(i)} (1-\alpha_{ij}) m_{ij} & \text{for } k = i \\ (\alpha_{ik} - 1) m_{ik} & \text{for } k \in K(i) \\ 0 & \text{else,} \end{cases} \tag{5.130}$$

which constitutes a symmetric matrix with zero row sums, non-negative diagonal elements and non-positive off-diagonal elements. Then by Gershgorin's theorem, P is positive semidefinite, rendering $G = M_L - H = M_C + P$ positive definite. □

Theorem 5.37. *There exists some $\delta > 0$ such that for all $\tau \in (0,\delta)$ there exists a unique $h(\tau)$ with $F(h(\tau),\tau) = 0$.* △

Proof. According to Corollary 5.30, our function F is locally Lipschitz continuous on $\mathbb{R}^N \times \mathbb{R}$ and in particular Lipschitz continuous on a neighbourhood of the point $(0,0)$ at which $F(0,0) = 0$ holds (see (5.73)). Proposition 5.36 ensures that $\Pi_h \partial F(0,0)$ is non-singular, because all matrices in this set are symmetric positive definite. Now from the implicit function theorem (Theorem 5.14) it follows that there exist neighbourhoods of U of $h = 0$ and V of $\tau = 0$ such that for all $\tau \in V$ there exists a unique $h(\tau)$ such that $F(h(\tau),\tau) = 0$ for $\tau \in V$. The function $V \ni \tau \mapsto h(\tau)$ is even Lipschitz continuous. □

5.4.3.5 Non-Singularity Away from the Point $(h,\tau) = (0,0)$

So far, the analysis has been relying heavily on the fact that the α_{ij} become locally constant at points of differentiability for small enough (h,τ), which essentially turns the nonlinear function $\overline{f}$ into a linear one. The price for this is that the above existence and uniqueness result possibly holds only for a very small time-step bound $\delta > 0$. In fact, the argument that $\alpha_{ij} \equiv 1$ locally around $(h,\tau) = (0,0)$ in the case $0 \notin \{Q_i^+, Q_j^-, Q_i^-, Q_j^+\}$ only works assuming that $|P_i^+| \le |Q_i^+|, |P_j^-| \le |Q_j^-|, |P_i^-| \le |Q_i^-|$ and $|P_j^+| \le |Q_j^+|$, which can be guaranteed only by assuming $\|(h,\tau)\|$ to be small enough. Obviously, the moduli of the quantities Q can be very small. Such a tiny neighbourhood of $(h,\tau) = (0,0)$ is then likely to be departed from by solutions to $F(h(\tau),\tau)$ at reasonable time-step choices for τ.

Therefore, a result guaranteeing regularity of $\nabla_h(F(h,\tau))$ at arbitrary points $(h,\tau) \in \mathbb{R}^N \times \mathbb{R}_{>0}$ would be desirable.

Since the implicit function theorem for Lipschitz functions (Theorem 5.14) requires the regularity of convex combinations of limits of Jacobians, it seems advisable and practical to seek not only regularity of such Jacobians, but some notion of regularity that is preserved under convex combinations. Such classes are, for example, the class of strictly or irreducibly diagonally dominant matrices (by rows or columns) with positive diagonal (see Lemma 5.38 below) or the class of positive definite matrices. To obtain such Jacobians, unfortunately, we will have to set $\theta = 0$ (explicit Euler stepping) because the additional terms $\theta\tau y_{ij}$ in the definition of c_{ij} in (5.70), which are positive for $i \neq j$ and negative for $i = j$, may destroy the favourable symmetric positive definite or diagonal dominance properties of the regarded matrices.

5.4.3.5.1 Some Facts about Matrix Classes of Interest. This paragraph is a collection of simple lemmata used in the next paragraph.

Lemma 5.38 (Positive combinations of diagonally dominant matrices)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be diagonally dominant (by rows or columns) with non-negative diagonal entries and $C := \lambda A + \mu B$ for some $\lambda, \mu > 0$. Then the following assertions hold:*

(i) C is diagonally dominant.

(ii) If A or B is strictly diagonally dominant, so is C.

(iii) If A and B have positive diagonals, A has non-negative and B has non-positive off-diagonal, and in each row (column) there is some cancellation, i.e. for $i = 1, \dots, n$ there exists $j \neq i$ such that $a_{ij} > 0$ and $b_{ij} < 0$, then C is strictly diagonally dominant.

(iv) If $A, B \geq 0$ and A or B is irreducibly diagonally dominant, so is C. △

Proof. We may restrict to diagonal dominance by rows; then we easily compute for $i = 1, \dots, n$:

$$\lambda a_{ii} + \mu b_{ii} - \sum_{\substack{j=1\\ j\neq i}}^{n} |\lambda a_{ij} + \mu b_{ij}| \geq \lambda \left(a_{ii} - \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \right) + \mu \left(b_{ii} - \sum_{\substack{j=1\\ j\neq i}}^{n} |b_{ij}| \right) \geq 0, \tag{5.131}$$

and this inequality is strict for rows i such that the corresponding inequality for A or B is strict for row i. Let now A have non-negative and B have non-positive off-diagonal and let there be cancellation in each row. Note that $|\lambda a_{ij} + \mu b_{ij}| \leq |\lambda a_{ij}| + |\mu b_{ij}|$ for $j \neq i$ and equality holds if and only if $0 \in \{a_{ij}, b_{ij}\}$. Setting $\mathcal{K}_i := \{j \in \{1, \dots, n\} : a_{ij} > 0 \text{ and } b_{ij} < 0\}$, $\mathcal{L}_i := \{1, \dots, n\} \setminus \mathcal{K}_i \cup \{i\}$ and

$$\Delta_i := \sum_{j\in\mathcal{K}_i} (|\lambda a_{ij}| + |\mu b_{ij}| - |\lambda a_{ij} + \mu b_{ij}|) > 0, \tag{5.132}$$

we see that

$$\lambda a_{ii} + \mu b_{ii} - \sum_{j\neq i} |\lambda a_{ij} + \mu b_{ij}| = \underbrace{\lambda a_{ii} + \mu b_{ii} - \sum_{j\neq i} (|\lambda a_{ij}| + |\mu b_{ij}|)}_{\geq 0} + \Delta_i > 0 \tag{5.133}$$

in each row i because of cancellation. For the last assertion, let w.l.o.g. A be irreducibly diagonally dominant. Then $c_{ij} > 0$ for each pair (i, j) with $a_{ij} > 0$, hence C is irreducible. □

Lemma 5.39. *Let $A, B \in \mathbb{R}^{n\times n}$ be such that $A \geq 0$ is diagonally dominant by columns. For $i = 1, \ldots, n$ set*

$$\Delta_i(A) := a_{ii} - \sum_{j\neq i} a_{ji} \geq 0 \tag{5.134}$$

$$\Sigma_i(B) := \sum_{j=1}^{n} b_{ji}. \tag{5.135}$$

If $A + B \geq 0$ and

$$\Sigma_i(B) \leq \Delta_i(A) + 2b_{ii} \quad \text{for } i = 1, \ldots, n, \tag{5.136}$$

then $A + B$ is diagonally dominant by columns. In particular, this holds if B has vanishing column sums and non-negative diagonal elements. △

Proof. For $i = 1, \ldots, n$ we see that

$$a_{ii} + b_{ii} - \sum_{j\neq i}(a_{ji} + b_{ji}) = \Delta_i(A) + 2b_{ii} - \Sigma_i(B) \geq 0. \tag{5.137}$$

□

Proposition 5.40. *Let $A \in \mathbb{R}^{n\times n}$ be positive definite, $\lambda_1 > 0$ the minimal eigenvalue of its symmetric part and $E \in \mathbb{R}^{n\times n}$ a perturbation. Then $A + E$ is positive definite if $\|E\|_2 < \lambda_1$.* △

Proof. For a matrix $M \in \mathbb{R}^{n\times n}$, denote by $M_{\text{sym}} := (M + M^T)/2$ its symmetric part and by $M_{\text{skew}} := (M - M^T)/2$ its skew-symmetric part. We need to show that $(A + E)_{\text{sym}} = A_{\text{sym}} + E_{\text{sym}}$ is symmetric positive definite. By our premise we have

$$\max\{|\mu| : \mu \in \sigma(E_{\text{sym}})\} = \|E_{\text{sym}}\|_2 = \left\|\frac{E + E^T}{2}\right\|_2 \leq \|E\|_2 < \lambda_1. \tag{5.138}$$

If μ_1 denotes the minimal eigenvalue of E_{sym}, we therefore obtain $\mu_1 > -\lambda_1$. For the minimal eigenvalue of $A_{\text{sym}} + E_{\text{sym}}$ we can now infer using Rayleigh quotients:

$$\begin{aligned}\min\{\lambda : \lambda \in \sigma(A_{\text{sym}} + E_{\text{sym}})\} &= \min_{\|x\|_2=1} x^T(A_{\text{sym}} + E_{\text{sym}})x \\ &\geq \min_{\|x\|_2=1} x^T A_{\text{sym}} x + \min_{\|x\|_2=1} x^T E_{\text{sym}} x = \lambda_1 + \mu_1 > 0.\end{aligned} \tag{5.139}$$

□

Lemma 5.41. *The mass matrix M_C is strictly diagonally dominant for $d = 1$, weakly diagonally dominant for $d = 2$ and not diagonally dominant for $d = 3$.* △

Proof. For $d \in \{1,2,3\}$, let $\hat{T}^d := \mathrm{conv}\{e_1,\dots,e_d\}$ be the standard simplex. It is known that $|\hat{T}^d| = 1/d!$ and elementary calculus shows

$$\int_{\hat{T}^1} x^2 \, d\lambda^1 = \int_0^1 x^2 \, dx = \frac{1}{3} = \frac{1}{3}|\hat{T}^1| \tag{5.140}$$

$$\int_{\hat{T}^2} x^2 \, d\lambda^2 = \int_0^1 \int_0^{1-x} x^2 \, dy \, dx = \frac{1}{12} = \frac{1}{6}|\hat{T}^2| \tag{5.141}$$

$$\int_{\hat{T}^3} x^2 \, d\lambda^3 = \int_0^1 \int_0^{1-x} \int_0^{1-x-y} x^2 \, dz \, dy \, dx = \frac{1}{60} = \frac{1}{10}|\hat{T}^3| \tag{5.142}$$

and

$$\int_{\hat{T}^1} x(1-x) \, d\lambda^1 = \int_0^1 x(1-x) \, dx = \frac{1}{6} = \frac{1}{6}|\hat{T}^1| \tag{5.143}$$

$$\int_{\hat{T}^2} xy \, d\lambda^2 = \int_0^1 \int_0^{1-x} xy \, dy \, dx = \frac{1}{24} = \frac{1}{12}|\hat{T}^2| \tag{5.144}$$

$$\int_{\hat{T}^3} xy \, d\lambda^3 = \int_0^1 \int_0^{1-x} \int_0^{1-x-y} xy \, dz \, dy \, dx = \frac{1}{120} = \frac{1}{20}|\hat{T}^3|. \tag{5.145}$$

By affine transformations, the first and last term of each line are equal also for arbitrary d-simplices $T^d = \mathrm{conv}\{p_1,\dots,p_{d+1}\} \subset \mathbb{R}^d$ if x and $1-x$ ($d=1$) or x and y ($d \in \{2,3\}$) are replaced by φ_k and φ_l, respectively, for $k \neq l, k, l \in \{1,\dots,d+1\}$, where φ_k is the linear standard basis function on T^d associated with node p_k.
Let now p be a node of $\mathcal{T}$ with basis function φ. The assertion follows from the fact that, per element $T \in \mathcal{T}$ containing p, the term $\int_{T^d} \varphi^2 \, dx$ contributes to the diagonal of M_C only once but terms of the form $\int_{T^d} \varphi\psi \, dx$ (ψ being the basis function of another node of T) contribute d times to the off-diagonal entries of the row of M_C associated with p. □

5.4.3.5.2 Investigation of Generalised Jacobians Away from $(h,\tau) = (0,0)$. The goal of this paragraph is to find out whether non-singularity of $\Pi_h \partial F(h,\tau)$ can be guaranteed at arbitrary $(h,\tau) \in \mathbb{R}^N \times \mathbb{R}_{>0}$. The computation and analysis of Jacobians at points in $D_F = D_{\overline{f}}$ is complicated now by multiple issues:

- If convergence $\tau_n \to 0$ does not hold, then $c_{ij}(\tau_n) \not\to m_{ij}$ unless $\theta = 0$.
- The terms α_{ij} can no longer be argued to be constant around differentiability points $(h,\tau) \in D_F \cap \mathbb{R}^N \times \mathbb{R}_{>0}$. Therefore they have to be differentiated, too.
- From differentiability of the components $\overline{f}_i = \sum_{j \in K(i)} \alpha_{ij} f_{ij}$ at some point $(h,\tau) \in D_F$, differentiability of the individual summands does not follow.
- From differentiability of the product $\alpha_{ij} f_{ij}$ at some point, differentiability of the first factor (at least in some direction) is not immediate either.

The objective is thus to find arguments that allow for sum and product rules in order to break down the problem of determining the Jacobians. The following proposition is a crucial ingredient:

Proposition 5.42 (Invariance of generalised Jacobian under removal of null sets, [FP87])**.** *Let $r > 0$, $B_r(x) \subset \mathbb{R}^n$ the open r-ball around $x \in \mathbb{R}^n$, $f : B_r(x) \to \mathbb{R}^m$ locally Lipschitz continuous and $S \subset B_r(x)$ a set of Lebesgue measure zero. Then with the redefined Bouligand subdifferential*

$$\partial_B^S f(x) := \{G \in \mathbb{R}^{m\times n} : \exists\ (x_n)_{n\in\mathbb{N}} \subset D_f \setminus S \text{ with } x_n \to x \text{ and } \nabla F(x_n) \to G\} \tag{5.146}$$

it still holds that $\partial^S f(x) := \operatorname{conv}(\partial_B^S f(x)) = \partial f(x)$. △

A null set we certainly want to remove from $D_{\overline{f}}$ is the set of points such that one of the terms $\alpha_{ij} f_{ij}, i \in \{1,\dots,N\}, j \in K(i)$ is not differentiable. This way, differentiability of the summands defining the components of $\overline{f}$ is guaranteed and a sum rule for differentiation can be applied. Hence, set

$$N_{ij} := \left\{(h,\tau) \in \mathbb{R}^{N+1} : \alpha_{ij} f_{ij} \text{ is not differentiable at } (h,\tau)\right\} \tag{5.147}$$

and

$$N_1 := \bigcup_{\substack{i\in\{1,\dots,N\}\\ j\in K(i)}} N_{ij}. \tag{5.148}$$

The sets N_{ij} (and thus N_1) are null sets according to Rademacher's theorem since the $\alpha_{ij} f_{ij}$ are locally Lipschitz continuous on $\mathbb{R}^{N+1}$. The second null set should be the set

$$N_2 := \bigcup_{\substack{i\in\{1,\dots,N\}\\ j\in K(i)}} H_{ij}^0 \subset \mathbb{R}^N \times \mathbb{R}_{>0}, \tag{5.149}$$

where

$$H_{ij}^0 = \left\{(h,\tau) \in \mathbb{R}^N \times \mathbb{R}_{>0} : f_{ij}(h,\tau) = 0\right\}. \tag{5.150}$$

Lemma 5.43. *For $i \in \{1,\dots,N\}$, $j \in K(i)$ the set $H_{ij}^0 \subset \mathbb{R}^N \times \mathbb{R}_{>0}$ is a set of Lebesgue measure zero.* △

Proof. For arbitrary fixed $\tau \in \mathbb{R}_{>0}$, the set $H_{ij}^{0,\tau} := \{h \in \mathbb{R}^N : f(h,\tau) = 0\}$ is an affine hyperplane of $\mathbb{R}^N \times \{\tau\}$, i.e. an $(N-1)$-dimensional affine subspace:

$$f_{ij}(h,\tau) = 0 \iff \underbrace{c_{ij}(\tau)}_{>0}(h_i - h_j) = -g_{ij}(\tau) \iff h_i - h_j = -g_{ij}(\tau)/c_{ij}(\tau). \tag{5.151}$$

It follows that $H_{ij}^{0,\tau}$ is a set of vanishing N-dimensional Lebesgue measure in $\mathbb{R}^N \times \{\tau\}$ for each $\tau \geq 0$ and from Fubini's theorem we obtain the assertion by integrating the characteristic function of $H_{ij}^{0,\tau}$ as follows:

$$\lambda^{N+1}(H_{ij}^0) = \int_{\mathbb{R}^N\times\mathbb{R}_{>0}} \chi_{H_{ij}^0}\, d\lambda^{N+1} = \int_0^\infty \int_{\mathbb{R}^N\times\{\tau\}} \chi_{H_{ij}^{0,\tau}}\, dh d\tau = 0. \tag{5.152}$$

□

Discarding the set N_2 offers the advantage that the existence of an explicit formula for the derivatives of α_{ij} along coordinate directions $e_1,\dots,e_N$ (for fixed τ_0) and a product rule can be shown:

Lemma 5.44. *Let N_1 and N_2 be the null sets defined in* (5.148) *and* (5.149)*, respectively.*

(i) For any $(h_0, \tau_0) \in \mathbb{R}^N \times \mathbb{R}_{>0} \setminus N_2$, $i, m \in \{1, \dots, N\}$, $j \in K(i)$, the function α_{ij} has well-defined one-sided directional derivatives at (h_0, τ_0) in both coordinate directions e_m and $-e_m$.

(ii) If in addition $(h_0, \tau_0) \in \mathbb{R}^N \times \mathbb{R}_{>0} \setminus (N_1 \cup N_2)$, then these two one-sided derivatives agree and the following product rule holds:

$$\frac{\partial}{\partial h_m}(\alpha_{ij} f_{ij})(h_0, \tau_0) = \left(\frac{\partial}{\partial h_m}\alpha_{ij}(h_0, \tau_0)\right) f_{ij}(h_0, \tau_0) + \alpha_{ij}(h_0, \tau_0)\frac{\partial}{\partial h_m} f_{ij}(h_0, \tau_0), \tag{5.153}$$

where the partial derivatives of α_{ij} can be computed explicitly:

- *Case 1: $f_{ij}(h_0, \tau_0) > 0$.*

(a) If $\alpha_{ij}(h_0, \tau_0) = Q_i^+ / P_i^+(h_0, \tau_0)$, then

$$\frac{\partial \alpha_{ij}}{\partial h_m}(h_0, \tau_0) \in \left\{0, \frac{-Q_i^+}{P_i^+(h_0, \tau_0)^2}\frac{\partial P_i^+}{\partial h_m}(h_0, \tau_0)\right\} \tag{5.154}$$

with

$$\frac{\partial P_i^+}{\partial h_m}(h_0, \tau_0) = \begin{cases} \sum\limits_{\substack{k \in K(i) \\ f_{ik}(h_0, \tau_0) > 0}} c_{ik}(\tau_0) & \text{for } m = i \\ -c_{im}(\tau_0) & \text{for } f_{im}(h_0, \tau_0) > 0 \\ 0 & \text{else.} \end{cases} \tag{5.155}$$

(b) If $\alpha_{ij}(h_0, \tau_0) = Q_j^- / P_j^-(h_0, \tau_0)$, then

$$\frac{\partial \alpha_{ij}}{\partial h_m}(h_0, \tau_0) \in \left\{0, \frac{-Q_j^-}{P_j^-(h_0, \tau_0)^2}\frac{\partial P_j^-}{\partial h_m}(h_0, \tau_0)\right\} \tag{5.156}$$

with

$$\frac{\partial P_j^-}{\partial h_m}(h_0, \tau_0) = \begin{cases} \sum\limits_{\substack{k \in K(j) \\ f_{jk}(h_0, \tau_0) < 0}} c_{jk}(\tau_0) & \text{for } m = j \\ -c_{jm}(\tau_0) & \text{for } f_{jm}(h_0, \tau_0) < 0 \\ 0 & \text{else.} \end{cases} \tag{5.157}$$

(c) If $Q_i^+ / P_i^+(h_0, \tau_0)$, $Q_j^- / P_j^-(h_0, \tau_0) > 1$, then

$$\frac{\partial}{\partial h_m}\alpha_{ij}(h_0, \tau_0) = 0 \qquad \text{for } m = 1, \dots, N. \tag{5.158}$$

- *Case 2: $f_{ij}(h_0, \tau_0) < 0$.*

(a) If $\alpha_{ij}(h_0, \tau_0) = Q_i^- / P_i^-(h_0, \tau_0)$, then

$$\frac{\partial \alpha_{ij}}{\partial h_m}(h_0, \tau_0) \in \left\{0, \frac{-Q_i^-}{P_i^-(h_0, \tau_0)^2}\frac{\partial P_i^-}{\partial h_m}(h_0, \tau_0)\right\} \tag{5.159}$$

with

$$\frac{\partial P_i^-}{\partial h_m}(h_0,\tau_0) = \begin{cases} \sum\limits_{\substack{k\in K(i)\\ f_{ik}(h_0,\tau_0)<0}} c_{ik}(\tau_0) & \text{for } m=i \\ -c_{im}(\tau_0) & \text{for } f_{im}(h_0,\tau_0)<0 \\ 0 & \text{else.} \end{cases} \tag{5.160}$$

(b) If $\alpha_{ij}(h_0,\tau_0) = Q_j^+/P_j^+(h_0,\tau_0)$*, then*

$$\frac{\partial \alpha_{ij}}{\partial h_m}(h_0,\tau_0) \in \left\{0, \frac{-Q_j^+}{P_j^+(h_0,\tau_0)^2}\frac{\partial P_j^+}{\partial h_m}(h_0,\tau_0)\right\} \tag{5.161}$$

with

$$\frac{\partial P_j^+}{\partial h_m}(h_0,\tau_0) = \begin{cases} \sum\limits_{\substack{k\in K(j)\\ f_{jk}(h_0,\tau_0)>0}} c_{jk}(\tau_0) & \text{for } m=j \\ -c_{jm}(\tau_0) & \text{for } f_{jm}(h_0,\tau_0)>0 \\ 0 & \text{else.} \end{cases} \tag{5.162}$$

(c) If $Q_i^-/P_i^-(h_0,\tau_0)$*,* $Q_j^+/P_j^+(h_0,\tau_0) > 1$*, then*

$$\frac{\partial}{\partial h_m}\alpha_{ij}(h_0,\tau_0) = 0 \qquad \text{for } m=1,\dots,N. \tag{5.163}$$

△

Proof. W.l.o.g. we can assume case 1: $f_{ij}(h_0,\tau_0) > 0$, so that $\alpha_{ij} = \min(R_i^+, R_j^-)$ with

$$R_i^+ = \min\left(\frac{Q_i^+}{P_i^+}, 1\right) \qquad \text{and} \qquad R_j^- = \min\left(\frac{Q_j^-}{P_j^-}, 1\right) \tag{5.164}$$

on a neighbourhood of (h_0,τ_0).

(1) If $Q_i^+/P_i^+(h_0,\tau_0)$, $Q_j^-/P_j^-(h_0,\tau_0) > 1$, we have $\alpha_{ij}(h,\tau_0) \equiv 1$ on a neighbourhood of h_0 and (5.158) follows.

(2) If $0 \in \{Q_i^+, Q_j^-\}$, then $\alpha_{ij}(\cdot,\tau_0) \equiv 0$ and (5.154) or (5.156) holds trivially.

(3) If $R_i^+(h_0,\tau_0) < R_j^-(h_0,\tau_0)$, then $\alpha_{ij}(\cdot,\tau_0) = Q_i^+/P_i^+(h,\tau_0)$ with

$$P_i^+(h,\tau_0) = \sum_{\substack{k\in K(i)\\ f_{ik}(h_0,\tau_0)>0}} c_{ik}(\tau_0)(h_i - h_k) + g_{ik}(\tau_0) \tag{5.165}$$

on a neighbourhood of h_0, so that $\alpha_{ij}(\cdot,\tau_0)$ is differentiable on this neighbourhood and

$$\frac{\partial \alpha_{ij}}{\partial h_m^+}(h_0,\tau_0) = \frac{\partial \alpha_{ij}}{\partial h_m^-}(h_0,\tau_0) \tag{5.166}$$

are given by (5.154) and (5.155).

(4) The case $R_j^-(h_0,\tau_0) < R_i^+(h_0,\tau_0)$ is treated analogously.

(5) If $0 < R_i^+(h_0,\tau_0) = \frac{Q_i^+}{P_i^+}(h_0,\tau_0) = \frac{Q_j^-}{P_i^-}(h_0,\tau_0) = R_j^-(h_0,\tau_0)$, define

$$\mathcal{J}^+(h) := \{j \in K(i) : f_{ij}(h,\tau_0) > 0\} \tag{5.167}$$
$$\mathcal{J}^-(h) := \{k \in K(j) : f_{jk}(h,\tau_0) < 0\} \tag{5.168}$$

and let $V \subset \mathbb{R}^N$ be a neighbourhood of h_0 such that $\mathcal{J}^\pm(h) = \mathcal{J}^\pm(h_0)$ for all $h \in V$. Furthermore, define

$$H^{0,+,-} := \left\{h \in \mathbb{R}^N : Q_i^+ P_j^-(h,\tau_0) - Q_j^- P_i^+(h,\tau_0) \overset{=}{\gtrless} 0\right\}. \tag{5.169}$$

Then clearly $h_0 \in H^0 \cap V$ and since the map $Q_i^+ P_j^-(\cdot,\tau) - Q_j^- P_i^+(\cdot,\tau)$ is affine on V we have that $H^0 \cap V$ is the intersection of V with a hyperplane and $H^\pm \cap V$ are the intersections of V and the adjoining half-spaces.

Two subcases need to be considered:

(a) $R_i^+(h_0,\tau_0) = R_j^-(h_0,\tau_0) < 1$. Then (possibly after shrinking V) we have $R_i^+ = Q_i^+/P_i^+$ and $R_j^- = Q_j^-/P_j^-$ on V and

$$\alpha_{ij}(\cdot,\tau_0) = Q_i^+/P_i^+(\cdot,\tau_0) \quad \text{on } (H^0 \cup H^+) \cap V \tag{5.170}$$
$$\alpha_{ij}(\cdot,\tau_0) = Q_j^-/P_j^-(\cdot,\tau_0) \quad \text{on } (H^0 \cup H^-) \cap V, \tag{5.171}$$

so that $\frac{\partial}{\partial h_m^\pm}\alpha_{ij}(h_0,\tau_0)$ can be computed by (5.154) – (5.155) if $h_0 \pm \delta e_m \in (H^0 \cup H^+) \cap V$ and by (5.156) – (5.157) if $h_0 \pm \delta e_m \in (H^0 \cup H^-) \cap V$ for small $\delta > 0$.

(b) $R_i^+(h_0,\tau_0) = R_j^-(h_0,\tau_0) = 1$. Then $h_0 \in H^0 \cap H_i^0 \cap H_j^0 \cap V$, where we define

$$H_i^{0,+,-} := \left\{h \in \mathbb{R}^N : Q_i^+ \overset{=}{\gtrless} P_i^+(h,\tau_0)\right\} \tag{5.172}$$
$$H_j^{0,+,-} := \left\{h \in \mathbb{R}^N : Q_j^- \overset{=}{\gtrless} P_j^-(h,\tau_0)\right\}. \tag{5.173}$$

Now if $h_0 \pm \delta e_m \in (H_i^+ \cup H_i^0) \cap (H_j^- \cup H_j^0) \cap V$ for small $\delta > 0$, we have $\frac{\partial}{\partial h_m^\pm}\alpha_{ij}(\cdot,\tau_0) = 0$, otherwise we can argue as in case (a).

It remains to show that, if $(h_0,\tau_0) \in \mathbb{R}^N \times \mathbb{R}_{>0} \setminus (N_1 \cup N_2)$, then

$$\frac{\partial}{\partial h_m^+}\alpha_{ij}(h_0,\tau_0) = \frac{\partial}{\partial h_m^-}\alpha_{ij}(h_0,\tau_0). \tag{5.174}$$

The one-sided derivatives can only differ in case (5). In case (5)(a), (5.174) holds by construction of H^0 if $h_0 + \delta e_m, h_0 - \delta e_m \in V \cap H^0$ for small $\delta > 0$. Otherwise we have $h_0 \pm \delta e_m \in V \cap H^\pm$ or $h_0 \pm \delta e_m \in V \cap H^\mp$ and hence

$$\frac{\partial}{\partial h_m^+}\alpha_{ij}(h_0,\tau_0) = \frac{-Q_i^+}{P_i^+(h_0,\tau_0)^2}\frac{\partial P_i^+}{\partial h_m}(h_0,\tau_0) \tag{5.175}$$

or

$$\frac{\partial}{\partial h_m^-}\alpha_{ij}(h_0,\tau_0) = \frac{-Q_i^+}{P_i^+(h_0,\tau_0)^2}\frac{\partial P_i^+}{\partial h_m}(h_0,\tau_0) \tag{5.176}$$

and the product rule for $\frac{\partial}{\partial h_m^\pm}\alpha_{ij}f_{ij}(h_0,\tau_0)$ in conjunction with $f_{ij}(h_0,\tau_0)\neq 0$ and differentiability of f_{ij} gives

$$\frac{\partial}{\partial h_m}\alpha_{ij}(h_0,\tau_0) = \frac{\partial}{\partial h_m^+}\alpha_{ij}(h_0,\tau_0) = \frac{\partial}{\partial h_m^-}\alpha_{ij}(h_0,\tau_0) = \frac{-Q_i^+}{P_i^+(h_0,\tau_0)^2}\frac{\partial P_i^+}{\partial h_m}(h_0,\tau_0). \tag{5.177}$$

In case (5)(b) we can argue the same way to find that the left- and right-handed directional derivatives in coordinate directions must agree, but now $\frac{\partial}{\partial h_m}\alpha_{ij}(h_0,\tau_0)=0$ is possible. □

Lemma 5.44 justifies using the following product rule at points $(h,\tau)\in\mathbb{R}^N\times\mathbb{R}_{>0}\setminus(N_1\cup N_2)$:

$$\frac{\partial}{\partial h_m}\overline{f}_i(h,\tau) = \sum_{j\in K(i)}\alpha_{ij}(h,\tau)\frac{\partial}{\partial h_m}f_{ij}(h,\tau) + \sum_{j\in K(i)}\left(\frac{\partial}{\partial h_m}\alpha_{ij}(h,\tau)\right)f_{ij}(h,\tau) \tag{5.178}$$

and thus the decomposition $\nabla_h\overline{f}(h,\tau)=\mathcal{N}+\mathcal{O}$ with

$$\mathcal{N}_{im} := \begin{cases}\sum_{k\in K(i)}\alpha_{ik}(h,\tau)c_{ik}(\tau) & \text{for } m=i\\ -\alpha_{im}(h,\tau)c_{im}(\tau) & \text{for } m\neq i\end{cases},\quad \mathcal{O}_{im} := \sum_{j\in K(i)}\left(\frac{\partial}{\partial h_m}\alpha_{ij}(h,\tau)\right)f_{ij}(h,\tau). \tag{5.179}$$

Remark 5.45. Recall that $c_{ij}(\tau)=m_{ij}+\theta\tau y_{ij}=m_{ij}$ for $\theta=0$ independently of τ. △

Remark 5.46. The matrix $\mathcal{O}$ can be regarded as a perturbation of positive definiteness, since it is still true (for $\theta=0$) that $M_L-\mathcal{N}$ is positive definite. The matrix $\mathcal{O}$ cannot in general be expected to be symmetric, but $M_L-\mathcal{N}-\mathcal{O}$ might still be positive definite, which one could attempt to show by the positive definiteness perturbation result of Proposition 5.40. Bounds for the minimal eigenvalues of the mass matrix M_C can be computed as shown in [Fri72]. △

Assumption 5.47. Let us now focus on the case $\theta=0$ and $d=2$. △

Then we know from Lemma 5.41 that M_C is weakly diagonally dominant. We hope to use this property in order to show some non-singularity implying diagonal dominance property of the generalised Jacobian of $\nabla_h F(h,\tau)$ at $(h,\tau)\in\mathbb{R}^N\times\mathbb{R}_{>0}\setminus(N_1\cup N_2)$ that is preserved as $(h,\tau)\to(h_0,\tau_0)\in\mathbb{R}^N\times\mathbb{R}_{>0}$. Let us decompose $\mathcal{O}=\mathcal{O}^1+\mathcal{O}^2$ at a point $(h,\tau)\in\mathbb{R}^N\times\mathbb{R}_{>0}\setminus(N_1\cup N_2)$ with:

$$\mathcal{O}^1_{im} := \begin{cases}0\\ \sum_{j\in K(i)\setminus\{m\}}\left(\frac{\partial}{\partial h_m}\alpha_{ij}\right)f_{ij}\end{cases},\quad \mathcal{O}^2_{im} := \begin{cases}\sum_{j\in K(m)}\left(\frac{\partial}{\partial h_m}\alpha_{mj}\right)f_{mj} & \text{for } m=i\\ \left(\frac{\partial}{\partial h_m}\alpha_{im}\right)f_{im}\mathbb{1}_{\{m\in K(i)\}} & \text{for } m\neq i,\end{cases} \tag{5.180}$$

where

$$\mathbb{1}_{\{\text{statement}\}} := \begin{cases}1 & \text{if the statement is true}\\ 0 & \text{else.}\end{cases} \tag{5.181}$$

Lemma 5.48. *Let Assumption 5.47 hold. Then $\nabla_h F(h,\tau)$ is weakly diagonally dominant by columns at $(h,\tau)\in\mathbb{R}^N\times\mathbb{R}_{>0}\setminus(N_1\cup N_2)$.* △

Proof. We will make use of Lemma 5.39 with $A := M_C$ and $B := -\mathcal{O}^1$ to first show that $M_C - \mathcal{O}^1$ is weakly diagonally dominant and then prove the same for $M_L - \mathcal{N} - \mathcal{O}^1$ and finally for $\nabla_h F(h,\tau) = M_L - \mathcal{N} - \mathcal{O}$.

(1) $A \geq 0$ and A is weakly diagonally dominant since we assume $d = 2$.

(2) B has a vanishing diagonal by definition. Now we show that B has zero column sums. Note that

$$j \in K(i) \iff i \in K(j) \qquad \text{and} \qquad \alpha_{ij} = \alpha_{ji},\ f_{ij} = -f_{ji}. \tag{5.182}$$

With this it follows immediately from the definition that $\mathcal{O}^2$ has zero column sums, so that it suffices to show that $\mathcal{O}$ has zero column sums. But it follows, again from (5.182), that

$$\sum_{i=1}^{N} \sum_{j \in K(i)} \left(\frac{\partial}{\partial h_m} \alpha_{ij} \right) f_{ij} = 0. \tag{5.183}$$

(3) Let us now prove that $A + B \geq 0$. $a_{ii} + b_{ii} = a_{ii} = m_{ii} > 0$, so it remains to consider $i \neq m$. We omit the argument (h,τ). For $m \in K(i)$ we have

$$a_{im} + b_{im} = m_{im} - \sum_{j \in K(i)\setminus\{m\}} \left(\frac{\partial}{\partial h_m} \alpha_{ij} \right) f_{ij} \geq m_{im} - \sum_{j \in K(i)\setminus\{m\}} \left[\left(\frac{\partial}{\partial h_m} \alpha_{ij} \right) f_{ij} \right]^+. \tag{5.184}$$

According to Lemma 5.44 we may assume that

$$\alpha_{ij} \in \left\{ \frac{Q_i^+}{P_i^+}, \frac{Q_j^-}{P_j^-}, \frac{Q_j^+}{P_j^+}, \frac{Q_i^-}{P_i^-} \right\} \tag{5.185}$$

for each summand of the summation sign in the previous equation; otherwise $\frac{\partial}{\partial h_m} \alpha_{ij} = 0$ for $m = 1, \ldots, N$. We can then distinguish between four cases:

- Case 1: $f_{ij} > 0$, $\alpha_{ij} = \frac{Q_i^+}{P_i^+}$ and $m \neq i$. Then $\left(\frac{\partial}{\partial h_m} \alpha_{ij} \right) f_{ij} = \frac{Q_i^+}{(P_i^+)^2} m_{im} \mathbb{1}_{\{f_{im}>0\}} f_{ij} \geq 0$.
- Case 2: $f_{ij} > 0$, $\alpha_{ij} = \frac{Q_j^-}{P_j^-}$ and $m \neq j$. Then $\left(\frac{\partial}{\partial h_m} \alpha_{ij} \right) f_{ij} = \frac{Q_j^-}{(P_j^-)^2} m_{jm} \mathbb{1}_{\{f_{jm}<0\}} f_{ij} \leq 0$.
- Case 3: $f_{ij} < 0$, $\alpha_{ij} = \frac{Q_i^-}{P_i^-}$ and $m \neq i$. Then $\left(\frac{\partial}{\partial h_m^+} \alpha_{ij} \right) f_{ij} = \frac{Q_i^-}{(P_i^-)^2} m_{im} \mathbb{1}_{\{f_{im}<0\}} f_{ij} \geq 0$.
- Case 4: $f_{ij} < 0$, $\alpha_{ij} = \frac{Q_j^+}{P_j^+}$ and $m \neq j$. Then $\left(\frac{\partial}{\partial h_m} \alpha_{ij} \right) f_{ij} = \frac{Q_j^+}{(P_j^+)^2} m_{jm} \mathbb{1}_{\{f_{jm}>0\}} f_{ij} \leq 0$.

We see that we can ignore terms in (5.184) associated to the cases 2 and 4 (due to their sign). It now follows for $f_{im}(h,\tau) > 0$ and $f_{im}(h,\tau) < 0$ from

$$\frac{Q_i^+}{(P_i^+)^2} m_{im} \sum_{\substack{j \in K(i)\setminus\{m\} \\ \alpha_{ij} = Q_i^+/P_i^+}} f_{ij} \leq \frac{Q_i^+}{P_i^+} m_{im} \leq m_{im}, \tag{5.186}$$

$$\frac{Q_i^-}{(P_i^-)^2} m_{im} \sum_{\substack{j \in K(i)\setminus\{m\} \\ \alpha_{ij} = Q_i^-/P_i^-}} f_{ij} \leq \frac{Q_i^-}{P_i^-} m_{im} \leq m_{im}, \tag{5.187}$$

respectively, that $a_{im} + b_{im} \geq 0$ for $m \in K(i)$. Now we check the case $m \neq i$, $m \notin K(i)$. Then $a_{im} = m_{im} = 0$ and $\frac{\partial}{\partial h_m}\alpha_{ij}$ is non-vanishing only if

$$\alpha_{ij} \in \left\{ \frac{Q_j^+}{P_j^+}, \frac{Q_j^-}{P_j^-} \right\} \quad \text{and} \quad m \in K(j). \tag{5.188}$$

Since $m \notin K(i)$, it holds $m \neq j$ for all $j \in K(i)$ and therefore

$$b_{im} = -\sum_{j \in K(i)} \left(\frac{\partial}{\partial h_m}\alpha_{ij} \right) f_{ij} \geq 0 \tag{5.189}$$

by cases 2 and 4 above. This concludes the proof that $A + B \geq 0$.

(4) Lemma 5.39 yields that $A + B = M_C - \mathcal{O}^1$ is weakly diagonally dominant. To demonstrate that $M_L - \mathcal{N} - \mathcal{O}^1$ is diagonally dominant, we note that

$$M_L - \mathcal{N} - \mathcal{O}^1 = M_C - \mathcal{O}^1 + P \quad \text{with} \quad P = M_L - M_C - \mathcal{N}. \tag{5.190}$$

P is a matrix with zero column sums, non-negative diagonal and non-positive off-diagonal and thus diagonally dominant. Lemma 5.38 gives that $M_C - \mathcal{O}^1 + P$ is (weakly) diagonally dominant by columns as the sum of two diagonally dominant matrices with non-negative diagonals.

(5) In this final step, we show that $M_L - \mathcal{N} - \mathcal{O} = M_L - \mathcal{N} - \mathcal{O}^1 - \mathcal{O}^2$ is diagonally dominant by repeating the same argument used in the previous step, with P replaced by $-\mathcal{O}^2$. It was already shown above that $\mathcal{O}^2$ has zero column sums. To obtain that $\mathcal{O}^2$ has non-positive diagonal and non-negative off-diagonal elements, we need only ensure that

$$\left(\frac{\partial}{\partial h_m}\alpha_{jm} \right) f_{jm} \geq 0 \quad \text{for} \quad j \in K(m). \tag{5.191}$$

Once again, four cases must be considered:

- Case 1: $f_{jm} > 0$ and $\alpha_{jm} = \frac{Q_j^+}{P_j^+}$. Then $\left(\frac{\partial}{\partial h_m}\alpha_{jm} \right) f_{jm} = \frac{-Q_j^+}{(P_j^+)^2}(-m_{jm}) f_{jm} \geq 0$.
- Case 2: $f_{jm} > 0$ and $\alpha_{jm} = \frac{Q_m^-}{P_m^-}$. Then $\left(\frac{\partial}{\partial h_m}\alpha_{jm} \right) f_{jm} = \frac{-Q_m^-}{(P_m^-)^2} \left(\sum_{\substack{k \in K(m) \\ f_{mk} < 0}} m_{mk} \right) f_{jm} \geq 0$.
- Case 3: $f_{jm} < 0$ and $\alpha_{jm} = \frac{Q_j^-}{P_j^-}$. Then $\left(\frac{\partial}{\partial h_m}\alpha_{jm} \right) f_{jm} = \frac{-Q_j^-}{(P_j^-)^2}(-m_{jm}) f_{jm} \geq 0$.
- Case 4: $f_{jm} < 0$ and $\alpha_{jm} = \frac{Q_m^+}{P_m^+}$. Then $\left(\frac{\partial}{\partial h_m}\alpha_{jm} \right) f_{jm} = \frac{-Q_m^+}{(P_m^+)^2} \left(\sum_{\substack{k \in K(m) \\ f_{mk} > 0}} m_{mk} \right) f_{jm} \geq 0$.

Hence we have that $-\mathcal{O}^2$ is diagonally dominant by columns, being a matrix with zero column sums, non-negative diagonal and non-positive off-diagonal. The proof is thus complete. □

Unfortunately, the above proof gives only weak diagonal dominance and thus no criterion for non-singularity. A rather crude but easy fix to this would be to define more strictly the cut-off in the functions α_{ij}. Namely, instead of $\beta = 1$ let $\beta \in [0, 1)$ and redefine

$$\alpha_{ij} := \begin{cases} \min(R_i^+, R_j^-) & \text{for } f_{ij} \geq 0 \\ \min(R_i^-, R_j^+) & \text{for } f_{ij} < 0 \end{cases} \qquad \text{with} \qquad R_k^* := \begin{cases} \min\left(\frac{Q_k^*}{P_k^*}, \beta\right) & \text{for } P_k^* \neq 0 \\ \beta & \text{for } P_k^* = 0 \end{cases} \tag{5.192}$$

for $k \in \{i, j\}$ and $* \in \{+, -\}$. Then all the properties shown thus far for α_{ij} and $\alpha_{ij} f_{ij}$ involving piecewise smoothness, local Lipschitz continuity and formulae for directional derivatives hold analogously, but the occurrence of cancellation in the sum $M_C - \mathcal{O}^1 + P$ yields strict diagonal dominance:

Lemma 5.49. *Let Assumption 5.47 hold. With $\beta \in [0, 1)$ and the redefined functions α_{ij} from (5.192) it holds that $G := \nabla_h F(h, \tau)$ is strictly diagonally dominant by columns for $(h, \tau) \in \mathbb{R}^N \times \mathbb{R}_{>0} \setminus (N_1 \cup N_2)$. Specifically, for each column $m = 1, \ldots, N$ of this matrix there exists a constant $c_m(\beta) > 0$ independent of (h, τ) such that*

$$|g_{mm}| - \sum_{i \neq m} |g_{im}| \geq c_m > 0. \tag{5.193}$$

△

Proof. We can copy the proof of Lemma 5.48 verbatim up to equations (5.186) and (5.187), which can be strengthened due to the redefinition of the α_{ij}:

$$\frac{Q_i^+}{(P_i^+)^2} m_{im} \sum_{\substack{j \in K(i) \setminus \{m\} \\ \alpha_{ij} = Q_i^+ / P_i^+}} f_{ij} \leq \frac{Q_i^+}{P_i^+} m_{im} \leq \beta m_{im}, \tag{5.194}$$

$$\frac{Q_i^-}{(P_i^-)^2} m_{im} \sum_{\substack{j \in K(i) \setminus \{m\} \\ \alpha_{ij} = Q_i^- / P_i^-}} f_{ij} \leq \frac{Q_i^-}{P_i^-} m_{im} \leq \beta m_{im}, \tag{5.195}$$

so that

$$a_{im} + b_{im} \geq (1 - \beta) m_{im} \quad \text{for} \quad m \in K(i). \tag{5.196}$$

The matrix $A + B = M_C - \mathcal{O}^1 \geq 0$ is still non-negative with positive diagonal elements and both matrices $A + B$ and $-\mathcal{O}^2$ are still weakly diagonally dominant by columns with non-negative diagonal. What is new is that $A + B + P = M_L - \mathcal{N} - \mathcal{O}^1$ with $P := M_L - M_C - \mathcal{N}$ is now strictly diagonally dominant by columns. Since

$$p_{im} = \begin{cases} \sum_{j \in K(i)} (1 - \alpha_{ij}) m_{ij} & \text{for } m = i \\ (\alpha_{im} - 1) m_{im} & \text{for } m \neq i, \end{cases} \tag{5.197}$$

we have that P has positive diagonal and non-positive off-diagonal elements. When adding $Z := A + B$ and P, cancellation occurs due to the fact that $z_{im} \geq (1 - \beta) m_{im} > 0$ and $p_{im} \leq (\beta - 1) m_{im} < 0$ for $m \in K(i)$. Specifically, we set for $m = 1, \ldots, N$ the m-th "column dominance"

$$\Delta_m := \sum_{i \in K(m)} (|z_{im}| + |p_{im}| - |z_{im} + p_{im}|) \tag{5.198}$$

which allows the estimate and the (h,τ) independent definition

$$\Delta_m \geq 2(1-\beta)\sum_{i\in K(m)} m_{im} =: c_m. \tag{5.199}$$

As in the proof Lemma 5.38 (iii) we see that

$$z_{mm} + p_{mm} - \sum_{i\neq m} |z_{im} + p_{im}| \geq \Delta_m \geq c_m > 0 \tag{5.200}$$

and this carries over to $G := \nabla_h F(h,\tau) = Z + P - \mathcal{O}^2$. □

Proposition 5.50. *Under the condition $\theta = 0$, $d = 2$ and with the functions α_{ij} redefined as in* (5.192) *with some $\beta \in [0,1)$, the generalised Jacobian $\Pi_h \partial F(h,\tau)$ is non-singular for arbitrary $(h,\tau) \in \mathbb{R}^N \times \mathbb{R}_{>0}$.* △

Proof. Each element of $\Pi_h \partial F(h,\tau)$ can be obtained as the limit of a sequence $\nabla_h F(h_n,\tau_n)$ in $\mathbb{R}^N \times \mathbb{R}_{>0} \setminus (N_1 \cup N_2)$ with $(h_n,\tau_n) \to (h,\tau)$ as $n \to \infty$. All of these matrices were proven in Lemma 5.49 to be strictly diagonally dominant by columns with a "column dominance" at least $c_m > 0$ for each row $m = 1,\ldots,N$. The same must hold in the limit. □

5.4.3.5.3 Building In the Homogeneous Dirichlet Boundary Conditions. So far in Section 5.4, we have been ignoring the boundary conditions by basing our entire analysis on a non-linear blend of the high order and low order problems in (5.46). Thereby we have implicitly been solving the problem with Neumann rather than Dirichlet boundary conditions. Let us now show that the results obtained so far also hold when attention is paid to the boundary conditions.

Possibly after renumbering the nodes, we can assume that $\mathcal{I} := \{1,\ldots,M\} \subset \{1,\ldots,N\}$ for some $M < N$ is the set of interior node indices. Set

$$\mathcal{D} := \{1,\ldots,N\} \setminus \mathcal{I}, \tag{5.201}$$

the set of (Dirichlet) boundary nodes. A straightforward way to ensure that u^{n+1} vanishes at boundary nodes is to only solve for the interior node values $u^{n+1}_{\mathcal{I}}$ and to fix $u^{n+1}_{\mathcal{D}} = 0$ a priori. Since we then solve for the difference $h^\circ := (u^{n+1} - u^n)_{\mathcal{I}}$, this amounts to finding a unique $h^\circ = h^\circ(\tau) \in \mathbb{R}^M$ for $\tau > 0$ such that

$$\tilde{F}(h^\circ,\tau) = (M_L - \theta\tau L)^\circ h^\circ - \tau L^\circ u^n_{\mathcal{I}} - \tilde{f}(h^\circ,\tau) = 0, \tag{5.202}$$

where, as in the previous chapter, $A^\circ := A_{\mathcal{I}\mathcal{I}}$ for a matrix $A \in \mathbb{R}^{N\times N}$ and

$$\tilde{f} = \pi \circ \overline{f} \circ \iota \tag{5.203}$$

with

$$\pi : \mathbb{R}^N \to \mathbb{R}^M, \quad \pi(x) = x_{\mathcal{I}} \tag{5.204}$$

$$\iota : \mathbb{R}^M \times \mathbb{R}_{>0} \to \mathbb{R}^N \times \mathbb{R}_{>0}, \quad \iota(h^\circ,\tau) = \left(\begin{bmatrix} h^\circ \\ 0 \end{bmatrix}, \tau\right). \tag{5.205}$$

Slightly abusing notation, we set $\alpha_{ij}(h^\circ,\tau) := (\alpha_{ij}\circ\iota)(h^\circ,\tau)$ and $f_{ij}(h^\circ,\tau) := (f_{ij}\circ\iota)(h^\circ,\tau)$ and see immediately that $\alpha_{ij}f_{ij} \in PC^\infty(\mathbb{R}^M\times\mathbb{R}_{>0})$ and is locally Lipschitz continuous for each $i\in\mathcal{I}, j\in K(i)$. Furthermore, for $i,m\in\mathcal{I}$ and $j\in K(i)$ ($j\in\mathcal{D}$ is allowed!), the derivatives $\frac{\partial}{\partial h_m^\circ}\alpha_{ij}(h_0^\circ,\tau_0)$ and $\frac{\partial}{\partial h_m^\circ}f_{ij}(h_0^\circ,\tau_0)$ can still be computed according to the rules of Lemma 5.44 as long as $(h_0^\circ,\tau_0)\in\mathbb{R}^M\times\mathbb{R}_{>0}\setminus(N_1\cup N_2)$, where the null sets N_1 and N_2 are now defined by

$$N_1 := \bigcup_{\substack{i\in\mathcal{I}\\ j\in K(i)}} N_{ij} \quad\text{with}\quad N_{ij} = \{(h^\circ,t)\in\mathbb{R}^{M+1} : \alpha_{ij}f_{ij} \text{ is not differentiable at } (h^\circ,\tau)\}, \tag{5.206}$$

$$N_2 := \bigcup_{\substack{i\in\mathcal{I}\\ j\in K(i)}} H_{ij}^\circ \quad\text{with}\quad H_{ij}^\circ = \{(h^\circ,\tau)\in\mathbb{R}^M\times\mathbb{R}_{>0} : f_{ij}(h^\circ,\tau)=0\}. \tag{5.207}$$

Then at $(h_0^\circ,\tau_0)\in\mathbb{R}^M\times\mathbb{R}_{>0}\setminus(N_1\cup N_2)$ we can decompose analogously as before

$$\nabla_{h^\circ}\tilde{f}(h^\circ,\tau) = \mathcal{A}+\mathcal{B} = \mathcal{A}+\mathcal{B}^1+\mathcal{B}^2 \tag{5.208}$$

with $\mathcal{A},\mathcal{B},\mathcal{B}^1,\mathcal{B}^2\in\mathbb{R}^{M\times M}$ defined by

$$\mathcal{A}_{im} := \begin{cases}\sum_{k\in K(i)}\alpha_{ik}(h^\circ,\tau)c_{ik}(\tau) & \text{for } m=i\\ -\alpha_{im}(h^\circ,\tau)c_{im}(\tau) & \text{for } m\neq i\end{cases} \tag{5.209}$$

$$\mathcal{B}_{im} := \sum_{j\in K(i)}\left(\frac{\partial}{\partial h_m^\circ}\alpha_{ij}(h^\circ,\tau)\right)f_{ij}(h^\circ,\tau) \tag{5.210}$$

and

$$\mathcal{B}^1_{im} := \begin{cases}0\\ \sum_{j\in K(i)\setminus\{m\}}\left(\frac{\partial}{\partial h_m^\circ}\alpha_{ij}\right)f_{ij}\end{cases}, \quad \mathcal{B}^2_{im} := \begin{cases}\sum_{j\in K(m)}\left(\frac{\partial}{\partial h_m^\circ}\alpha_{mj}\right)f_{mj} & \text{for } m=i\\ \left(\frac{\partial}{\partial h\circ_m}\alpha_{im}\right)f_{im}\mathbb{1}_{\{m\in K(i)\}} & \text{for } m\neq i.\end{cases} \tag{5.211}$$

Again, let Assumption 5.47 hold, i.e. $d=2$ and $\theta=0$. We now need to show four things to ensure diagonal dominance by columns of $\nabla\tilde{F}(h^\circ,\tau) = M_L^\circ - \nabla_{h^\circ}\tilde{f}(h^\circ,\tau)$:

(i) $M_C^\circ - \mathcal{B}^1 \geq 0$

(ii) $\Sigma_m(-\mathcal{B}^1) \leq \Delta_m(M_C^\circ)$ for $m=1,\dots,M$, with the notation of Lemma 5.39.

(iii) $-\mathcal{B}^2$ is diagonally dominant by columns with non-negative diagonal elements.

(iv) $\mathcal{P} := M_L^\circ - M_C^\circ - \mathcal{A}$ is diagonally dominant by columns with non-negative diagonal elements.

Then (i) and (ii) in conjunction with Lemma 5.39 imply that $M_C^\circ - \mathcal{B}^1$ is diagonally dominant by columns and, together with (iii) and (iv), that $M_L^\circ - \nabla_{h^\circ}\tilde{f}(h^\circ,\tau)$ is diagonally dominant by columns.

Ad (i). This was already proven in Lemma 5.48: Note that $(M_C^\circ)_{im} = (M_C)_{im}$ and $(\mathcal{B}^1)_{im} = (\mathcal{O}^1)_{im}$ for $i,m\in\mathcal{I}$.

Ad (ii). Since $M_C \geq 0$ has weakly diagonally dominant columns, we have for $m = 1, \ldots, M$

$$\Delta_m(M_C^\circ) = \sum_{j \in K(m) \cap \mathcal{D}} m_{jm} \geq 0. \tag{5.212}$$

On the other hand, by the same symmetry argument used in the proof of Lemma 5.48 to show that $-\mathcal{O}^1$ has zero column sums, we obtain for $m = 1, \ldots, M$ that

$$\Sigma_m(-\mathcal{B}^1) = -\sum_{i \in \mathcal{I}} \sum_{j \in K(i) \cap \mathcal{D}} \left(\frac{\partial}{\partial h_m^\circ} \alpha_{ij}(h^\circ, \tau) \right) f_{ij}(h^\circ, \tau). \tag{5.213}$$

We omit the argument (h°, τ) and estimate

$$\begin{aligned}
\Sigma_m(-\mathcal{B}^1) &\leq -\sum_{i \in \mathcal{I}} \sum_{j \in K(i) \cap \mathcal{D}} \left[\left(\frac{\partial}{\partial h_m^\circ} \alpha_{ij} \right) f_{ij} \right]^- \\
&= \sum_{i \in I} \sum_{\substack{j \in K(i) \cap \mathcal{D} \\ \alpha_{ij} = Q_j^-/P_j^- \\ f_{jm} < 0}} \left(\frac{-Q_j^-}{(P_j^-)^2} f_{ij} \right) m_{jm} + \sum_{i \in I} \sum_{\substack{j \in K(i) \cap \mathcal{D} \\ \alpha_{ij} = Q_j^+/P_j^+ \\ f_{jm} > 0}} \left(\frac{-Q_j^+}{(P_j^+)^2} f_{ij} \right) m_{jm} \\
&= \sum_{\substack{j \in K(m) \cap \mathcal{D} \\ f_{jm} < 0}} m_{jm} \frac{Q_j^-}{(P_j^-)^2} \sum_{\substack{i \in \mathcal{I} \\ j \in K(i) \\ \alpha_{ij} = Q_j^-/P_j^-}} f_{ji} + \sum_{\substack{j \in K(m) \cap \mathcal{D} \\ f_{jm} > 0}} m_{jm} \frac{Q_j^+}{(P_j^+)^2} \sum_{\substack{i \in \mathcal{I} \\ j \in K(i) \\ \alpha_{ij} = Q_j^+/P_j^+}} f_{ji} \\
&\leq \sum_{\substack{j \in K(m) \cap \mathcal{D} \\ f_{jm} < 0}} m_{jm} \frac{Q_j^-}{P_j^-} + \sum_{\substack{j \in K(m) \cap \mathcal{D} \\ f_{jm} > 0}} m_{jm} \frac{Q_j^+}{P_j^+} \leq \sum_{j \in K(m) \cap \mathcal{D}} m_{jm} = \Delta_m(M_C^\circ),
\end{aligned} \tag{5.214}$$

where we have used the calculations of the first Cases 1 – 4 in the proof of Lemma 5.48 for the first equality sign.

Ad (iii). Using what we have proven for $-\mathcal{O}^2$ in Lemma 5.48, we see that $Z := -\mathcal{B}^2$ is diagonally dominant by columns with

$$|z_{ii}| - \sum_{j \neq i} |z_{ji}| = -\sum_{j \in K(i) \cap \in \mathcal{D}} \mathcal{O}_{ji}^2 \geq 0. \tag{5.215}$$

Ad (iv). The elements of $\mathcal{P}$ for $i, m \in \{1, \ldots, M\}$ are given by

$$\mathcal{P}_{im} = \begin{cases} \sum_{k \in K(m)} m_{mk}(1 - \alpha_{mk}) \geq 0 & \text{for } m = i \\ (\alpha_{im} - 1) m_{im} \leq 0 & \text{for } m \neq i, \end{cases} \tag{5.216}$$

so that for $m = 1, \ldots, M$

$$\sum_{i \in \mathcal{I}} \mathcal{P}_{im} = \sum_{k \in K(m) \cap \mathcal{D}} (1 - \alpha_{mk}) m_{mk} \geq 0. \tag{5.217}$$

Again, we are only able to show weak diagonal dominance and an argument to either show strict dominance in every column or irreducibility and strictness in one column may be possible but would have to be attained by a more careful analysis.
If the α_{ij} are redefined to be cut off at values $\beta \in [0,1)$ instead of $\beta = 1$ as (5.192), then the addition of $M_C^\circ - \mathcal{B}^1$ and $\mathcal{P}$ introduces cancellation such that $\nabla \tilde{F}(h^\circ, \tau)$ is strictly diagonally dominant by columns with uniform positive strictness in $\mathbb{R}^M \times \mathbb{R}_{>0} \setminus (N_1 \cup N_2)$ and we obtain

Proposition 5.51. *Under the condition $\theta = 0$, $d = 2$ and with the functions α_{ij} redefined as in (5.192) with some $\beta \in [0,1)$, the generalized Jacobian $\Pi_h \partial \tilde{F}(h^\circ, \tau)$ is non-singular for arbitrary $(h^\circ, \tau) \in \mathbb{R}^M \times \mathbb{R}_{>0}$.* △

5.4.4 Semi-smooth Newton Method

We now take some material from [Ulb02, Chapter 2] to show that our $F := F(\cdot, \tau)$ defined in (5.72) is semi-smooth for all $\tau \geq 0$. This property is needed to prove convergence of the following algorithm:

Algorithm 5.52 (Semi-smooth Newton Method)**.** Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz, $x_0 \in \mathbb{R}^n$. Set $k := 0$.

(1) Unless a stopping criterion is met, solve

$$G(x_k) d_k = -f(x_k) \tag{5.218}$$

for d_k, where $G(x_k) \in \partial f(x_k)$.

(2) Set $x_{k+1} := x_k + d_k$ and $k = k + 1$ and go to step (1). △

Definition 5.53 (Semi-smoothness, higher order semi-smoothness)**.** Let $V \subset \mathbb{R}^n$ be open, $f : V \to \mathbb{R}^m$ and $x \in V$.

(i) f is called *semi-smooth at x* if it is locally Lipschitz around x and one of the three equivalent conditions holds:

(a) The limit

$$\lim_{\substack{G \in \partial f(x + t\tilde{d}) \\ \tilde{d} \to d,\ t \downarrow 0}} G\tilde{d} \tag{5.219}$$

exists for all all $d \in \mathbb{R}^n$.

(b) All one-sided directional derivatives $f'(x; d)$ $(d \in \mathbb{R}^n)$ at x exist and

$$\sup_{G \in \partial f(x+d)} \|f(x + d) - f(x) - Gd\| = \mathcal{O}(\|d\|) \quad \text{as } d \to 0. \tag{5.220}$$

(c) All one-sided directional derivatives $f'(x; d)$ $(d \in \mathbb{R}^n)$ at x exist and

$$\sup_{G \in \partial f(x+d)} \big\|Gd - f'(x; d)\big\| = \mathcal{O}(\|d\|) \quad \text{as } d \to 0. \tag{5.221}$$

(ii) f is called *α-order semi-smooth* at x for $\alpha \in (0, 1]$ if it is locally Lipschitz around x, $f'(x; \cdot)$ exists and one of the two equivalent conditions holds:

(a)
$$\sup_{G\in\partial f(x+d)} \|f(x+d) - f(x) - Gd\| = \mathcal{O}(\|d\|^{1+\alpha}) \quad \text{as } d \to 0. \tag{5.222}$$

(b)
$$\sup_{G\in\partial f(x+d)} \left\|Gd - f'(x;d)\right\| = \mathcal{O}(\|d\|^{1+\alpha}) \quad \text{as } d \to 0. \tag{5.223}$$

△

Remark 5.54. The equivalences of the above definition are the subject of Propositions 2.7 and 2.14 of [Ulb02]. △

Theorem 5.55 (Local convergence of semi-smooth Newton method)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz continuous and semi-smooth at x^*, $f(x^*) = 0$ and $\partial f(x^*)$ non-singular. Then there exists $\varepsilon > 0$ such that for all initial $x_0 \in B_\varepsilon(x^*)$ the sequence $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 5.52 is well-defined, converges to x^* and satisfies*

$$\|x_{k+1} - x^*\| = o(\|x_k - x^*\|) \quad \text{for } k \to \infty. \tag{5.224}$$

If in addition f is α-order semi-smooth at x^* for some $\alpha \in (0, 1]$ then the convergence rate improves:

$$\|x_{k+1} - x^*\| = \mathcal{O}(\|x_k - x^*\|^{1+\alpha}) \quad \text{for } k \to \infty. \tag{5.225}$$

△

Proof. We note, as is proven in [Hin10, Theorem 2.8.], that upper semicontinuity of ∂f implies the existence of $C, \delta > 0$ such that G is invertible with $\left\|G^{-1}\right\| \leq C$ for all $G \in \partial f(x)$, $x \in B_\delta(x^*)$. The rest of this proof is taken from [Ulb02, Propositions 2.12 and 2.18].

(1) Set $e_k := x_k - x^*$ and let $G_k \in \partial f(x_k)$ be the choice made in Algorithm 5.52. Then $G_k d_k = -f(x_k)$ and

$$G_k e_{k+1} = G_k(d_k + e_k) = -f(x_k) + G_k e_k = -(f(x^* + e_k) - f(x^*) - G_k e_k). \tag{5.226}$$

Setting $k = 0$ and employing (5.220) we see that we can choose $x_0 \in B_\delta(x^*)$ such that $G_0 e_1 \leq (2C)^{-1}\|e_0\|$. Then

$$\|e_1\| \leq \|G_0^{-1}\|\|G_0 e_1\| \leq \frac{1}{2}\|e_0\|, \tag{5.227}$$

from which q-linear convergence $e_k \to 0$, $k \to \infty$ follows by induction.

(2) Having proved convergence, we can use (5.226) and (5.220) once again to obtain q-superlinear convergence:

$$\|G_k e_{k+1}\| = o(\|e_k\|) \quad \text{as } \|e_k\| \to 0. \tag{5.228}$$

(3) Now consider the case that f is α-order semi-smooth for some $\alpha \in (0, 1]$. From (5.226) and (5.222) follows

$$\|G_k e_{k+1}\| = \mathcal{O}(\|e_k\|^{1+\alpha}) \quad \text{as } \|e_k\| \to 0. \tag{5.229}$$

As before, q-linear convergence follows by induction for $\|x^* - x_0\|$ small enough and this time we obtain the improved rate

$$\|e_{k+1}\| \leq \left\|G_k^{-1}\right\| \|G_k e_{k+1}\| = \mathcal{O}(\|e_k\|^{1+\alpha}). \tag{5.230}$$

□

Proposition 5.56 ([Ulb02, Proposition 2.26])**.** *Let $V \subset \mathbb{R}^n$. If $f \in PC^1(V, \mathbb{R}^m)$, then f is semi-smooth on V. If $f \in PC^2(V, \mathbb{R}^m)$, then f is even 1-order semi-smooth on V.* △

Corollary 5.57. *For any fixed $\tau \geq 0$, the function $F(h) := F(h, \tau) = (M_L - \theta\tau L)h - \overline{f}(h, \tau) - \tau L u^n$ is 1-order semi-smooth on $\mathbb{R}^N$. In particular, if h^* is such that $F(h^*) = 0$ and $\partial F(h^*)$ is non-singular, then the semi-smooth Newton method from Algorithm 5.52 converges locally at quadratic rate.* △

Proof. Combine Lemma 5.27, Proposition 5.56 and Theorem 5.55. □

6. Conclusion

In this thesis we have examined transient convection-diffusion equations with dominant convection and homogeneous Dirichlet boundary conditions on triangulated bounded domains $\Omega \subset \mathbb{R}^d$ for $d \geq 1$ both from a theoretical and numerical analysis point of view.

On the purely theoretical side, we have shown uniform converge of regular solutions on certain subdomains of C^2 domains Ω to the solution of the reduced problem as $\epsilon \downarrow 0$.

In the case $d = 1$, with the parabolic maximum principle and Proposition 4.2 about the decreasing nature of total variation of classical solutions we have seen that the oscillations observed in the V_0^1 finite element discretisations are indeed unphysical as they lead to values exceeding the range of the initial boundary conditions and to an increase in total variation. The criterion given by Harten's lemma (Proposition 4.5) has allowed us to show that the physically plausible manipulation of upwinding the convective C part of the stiffness matrix $C + \epsilon D$ makes the mass-lumped explicit Euler scheme TVD and thus free of oscillations under merely a CFL-like condition on the time step size τ. Later we have seen that, as a linear LED scheme, it cannot be of consistency order greater than 1, which is often called Godunov's order barrier and manifests itself in strongly smeared solutions.

For $d = 2$ it was then proved that, for elementwise constant divergence-free fields b, the mass-lumped semi-discrete method can be interpreted as a finite volume scheme over the barycentric dual mesh with a central flux approximation for the convective flux and that performing an upwinding of this finite volume scheme amounts to the addition of a discrete diffusion/upwinding matrix Y just large enough to cancel all non-negative off-diagonal entries of $-C$. We have thus elucidated somewhat this purely algebraic step. Projecting a general $b \in W^{1,\infty}(\Omega)$ onto the lowest order Raviart-Thomas space $\mathcal{RT}_0$, we have argued that the method resulting from such an upwinding belongs to the class of upwind finite element methods of Baba and Tabata.
Furthermore, we have seen that the resulting scheme is local extremum diminishing (LED) for $d = 2$ if $\mathcal{T}$ is a Delaunay triangulation, since in this case the diffusive part $-\epsilon D$ has non-negative off-diagonal elements. Unfortunately, the concept of total variation for functions from $V_0^1(\mathcal{T})$ loses its meaning in two dimensions, so that even characterising what "being oscillatory" means for a scheme becomes less obvious than in 1D. Upwinding on an irregular triangular mesh still permits spurious oscillations when the directions of advection and a gradient in the discrete solution are transverse, which was demonstrated in Remark 4.32. However, the LED property of the semi-discrete upwinded scheme ensures that a discrete version of the weak maximum principle is respected.

The remainder of the work was dedicated to the non-linear variant of FCT suggested in [Kuz10] which aims at restoring the high order finite element scheme to the greatest extent while suppressing the emergence and enhancement of oscillations. By reformulating the problem in the form $F(h(\tau), \tau) = 0$ and showing that F is piecewise smooth and thus locally Lipschitz continuous, the application of an implicit function theorem for Lipschitz continuous functions was possible.
The required regularity of the generalised Jacobian $\Pi_h \partial F$ at points (h_0, τ_0) with $F(h_0, \tau_0) = 0$ at which one wishes to extend this equation to τ from some interval $[\tau_0, \tau_0 + \delta]$ was successful for $(h_0, \tau_0) = (0, 0)$, since the nonlinear function $F(\cdot, 0)$ becomes affine on a neighbourhood of $h_0 = 0$. Hence we have shown that for some positive time-step bound $\delta > 0$ there exists a unique solution $h(\tau)$ to this non-linear problem for time steps $\tau \leq \delta$.
The attempt undertaken for the explicit Euler method ($\theta = 0$) to extend this result and show that the generalised Jacobians $\Pi_h \partial F(h, \tau)$ are non-singular for any (h, τ) with $\tau > 0$ was complicated by the fact that the mass matrix M_C is strictly diagonally dominant only for $d = 1$. Therefore we have only been able to show weak diagonal dominance, but no strict or irreducible diagonal dominance. It seems that this desired result is either formulated too strongly or that, at least for $d = 3$ or $\theta > 0$, a different kind of non-singularity of $\Pi_h \partial F(h, \tau)$ has to be proved, e.g. positive definiteness.

If a $h(\tau)$ satisfying $F(h(\tau), \tau) = 0$ exists and $\Pi_h \partial F(h(\tau), \tau)$ is non-singular, however, then the semi-smoothness of $F(\cdot, \tau)$ implied by its piecewise smoothness guarantees locally quadratic convergence of the semi-smooth Newton method.

Apart from solvability, other questions regarding the analysis of the non-linear FCT method still need to be addressed properly, most fundamentally whether or under which time-step restrictions it diminishes local extrema and what its rate of convergence in terms of powers of h is.

A. Tools from the Theory of Finite Elements

Theorem A.1 (Trace inequality)**.** *Let $\Omega \subset \mathbb{R}^d$ be a domain with Lipschitz boundary and $p \in [1, \infty]$. Then*

$$\|v\|_{L^p(\partial\Omega)} \leq C(\Omega) \|v\|_{L^p(\Omega)}^{1-1/p} \|v\|_{W^{1,p}(\Omega)}^{1/p} \quad \text{for all } v \in W^{1,p}(\Omega).$$

△

Proof. See [BS08], Theorem 1.6.6. □

Corollary A.2 (Scaled L^1 trace inequality for simplices)**.** *Let $T \subset \mathbb{R}^d$ be a d-simplex and S one of its sides. Then for $v \in W^{1,1}(T)$ it holds that*

$$\|v\|_{L^1(\partial S)} \leq C(d, \sigma_T)(h_T^{-1} \|v\|_{L^1(T)} + \|\nabla v\|_{L^1(T)}) \tag{A.1}$$

△

Proof. This follows from by transformation onto the standard simplex $\hat{T}$, the trace inequality in Theorem A.1 and transformation back onto T. □

Lemma A.3 (Interpolation error on simplices)**.** *Let $T \subset \mathbb{R}^d$ be a d-simplex, $\hat{T}$ the standard d-simplex and $F : \hat{T} \to T, F(\hat{x}) = A\hat{x} + \tau$ an affine bijection. Let $k, m \in \mathbb{N}_0$ and $p, q \in [1, \infty]$ be such that the continuous embedding*

$$W^{k+1,p}(\hat{T}) \hookrightarrow W^{m,q}(\hat{T})$$

by inclusion exists and let $\hat{I} \in L(W^{k+1,p}(\hat{T}), W^{m,q}(\hat{T}))$ be a bounded linear operator with

$$\hat{I}|_{\mathbb{P}^k(\hat{T})} = \mathrm{id}_{\mathbb{P}^k(\hat{T})}$$

(an interpolation operator). Then for the interpolation operator $I \in L(W^{k+1,p}(T), W^{m,q}(T))$ given by $Iu \circ F = \hat{I}(u \circ F)$ the error estimate

$$|u - Iu|_{W^{m,q}(T)} \leq C\sigma(T)^{m - d\min\{0, \frac{1}{q} - \frac{1}{p}\}} h_T^{k+1-m+d\left(\frac{1}{q} - \frac{1}{p}\right)} |u|_{W^{k+1,p}(T)} \tag{A.2}$$

holds for each $u \in W^{k+1,p}(T)$, and $C = C(k, m, p, q, \hat{T}, \|\hat{I}\|)$. △

Proof. See [Dzi10, Satz 3.31]. □

B. Tools from the Theory of Ordinary Differential Equations

Theorem B.1 (Gronwall's inequality in integral form, [Joh16, Lemma A.53]). *Let $T \in (0, \infty]$, $f, g \in L^\infty((0,T), \mathbb{R})$ and $\lambda \in L^1((0,T), \mathbb{R}_{\geq 0})$. If the implict estimate*

$$f(t) \leq g(t) + \int_0^t \lambda(s) f(s)\ ds \quad \text{for a.e. } t \in [0,T] \tag{B.1}$$

holds, then so does explicit estimate

$$f(t) \leq g(t) + \int_0^t \exp\left(\int_s^t \lambda(\tau)\ d\tau\right) \lambda(s) g(s)\ ds \quad \text{for a.e. } t \in [0,T]. \tag{B.2}$$

If g is continuous and montonically increasing, then

$$f(t) \leq \exp\left(\int_0^t \lambda(\tau)\ d\tau\right) g(t). \tag{B.3}$$

△

Theorem B.2 (Gronwall's inequality in differential form, [Eva10, Appendix B.2 j]). *Let $f \in C([0,T], \mathbb{R}_{\geq 0})$ be absolutely continuous and $g, \lambda \in L^1([0,T], \mathbb{R}_{\geq 0})$ such that*

$$f'(t) \leq \lambda(t) f(t) + g(t) \quad \text{for a.e. } t \in [0,T]. \tag{B.4}$$

Then

$$f(t) \leq \exp\left(\int_0^t \lambda(s)\ ds\right)\left(f(0) + \int_0^t g(s)\ ds\right) \quad \text{for all } t \in [0,T]. \tag{B.5}$$

△

Theorem B.3 (Carathéodory's local existence and uniqueness theorem). *Let $t_0 \in \mathbb{R}$, $y_0 \in \mathbb{R}^n$, $b, T > 0$, $\Omega_y := B_b(y_0)$ and $\Omega_T := (t_0, t_0 + T) \times \Omega_y$. Consider a function $f : \Omega_T \to \mathbb{R}^n$ meeting the following* Carathéodory conditions*:*

(a) For each fixed $y \in \Omega_y$ the function $t \mapsto f(t,y)$ is measurable.

(b) For each fixed $t \in (t_0, t_0 + T)$ the function $y \mapsto f(t,y)$ is continuous.

(c) There exists a function $F \in L^1((t_0, t_0+T), \mathbb{R}_{\geq 0})$ *such that* $\|f(t,y)\| \leq F(t)$ *for any* $(t,y) \in \Omega_T$.

Then the following statements hold true:

(i) There exists some $a \in (0,T]$ *and an absolutely continuous function* $y : [t_0, t_0+a] \to \mathbb{R}^n$ *solution to the integral form of the initial value problem* $y'(t) = f(t, y(t))$ *and* $y(0) = y_0$, *i.e.*

$$y(t) = y_0 + \int_0^t f(s, y(s))\, ds \quad \text{for all } t \in [t_0, t_0+a]. \tag{B.6}$$

(ii) At each $t \in [t_0, t_0+a]$ *such that the integrand is continuous at* t, *the strong form*

$$y'(t) = f(t, y(t)) \tag{B.7}$$

of the ordinary differential equation holds at t.

(iii) If in addition to the above Carathédory conditions the Lipschitz condition

$$\|f(t, y_1) - f(t, y_2)\| \leq G(t)\, \|y_1 - y_2\| \tag{B.8}$$

holds on Ω_T *for some* $G \in L^1((t_0, t_0+T), \mathbb{R}_{\geq 0})$, *then the solution on* $[t_0, t_0+a]$ *is unique.* △

Proof. References to proofs can be found in [Joh16, Theorem A.50]. □

Bibliography

[Alt16] Hans Wilhelm Alt. *Linear Functional Analysis*. 1st ed. Universitext. Springer London, 2016.

[Bar92] Timothy J. Barth. "Aspects of Unstructured Grids and Finite-Volume Solvers for the Euler and Navier-Stokes Equations". In: *AGARD Report 787: Special course on Unstructured Grid Methods for Advection Dominated Flows*. NATO, May 1992.

[BB73] Jay P. Boris and David L. Book. "Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works". In: *Journal of Computational Physics* 11.1 (1973), pp. 38 –69.

[BB76] J.P Boris and D.L Book. "Flux-corrected transport. III. Minimal-error FCT algorithms". In: *Journal of Computational Physics* 20.4 (1976), pp. 397 –431.

[BBH75] D.L. Book, J.P. Boris, and K. Hain. "Flux-corrected transport II: Generalizations of the method". In: *Journal of Computational Physics* 18.3 (1975), pp. 248 –283.

[BS08] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. 3rd ed. Vol. 15. Texts in Applied Mathematics. Springer, 2008.

[BT81] Kinji Baba and Masahisa Tabata. "On a conservation upwind finite element scheme for convective diffusion equations". In: *RAIRO. Anal. numér.* 15.1 (1981), pp. 3–25.

[Cla75] Frank H. Clarke. "Generalized Gradients and Applications". In: *Transactions of the American Mathematical Society* 205 (1975), pp. 247–262.

[CR73] P.G. Ciarlet and P.-A. Raviart. "Maximum principle and uniform convergence for the finite element method". In: *Computer Methods in Applied Mechanics and Engineering* 2 (1973), pp. 17–31.

[Dzi10] Gerhard Dziuk. *Theorie und Numerik partieller Differentialgleichungen*. 1st ed. De Gruyter Studium. De Gruyter, 2010.

[Eva10] Lawrence C. Evans. *Partial Differential Equations*. 2nd ed. Vol. 19. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, 2010.

[FP87] M. Fabian and D. Preiss. "On the Clarke's generalized jacobian". In: *Proceedings of the 14th Winter School on Abstract Analysis*. Palermo: Circolo Matematico di Palermo, 1987, [305]–307.

[Fri64] Avner Friedman. *Partial Differential Equations Of Parabolic Type*. Reprint 1983. Robert E. Krieger Publishing Company, 1964.

[Fri72] I. Fried. “Bounds on the extremal eigenvalues of the finite element stiffness and mass matrices and their spectral condition number”. In: *Journal of Sound and Vibration* 22.4 (1972), pp. 407 –418.

[GFLRT83] H. Goering, A. Felgenhauer, G. Lube, H.-G. Roos, and L. Tobiska. *Singularly Perturbed Differential Equations*. Vol. 13. Akademie-Verlag Berlin, 1983.

[Har83] Ami Harten. “High resolution schemes for hyperbolic conservation laws”. In: *Journal of Computational Physics* 49.3 (1983), pp. 357 –393.

[Hin10] Michael Hintermüller. *Semismooth Newton Methods and Applications*. Oberwolfach-Seminar on Mathematics of PDE-Constrained Optimization. 2010. URL: `https://www.math.uni-hamburg.de/home/hinze/Psfiles/Hintermueller_OWNotes.pdf` (visited on 02/12/2017).

[HV03] Willem Hundsdorfer and Jan Verwer. *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. 1st ed. Vol. 33. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2003.

[Jam95] A. Jameson. “Analysis and design of numerical schemes for gas dynamics, 1: artificial diffusion, upwind biasing, limiters and their effect on accuracy and multigrid convergence”. In: *International Journal of Computational Fluid Dynamics* 4 (1995), pp. 171–218.

[Joh16] Volker John. *Finite Element Methods for Incompressible Flow Problems*. Vol. 51. Springer Series in Computational Mathematics. Springer International Publishing, 2016.

[KK05] S. Korotov and M. Křížek. “Global and local refinement techniques yielding nonobtuse tetrahedral partitions”. In: *Computers & Mathematics with Applications* 50.7 (2005). Numerical Methods and Computational Mechanics, pp. 1105 –1113.

[KPP12] Eryk Kopczyński, Igor Pak, and Piotr Przytycki. “Acute triangulations of polyhedra and $\mathbb{R}^N$”. In: *Combinatorica* 32.1 (2012), pp. 85–110.

[Kum91] Bernd Kummer. “An implicit-function theorem for $C^{0,1}$-equations and parametric $C^{1,1}$-optimization”. In: *Journal of Mathematical Analysis and Applications* 158.1 (1991), pp. 35 –46.

[Kuz10] Dmitri Kuzmin. *A Guide to Numerical Methods for Transport Equations*. 2010. URL: `http://www.mathematik.uni-dortmund.de/~kuzmin/Transport.pdf` (visited on 01/13/2018).

[MKK07] M. Möller, D. Kuzmin, and D. Kourounis. “Implicit FEM-FCT algorithms and discrete Newton methods for transient convection problems”. In: *International Journal for Numerical Methods in Fluids* 57.6 (2007), pp. 761–792.

[QV94] Alfio Quarteroni and Alberto Valli. *Numerical Approximation of Partial Differential Equations*. 1st ed. Vol. 23. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 1994.

[Sat69] D.H. Sattinger. “On the Total Variation of Solutions of Parabolic Equations.” In: *Mathematische Annalen* 183 (1969), pp. 78–92.

[Sch12] Stefan Scholtes. *Introduction to Piecewise Differentiable Equations*. SpringerBriefs in Optimization. Springer, 2012.

[Sel93] V. Selmin. “The node-centred finite volume approach: Bridge between finite differences and finite elements”. In: *Computer Methods in Applied Mechanics and Engineering* 102.1 (1993), pp. 107 –138.

[Thi82] Lionel Thibault. “On generalized differentials and subdifferentials of Lipschitz vector-valued functions”. In: *Nonlinear Analysis: Theory, Methods & Applications* 6.10 (1982), pp. 1037 –1053.

[Ulb02] Michael Ulbrich. “Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces”. habilitation. Technische Universität München, submitted: 2001, accepted: 2002.

[Var00] Richard S. Varga. *Matrix Iterative Analysis*. 2nd ed. Springer Series in Computational Mathematics 27. Springer-Verlag Berlin Heidelberg, 2000.

[Zal79] Steven T. Zalesak. “Fully multidimensional flux-corrected transport algorithms for fluids”. In: *Journal of Computational Physics* 31.3 (1979), pp. 335 –362.

Statement of Authorship (Selbstständigkeitserklärung)

Ich, Paul Korsmeier, geboren am 19. Dezember 1991, Matrikelnummer 5040624, erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Masterarbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe.

Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch bei keiner anderen Universität als Prüfungsleistung eingereicht.

Berlin, 30. Juli 2018

_____

Paul Korsmeier